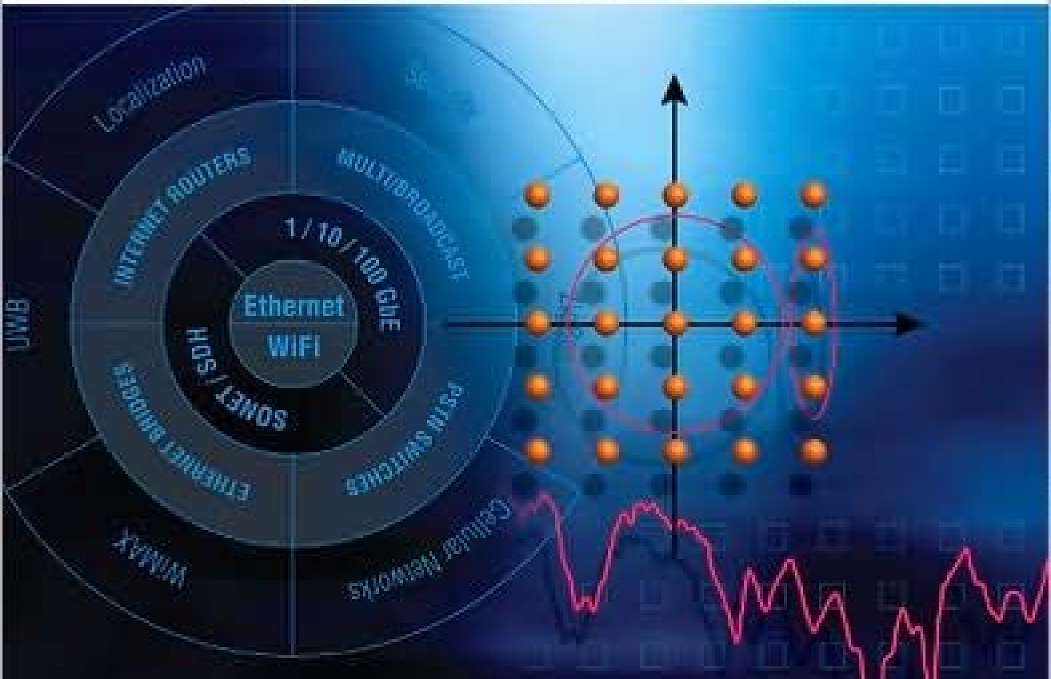
Networking Fundamentals

Wide, Local and Personal Area Communications



Kaveh Pahlavan
Prashant Krishnamurthy

WILEY

NETWORKING FUNDAMENTALS

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

NETWORKING FUNDAMENTALS

Wide, Local and Personal Area Communications

KAVEH PAHLAVAN

Worcester Polytechnic Institute, USA

PRASHANT KRISHNAMURTHY

University of Pittsburgh, USA

WILEY

A John Wiley and Sons, Ltd, Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Pahlavan, Kaveh, 1951-
Networking fundamentals : wide, local, and personal area communications /
Kaveh Pahlavan.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-99289-0 (cloth) – ISBN 978-0-470-99290-6 (pbk.) 1.
Computer networks. I. Title.
TK5105.5.P343 2009
004.6–dc22
2009004131

A catalogue record for this book is available from the British Library.

ISBN 9780470992890 (H/B) 9780470992906 (PBK)

Typeset in 10/12pt Times by Thomson Digital, Noida, India.
Printed and bound in Great Britain by Antony Rowe, Chippenham, UK

CONTENTS

ABOUT THE AUTHORS

Kaveh Pahlavan, is a Professor of Electrical and Computer Engineering (ECE), a Professor of Computer Science (CS), and Director of the Center for Wireless Information Network Studies, Worcester Polytechnic Institute (WPI), Worcester, MA. He is also a visiting Professor of Telecommunication Laboratory and CWC, University of Oulu, Finland. His area of research is location-aware broadband wireless indoor networks. He has contributed to numerous seminal technical publications and patents in this field. He is the principal author of the *Wireless Information Networks* (with Allen Levesque), John Wiley and Sons, 1995 and *Principles of Wireless Networks – A Unified Approach* (with P. Krishnamurthy), Prentice Hall, 2002. He has been a consultant to a number of companies, including CNR Inc., GTE Laboratories, Steinbrecher Corp., Simplex, Mercury Computers, WINDATA, SieraComm, 3COM, and Codex/Motorola in Massachusetts; JPL, Savi Technologies, RadioLAN in California; Aironet in Ohio; United Technology Research Center in Connecticut; Honeywell in Arizona; Nokia, LK-Products, Elektrobit, TEKES, the Finnish Academy in Finland; and NTT in Japan. Before joining WPI, he was the director of advanced development at Infinite Inc., Andover, MA, working on data communications. He started his career as an assistant professor at Northeastern University, Boston, MA. He is the Editor-in-Chief of the *International Journal on Wireless Information Networks*. He was the founder, the program chairman, and organizer of the IEEE Wireless LAN Workshop, Worcester, in 1991 and 1996 and the organizer and technical program chairman of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications, Boston, MA, 1992 and 1998. He has also been selected as a member of the Committee on Evolution of Untethered Communication, US National Research Council, 1997 and has led the US review team for the Finnish R&D Programs in Electronic and Telecommunication in 1999 and NETs project in 2003. For his contributions to the wireless networks he was the Westin Hadden Professor of Electrical and Computer Engineering at WPI during 1993–1996, was elected as a fellow of the IEEE in 1996 and became a fellow of Nokia in 1999. From May to December of 2000 he was the first Fulbright–Nokia scholar at the University of Oulu, Finland. Because of his inspiring visionary publications and his international conference activities for the growth of the wireless LAN industry, he is referred to as one of the founding fathers of the wireless LAN industry. In the past few years his research work has been the core for more than 25 patents by Skyhook Wireless, where he acts as the chief technical advisor. In January 2008 Steve Jobs announced that Skyhook Wireless's WiFi localization technology is used in iPhone. Details of his contributions to this field are available at www.cwins.wpi.edu.

Prashant Krishnamurthy is an associate professor with the graduate program in Telecommunications and Networking at the University of Pittsburgh. At Pitt, he regularly teaches courses on cryptography, network security, and wireless communications and networks. His research interests are wireless network security, wireless data networks, and position location in indoor wireless networks. He is the coauthor of the books *Principles of Wireless Networks – A Unified Approach* and *Physical Layer of Communication Systems* and is a co-editor of *Information Assurance: Dependability and Security in Networked Systems*. He served as the chair of the IEEE Communications Society Pittsburgh Chapter from 2000 to 2005. He obtained his PhD from Worcester Polytechnic Institute, Worcester, MA, in 1999.

PREFACE

Information networking has emerged as a multidisciplinary diversified area of research over the past few decades. From traditional wired telephony to cellular voice telephony and from wired access to wireless access to the Internet, information networks have profoundly impacted our lifestyle. At the time of writing, over 3 billion people are subscribed to cellular services and close to a billion residences have Internet connections. More recently, the popularity of smartphones enabling the fusion of computers, networking, and navigation for location-aware multimedia mobile networking has opened a new way of attachment between the human being and information networking gadgets. In response to this growth, universities and other educational institutions have to prepare their students in understanding these technologies.

Information networking is a multidisciplinary technology. To understand this industry and its technology, we need to learn a number of disciplines and develop an intuitive feeling for how these disciplines interact with one another. To achieve this goal, we describe important networking standards, classify their underlying technologies in a logical manner, and give detailed examples of successful technologies. The selection of detailed technical material for teaching in such a large and multidisciplinary field is very challenging, because the emphasis of the technology shifts in time. In the 1970s and 1980s the emphasis of industry and, subsequently, the interest in teaching networks were primarily based on queuing techniques [Kle75, Sch87, Ber87] because, at that time, medium access control was playing an important role in differentiating local-area network (LAN) technologies such as Ethernet, token ring and token bus. At that time, researchers and educators were interested in understanding the random behavior of traffic in contention access on computer resources and performance issues such as throughput and delay. The next generation of textbooks in the 1990s was around details of protocols used in the seven-layer ISO model, and they were written mostly by professors of computer science [Tan03, Pet07, Kur01]. During this period, authors with an electrical engineering education would introduce similar material with more emphasis on physical channels [e.g. Sta00]. These books described the Internet and asynchronous transfer mode as examples for wide-area networks (WANs) and provided details of a variety of LAN technologies at different levels of depth. They lacked adequate details in describing the cellular and other wireless networks that have been the center of attention in recent years for innovative networking.

The success of wireless information networks in the 1990s was a motivation behind another series of textbooks describing wide- and local-area wireless networks [Pah95, Goo97, Wal99, Rap03, Pah02]. The technical focus of these books was on describing wide-area cellular networks and local wireless networks. These books were written by professors of electrical engineering and computer engineering with different levels of emphasis on detailed descriptions of the lower layers issues and system engineering aspects describing details of implementation of wireless networks. These books do not cover the description of local wired technologies such as Ethernet or details of the implementation of bridges, switches, and routers which are used to form the Internet.

The emergence of wireless access and the dominance of the Ethernet in LAN technologies have shifted the innovations in networking towards the physical layer and characteristics of the medium. Currently, there is no single textbook that integrates all the aspects of current popular information networks together and places an emphasis on the details of physical layer aspects. In this book we pay attention to the physical layer while we provide fundamentals of information networking technologies which are used in wired and wireless networks designed for local- and wide-area operations. The book provides a comprehensive treatment of the wired Ethernet and Internet, as well as of cellular networks, wireless LANs (WLANs), and wireless personal area network (WPAN) technologies. The novelty of the book is that it places emphasis on physical layer issues related to formation and transmission of packets in a variety of networks. The structure and sequence of material for this book was first formed in a lecture series by the principal author at the graduate school of the Worcester Polytechnic Institute (WPI), Worcester, MA, entitled "Introduction to WANs and LANs." The principal author also taught shorter versions of the course at the University of Oulu, Finland, focused on the wired LANs and WLANs. The co-author of the book has taught material from this book at the University of Pittsburgh in several courses, such as "Mobile data networks," "Network security," and "Foundations of wireless communications." These courses were taught for students with electrical engineering, computer science, information science, and networking backgrounds, both from academia and industry.

We have organized the book as follows: we start with an overview of telecommunications followed by four parts each including several chapters. Part I contains Chapters 2–5 and explains the principles of design and analysis of information networks at the lowest layers. In particular, this part is devoted to the characteristics of the transmission media, applied transmission and coding, and medium access control. Part II and Part III are devoted to detailed descriptions of important WANs and LANs respectively. Part II describes the Internet and cellular networks and Part III covers popular wired LANs and WLANs, as well as WPAN technologies. Part IV describes security, localization, and sensor networking as other important aspects of information networks that have been important topics for fundamental research in recent years. The partitioned structure of the book allows flexibility in teaching the material. We believe that the most difficult part of the book for the students is Chapters 2–5, which provide a summary through mathematical descriptions of numerous technologies. Parts II and III of the book appear mathematically simpler but carry more details of how systems work. To make the difficult parts simpler for the students, an instructor can mix these topics as appropriate. For example, the lead author teaches similar material in one of his undergraduate courses in wireless networking by first introducing the channel behavior (Chapter 2), then describing assigned access methods (Chapter 5) before describing time-division multiple access cellular networks (Chapter 7). Then he introduces spread-spectrum coding techniques (Chapters 3 and 4) and code-division multiple access cellular networks (Chapter 7), and finally covers multidimensional constellations (Chapter 3) before discussing WLANs (Chapter 9). In fact, we believe that this is an effective approach for enabling the understanding of the fundamental concepts in students. Therefore, depending on the selection of the material, depth of the coverage, and background of the students, this book can be used for senior undergraduate or first- or second-year graduate courses in computer science, telecommunications, electrical and computer engineering, or electrical engineering departments as one course or a sequence of two courses.

The idea of writing this book and the first table of contents was developed between the lead author and Professor Mika Ylianttila of the University of Oulu, Finland, during the late summer of 2006 in Helsinki and Oulu, Finland, with some input from Dr. Sassan Iraji of Nokia. The basic idea was that the most interesting parts of the current networking issues evolve around physical layer aspects of wireless networks. Therefore, it is a good idea to write a new book describing the fundamentals of networks with emphasis on lower layers of communication protocols and technologies. To have a practical model, we started using the structure of the current authors' previous book, *Principles of Wireless Network – A Unified Approach*, and expanded that to include physical layer aspects of the wired medium as it is applied to the Ethernet and the Internet. Initially, we had the entire modulation and coding aspects of transmission in one chapter. Dr. Mohammad Heidari of the Center for Wireless Information Network Studies at WPI helped us to extract the coding parts from that chapter and prepare a first draft of a new chapter on coding based on class notes of the lead author for a similar lecture at the LANs and WANs course in WPI for which he had served as a teaching assistant. Dr. Heidari also committed to prepare the solutions for the book. The authors thank Dr. Heidari for his contribution in preparing Chapter 4 and editing Chapter 3 of the book and Dr. Ylianttila for his help in preparing the original proposal. In addition, the authors would like to express their appreciation to Dr. Allen Levesque, for his contributions in other books with the lead author which has indirectly impacted the formation of thoughts and the details of material presented in this book. The authors also acknowledge the indirect help of Professor Jacques Beneat, who prepared the solution manual of our other book, *Principles of Wireless Networks – A Unified Approach*. A significant number of those problems, and hence their solutions, are used in this book. The lead author also thanks Dr. Siamak Ayandeh and Brian Sylvester for their lectures in his classes on the Internet and Ethernet which has affected formation of material in Chapters 6 and 8 of the book, Ferit Akgul for his careful editing of the example and problems in the manuscript, and students of his Introduction to WANs and LANs course in Spring 2009 at WPI for their editorial comments on the first draft of the book. The second author expresses his gratitude to Dr. Richard Thompson, Dr. David Tipper, Dr. Martin Weiss, Dr. Sujata Banerjee, Dr. Taieb Znati, and Dr. Joseph Kabara of the Graduate Program in Telecommunications and Networking at Pitt. He has learnt a lot and obtained different perspectives on networking through his interaction and association with them. Similarly, we would like to express our appreciation to all graduates and affiliates of the CWINS laboratory at WPI and many graduates from the Telecommunications Program at Pitt whose work and interaction with the authors have directly or indirectly impacted on material presented in this book.

We have not directly referenced our referral to several resources on the Internet, notably Wikipedia. While there are people that question the accuracy of online resources, they have provided us with quick pointers to information, parameters, acronyms, and other useful references which helped us to build up a more comprehensive and up-to-date coverage of standards and technologies. We do acknowledge the benefits and usage of these resources.

In particular, we have used numerous articles in Wikipedia in many ways, as a quick check of facts and for links to references in many chapters of the book. We believe that such resources also benefit students tremendously.

The authors also would like to thank Katharine Unwin, Sarah Tilley, and Mark Hammond for the careful review of the manuscript and their useful comments, and for careful proof reading of the manuscript.

1

INTRODUCTION TO INFORMATION NETWORKS

1.1 INTRODUCTION

In the eyes of an engineer, the complexity of an industry relates to the challenges in implementation, size of the market for the industry, and the impact of the technology in this industry on human life. Everyday we use information networks and information technology more than any other technology – for social networking in both professional and personal aspects of our lives. The heart and the enabler of this technology is the *information networking industry*, which brings us mobile and fixed telephone services and connects us to the Internet in the home, office, and on the road, wherever we are. Although the infrastructure of the information networking industry is not seen by the public because it is mostly buried under the ground or it is propagating in the air invisibly, it is the most complex technology to implement, it owns the largest market size by far among all

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

industries, and it has enabled us to change our lifestyle to the extent that we often refer to our era as "the age of information technology."

To have an intuitive understanding of the size of the information networking industry it is good to know that the size of the budget of AT&T in the early 1980s, before its divestiture, was close to the budget of the fifth largest economy of the world. AT&T was the largest public switched telephone network (PSTN) company in the world and its core revenue at that time was generated from the plain old telephone service (POTS) that was first introduced in 1867. During the past two decades, the cellular telephone industry has augmented the income of the prosperous voice-oriented POTS with subscriber fees from more than 3 billion cellular telephone users worldwide. Today, the income of the wireless industry dominates the income of the wired telephone industry. While this income is still by far dominated by the revenue of the cellular phones, smart mobile terminals are gradually changing this characteristic by bringing more Internet-oriented applications into the terminal. In the mid 1990s, the Internet brought the computer and data communications and networking industry from a relatively smaller business-oriented office industry to an "everyday and everybody" use home-oriented industry that soon generated an income comparable to that of a voice-oriented POTS and the wireless industry. At the time of writing, the revenue of the information networking *industry* is dominated by the combined income of the wired and wireless telephone services over the PSTN and the Internet access industries, with total annual revenue of a few trillion dollars, which makes it one of the largest industries in the world.

At the start of the year 2008, the number of mobile phone subscriptions in the world had already passed 3 billion, with a worldwide average penetration rate of close to 50%, making the mobile phone the most widespread and adopted technology in the world. At the same time, close to a billion people around the world have access to the Internet or equivalent mobile Internet services using a mobile phone, a laptop, or a personal computer. In June 2007, Apple lunched the iPhone, which opened a new dimension in the integration of computer, communication, and navigation devices in a mobile handheld terminal. Around 4 million devices were sold during the first 200 days [MOB08], and in the first quarter of business it turned in to the third best-selling smartphone in the world [ARS08]. The third-generation (3G) iPhone introduced in July 2008 sold 1 million in the first 3 days [WAC08]. The iPhone provides built-in mobile user-friendly applications for video streaming, audio storage, and localization, as well as a number of popular applications, such as e-mail access, stock market reports, weather condition reports, calendar, and notebooks, on top of the traditional mobile phone and short messaging. This range of applications uses location-aware and secure broadband wireless access technologies provided through cellular networks, WiFi wireless local area networks (WLANs), and Bluetooth-based wireless personal-area networks (WPANs) to connect to the Internet and the PSTN backbones.

The authors believe that the information network technologies which enable smartphones to integrate traditional communication and Internet applications provide a suitable framework for teaching a basic course to introduce the fundamentals of modern information networks to students. The objective of this book is to provide a text for teaching these fundamentals at a senior undergraduate or first-year graduate-level course. In the rest of Section 1.1 we describe the elements of the information networks, a brief history of major events since the inception of the industry, a summary of the important standards, and a short description of long-haul standards to interconnect networks worldwide. The rest of this chapter provides a more detailed overview of evolution of important wide-area networks

(WANs) and local-area networks (LANs) followed by a brief description of the material presented in the rest of this book.

1.1.1 Elements of Information Networks

Figure 1.1 illustrates the fundamental elements of information networking. A network infrastructure interconnects user applications through telecommunication devices using network adaptors to provide them with means for exchanging information. Users are human beings, computers, or "things" such as a light bulb. Applications could be a simple telephone call, sending a short message, downloading a file, or listening to audio or video streams. The telecommunication devices range from a simple dumb terminal only translating user's message to an electrical signal used for communication, up to a smart terminal enabling multiple applications through a number of networking options. The network adaptor could be a connector to a pair of wires for a plain old telephone, a cellular phone, or a wireless local network chip set for a personal computer, a laptop or a smartphone, a low-power personal communication chip set for a light bulb, or a reader for a smart card. The information network infrastructure consists of a number of interconnecting elements that are connected primarily through point-to-point links. Switches include fixed and variable-rate connection-based circuit switches or connectionless packet-switching routers. The point-to-point links include a variety of fibers, coaxial cables, twisted-pair wires, and wireless technologies.

Figure 1.2 shows a view of the evolution of telecommunication devices and the networking concepts which have allowed these devices to interconnect with one another. The first communication device was the Morse pad invented for telegraph application in the nineteenth century to transfer coded short messages using Morse code. This was followed by the telephones devices used for voice communications. Both terminals were dumb and used for human-to-human communications. Shortly after the Second World War, the first dumb computer terminals were networked to start the era of computer-to-computer and human-to-computer networking which finally emerged into the Internet. Simplicity, flexibility, and lower cost of implementation of Internet technology opened a new frontier for the emergence of numerous popular applications and computer networking devices.

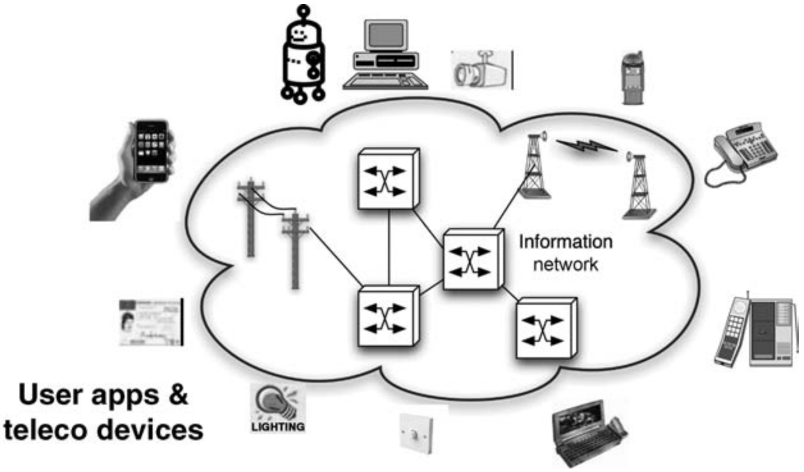


FIGURE 1.1 Elements of information networks.

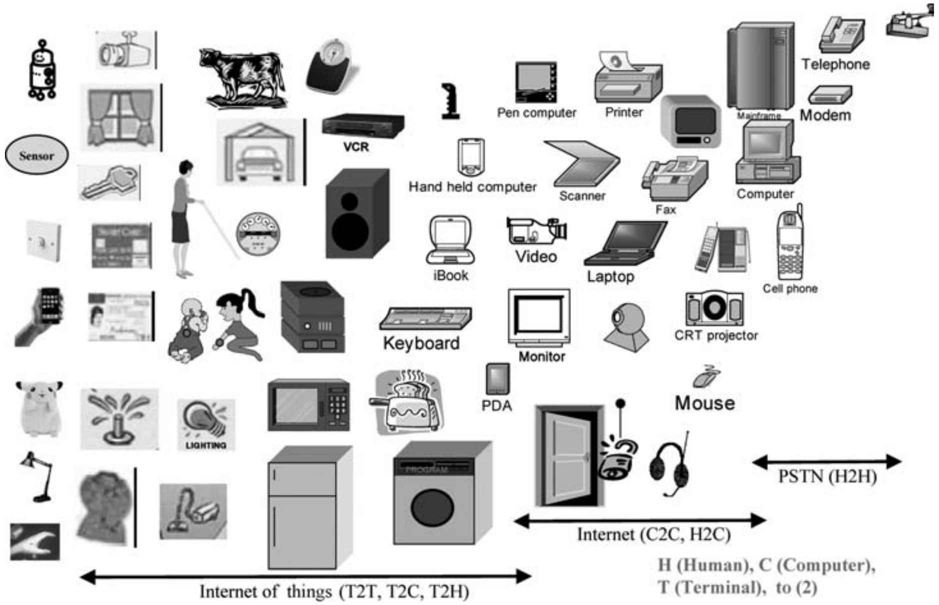


FIGURE 1.2 Evolution of telecommunication devices and information networks.

More recently, with the introduction of mobile wireless access to the PSTN and the Internet, a new window of opportunity for connecting "things" with the Internet and the so-called the "Internet of Things" has become popular. The Internet of Things allows things-to-human, things-to-computer and things-to-things networking, turning virtually everything to a communication device. Since these new devices have different data rates, power consumption, and cost requirements, a number of networking technologies have evolved in their support.

To support the evolution of telecommunication devices and applications, several information network infrastructures have evolved throughout the past 150 years. The largest of the existing backbone information networks are the PSTN, the Internet and the hybrid fiber cable (HFC). The PSTN was designed for the telephony application and it is also the backbone for popular cellular telephone networks. Cell phones are connecting to the PSTN using different cellular technologies. The Internet supports all computer communication data applications, and HFC was designed for cable TV distribution. To connect to the backbone networks, terminals are usually clustered in a local area to form a local network and the local network is connected to the backbone using another technology, which we can refer to as the access network. In this way, networking technologies are divided into core or backbone, access, and distribution or local networks.

In the PSTN industry, the local network is a private box exchange (PBX) used in office environments. It is a private switch allowing internal telephones of a company to talk with one another without the intervention of the core network. In the home, the local network for PSTN is the random tree wiring distribution in a residence which connects all rooms to the main line connected to the PSTN. Cordless telephones allow portable phones to connect to the home network. Cellular telephone companies add smaller base stations (BSs) called picocells inside large buildings, such as airports and shopping malls, and they are designing femtocell technologies for home applications to help cellular technologies to penetrate to indoor areas. In the case of the Internet, the local distribution network is a LAN. The Ethernet technology is dominating the wired LAN technology in offices, and WLANs complement them for mobile wireless access. In homes, the most popular Internet distribution network is the WLAN.

An access network is something between a WAN and a LAN. The infrastructure for the access network is a twisted-pair wire, a coaxial cable, a fiber line, a wireless terrestrial, or a satellite connection which connects the office or home to the backbone network. In the PSTN industry, this access network for the home is sometimes referred to as the last mile network. In the Internet industry it is sometimes referred to as metropolitan area networks (MANs). Home network access technologies are very important for the service providers because the high cost of the wiring to the home needs to be recuperated through the income of a single private subscriber.

1.1.2 Chronology of Information Networks

In the same way that the Greek philosophers of antiquity addressed the basic challenges in philosophy which laid the foundation for modern civilization, the emergence of the telegraph and the telephone in the early days of the information networking industry addressed the basic challenges facing information networking which have been carried on until modern times. Connection-based telephone conversations versus datagram-based connectionless messaging, digital versus analog, local versus wide area networking, wired versus wireless communications, and home versus office markets were all introduced in those early days. Over a century, engineers have discovered a number of technologies to enable these two services to adapt to the evolution of the terminals and to support ubiquitous operation.

To understand the sequence of events resulting in the evolution of information networks, it is useful to have a quick overview of the chronology of the events, which is presented in Table 1.1. Five years after Gauss and Weber's experiment to introduce wired telegraph for manually digitized data in 1834,[1] information networking started with the simple wired telegraph that used Morse code for digital data communication over long-distance wires between the two neighboring cities of Washington DC and Baltimore in 1839. It took 27 years, until 1866, for engineers to successfully extend this network over the ocean to make it a worldwide service. In 1900, 34 years after the challenging task of deploying cables in the ocean and 3 years after the first trial of the wireless telegraph, Marconi demonstrated *wireless* transoceanic telegraphy as the first wireless data application. It took over 150 years for this industry to grow into the wireless Internet. Bell started the telephone industry in 1867,[2] the first wired analog voice telecommunication service. It took 47 years for the telephone to become a transoceanic service in 1915, and it took almost 100 years for this industry to flourish into the wireless cellular telephone networks in the 1990s. The wireless telegraph was a point-to-point solution that eliminated the tedious task of laying very long wires in harsh environments. The telegraph was indeed a manual short messaging system (SMS) that

[1] Although focused on civil engineering, Rensselaer Polytechnic Institute (RPI), Troy, NY, the first engineering school in the USA, was established in 1824 as well. These events are indicators of the start of the industrial revolution and dominance of the engineers in shaping the future of the world.

[2] In the 1860s, in the dusts of the closing of the Civil War in the USA, a new wave of engineering and science schools mushroomed in the USA, starting with MIT (1862) in Boston and Worcester Polytechnic Institute (WPI; 1865) in Worcester, MA.

TABLE 1.1 A Brief History of Telecommunications

	Chronology of information networks
1839	First demonstration of telegraph between DC and Baltimore (Morse)
1876	Manually switched telephone for analog voice (Bell)
1900	Transoceanic wireless telegraph (Marconi)
1915	Transcontinental telephone (by Bell)
1946	First computer (U Penn)
1950	Voice-band modems for first computer networks using PSTN infrastructure
1968	Cable TV development and introduction of HFC
1969	ARPANET packet-switched network started (first node at UCLA)
1972	Demonstration of cellular systems (Motorola)
1973	Ethernet was invented (Metcalfe)
1980	IPv4 was released, fiber-optic systems were applied to the PSTN
1981	IEEE 802.3 adopted Ethernet
1986	IETF was formed
1990	GSM standard for TDMA cellular
1991	ATM Forum was founded
1994	Netscape was introduced and Internet became popular
1995	IS-95 standard for CDMA cellular, fast Ethernet at 100 Mb/s, first ATM specification
1996	IPv6 was defined by IETF
1997	IEEE 802.11 completed
1998	IEEE 802.1D MAC bridges and STA, gigabit Ethernet, 802.16 WMAN, IEEE 802.15.1 Bluetooth for WPAN
1999	IEEE 802.11b at 11 Mb/s
2000	3G IMT-2000 for wireless Internet access was introduced
2001	IEEE 802.11a for 54 Mb/s at 5 GHz using OFDM
2002	Mobile IP standard completed, UWB was used for high-speed WPAN
2003	IEEE 802.1Q virtual LAN, IEEE 802.11g using OFDM at 2.4 GHz, IEEE 802.15.4 ZigBee, 10 Gb/s Ethernet
2004	IEEE 802.1w rapid spanning tree algorithm for switches, IEEE 802.11d WiMAX for WMAN
2005	IEEE 802.11e mobile WiMAX
2006	RFID, sensor networks, “Internet of Things”
2007	IEEE 802.11n for 100 Mb/s
2008	iPhone, 100 Gb/s Ethernet

needed a skilled worker to decode the transmitted message. The wireless telephone network had to support numerous mobile users. While the challenge for wireless point-to-point communications is the design of the radio, the challenge for a wireless network is the design of a *system* that allows many mobile radios to work together.

The computer era started with the demonstration of the first digital computer at the University of Pennsylvania in 1946. Computers have revolutionized the traditional methods for information storage and processing that were in use since the dawn of literacy some 3000 years ago. The massive information stored and processed by computers needed communication networking. The first computer communication networks started after the Second World War by using voice-band modems operating over the PSTN infrastructure to exchange large amounts of data among computers located at great distances from one another. The need of the computer communication industry for massive information transfer and the high cost of leased lines provided by PSTN service providers to be used for voice-band data communications stimulated the evolution of sophisticated modem design technologies. By the late 1980s a number of modem design techniques were introduced which could achieve higher data rates over a fixed bandwidth channel [Pah88]]. These technologies laid the foundation for the design of the different physical layers that have been at the core of advancements in evolution of broadband access using cable modems and digital subscriber line (DSL), as well as LAN, WLAN, and WPAN technologies, in the 1990s and 2000s. In parallel with the growth of information theory to support higher data transmission rates, networking protocols emerged first for reliable transmission of data and later to facilitate the implementation of ever-growing computer applications.

About two decades after the emergence of circuit-switched computer communication networks, in 1969, the first wide-area packet-switched network called *Defense Advanced Research Projects Agency* Department Network (DARPAnet) was introduced, which later on became the Internet and gained popularity in the mid 1990s. A few years after the introduction of DARPAnet, in 1973, the first wired LAN, Ethernet, was invented, which dominated the LAN industry again in the mid 1990s. The Internet/Ethernet network core has dominated the networking industry, and numerous other technologies and applications have emerged around them. These technologies include a variety of Institute of Electrical and Electronics Engineers (IEEE) 802.11 WLANs, a number of IEEE 802.15 WPANs, several IEEE 802.16 wireless MANs (WMANs), a few IEEE 802.1 bridging technologies, and a number of transport control protocol (TCP)- and Internet protocol (IP)-based protocols defined by the Internet Engineering Task Force (IETF) which are highlighted in Table 1.1.

1.1.3 Standards Organizations for Information Networking

The increasing number of applications, and the information networking technologies to support them, has demanded standardization to facilitate the growth of the industry. Standards define specifications for the design of networks allowing multi-vendor operation which is essential for the growth of the industry. Figure 1.3 provides an overview of the standardization process in information networking. The standardization process starts in a special interest group of a standard developing body such as the IETF or IEEE 802.3, which defines the technical details of a networking technology as a standard for operation. The defined standard for implementation of the desired network is then moved for approval by a regional organization, such as the European Telecommunication Standards Institute

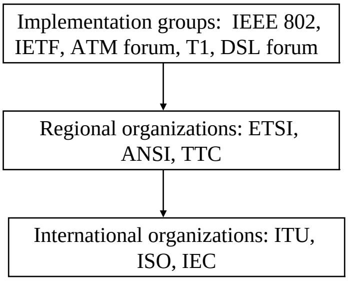

Implementation groups: IEEE 802, IETF, ATM forum, T1, DSL forum
Regional organizations: ETSI, ANSI, TTC
International organizations: ITU, ISO, IEC

FIGURE 1.3 Standard development process.

TABLE 1.2 Summary of Important Standard Organizations for Information Networking

Important Standards Organizations	
IEEE (Institute of Electrical and Electronics Engineers)	Publishes 802 series standards for LANs and 802.11, 802.15, and 802.16 for wireless applications
T1	Sponsored by Alliance for Telecommunications Information Solutions (ATIS) telecommunication standards body working on North American standards
ATM (Asynchronous Transfer Mode) Forum	An industrial group working on a standard for ATM networks
DSL (Digital Subscriber Loop) Forum	An industrial group working on xDSL services
CableLab	Industrial alliance in the USA to certify DOCSIS-compatible cable modems
IETF (Internet Engineering Task Force)	Publishes Internet standards that include TCP/IP and SNMP. It is not an accredited standards organization
FCC (Federal Communication Commission)	The frequency administration authority in the USA.
EIA/TIA (Electronic/ Telecommunication Industry Association)	US national standard for North American wireless systems
ANSI (American National Standards Institute)	Accepted 802 series and forwarded to ISO. Also published FDDI, HIPPI, SCSI, and Fiber Channel. Developed JTC models for wireless channels
ETSI (European Telecommunication Standards Institute)	Published GSM, HIPERLAN/1, and UMTS
CEPT (Committee of the European Post and Telecommunication)	Standardization body of the European Posts Telegraph and Telephone (PTT) ministries. Co-published GSM with ETSI
IEC (International Electrotechnical Commission)	Publishes jointly with ISO
ISO (International Standards Organization)	Ultimate international authority for approval of standards
ITU (International Telecommunication Union formerly CCITT)	International advisory committee under United Nations. The Telecommunication Sector, UTU-T, published ISDN and wide-area ATM standards. Also works on IMT-2000

(ETSI) or the American National Standards Institute (ANSI). The regional recommendation is finally submitted to world-level organizations, such as the International Telecommunications Union (ITU), International Standards Organization (ISO), or International Electrotechnical Commission (IEC), for final approval as an international standard. There are a number of standards organizations involved in information networking. Table 1.2 provides a summary of the important standards playing major roles in shaping the information networking industry which are also mentioned in this book.

The most important of the standard development organizations for technologies described in this book produces the IEEE 802-series standards for personal, local, and metropolitan area networking. The IEEE, the largest engineering organization in the world, publishes a number of technical journals and magazines and organizes numerous conferences worldwide. The IEEE 802 community is involved in defining standard specifications for information networks. The number 802 was simply the next free number that the IEEE could assign to a committee at the inception of the group on February 1980, although "80-2" is sometimes associated with the date of the first meeting. Regardless of the ambiguity of the name, the IEEE 802 community has played a major role in the evolution of information networks by introducing IEEE 802.3 Ethernet, IEEE 802.11 WLANs, IEEE 802.15 WPAN, and IEEE 802.16 WMAN and other standards which are discussed in detail in this book.

Another important standard development organization is the IETF, which was established in January 1986 to develop and promote Internet standard protocols around the TCP/IP suite for a variety of popular applications. In the 1990s, the Asynchronous Transfer Mode (ATM) Forum was an important standard development group trying to develop standards for connection-based fixed packet-length communications for the integration of all services. This philosophy was in contrast with Internet/Ethernet networking using connectionless communications with variable and long-length packets.

Telecommunication/Electronic Industry Association (TIA/EIA) is a US national standards body defining a variety of wire specifications used in LANs, MANs, and WANs. The TIA/EIA are trade associations in the USA representing several hundred telecommunications companies. The TIA/EIA has cooperated with the IEEE 802 community to define the media for most of the wired LANs used in fast and gigabit Ethernet. TIA/EIA also defines cellular telephone standards such as Interim Standard (IS-95) or cdmaOne and the IS-2000 or CDMA-2000 (respectively the second and third generation) cellular networks. ETSI and the Committee of the European Post and Telecommunications (CEPT) were the European standardization bodies publishing wireless networking standards, such as the global system for mobile communications (GSM) and the universal mobile telecommunications system (UMTS), in the EU.

The most important international standards organizations are the ITU, the ISO, and the IEC, which are all based in Geneva, Switzerland. Established in 1865, the ITU is an international advisory committee under the United Nations and its main charter includes telecommunication standardization and allocation of the radio spectrum. The telecommunication sector, ITU-T, has published for instance the integrated service data network (ISDN) and wide-area ATM standards, as well as International Mobile Telephone (IMT-2000) for 3G cellular networks. The World Administrative Radio Conference (WARC) was a technical conference of the ITU where delegates from member nations of the ITU met to revise or amend the entire international radio regulations pertaining to all telecommunication services throughout the world. The ISO and IEC are composed of the national standards bodies, one per member economy. These two standards often work with one another as the ultimate world standard organization. Established in 1947, the ISO nurtures worldwide proprietary industrial and commercial standards that often become law, either through treaties or national standards. The ISO seven-layer model for computer networking is one of the prominent examples of ISO standards. IEC, which started in 1906, is a nongovernmental international standards organization for "electrotechnology," which includes a vast number of standards from power generation, transmission and distribution to home appliances and office equipment, to telecommunication standards. The IEC publishes standards with the IEEE and develops standards jointly with the ISO and the ITU.

1.1.4 Evolution of Long-Haul Multiplexing Standards

In telecommunications and computer networks, multiplexing is used where multiple data streams are combined into one signal to share the expensive long-haul transmission resources. From a different perspective, multiplexing divides the physical capacity of the transmission medium into several logical channels, each carrying a data stream. The two most basic forms of multiplexing over point-to-point connections are time-division multiplexing (TDM) and frequency-division multiplexing (FDM). In optical communications, FDM is referred to as wavelength-division multiplexing (WDM). Multiple streams of digital data can also use code-division multiplexing (CDM), which has not been commercially successful over wired networks. Variable bit-rate digital bit streams may be transferred efficiently over a fixed-bandwidth channel similar to TDM by means of statistical multiplexing techniques such as ATM. If multiplexing is used for channel access then it is referred to as medium access control (MAC), in which case TDM becomes time-division multiple access (TDMA), FDM becomes frequency-division multiple access (FDMA), CDM becomes code-division multiple access (CDMA), and statistical multiplexing into something like carrier sense multiple access (CSMA). Multiplexing techniques are simple and they are part of the physical layer of the network. We discuss them in this section. MAC protocols are more complex and we describe them in Chapter 5.

In PSTN, the home user's telephone line carrying telephone or DSL data typically ends at the remote concentrator boxes distributed in the streets, where these lines are multiplexed and carried to the central switching office on significantly fewer numbers of wires and for much further distances than a customer's line can practically go. Fiber multiplexing lines, mostly using ATM protocols, are commonly used as the backbone of the network which connects POTS lines with the rest of the PSTN and carry data provided by DSLs. As a result, PSTN has been the driver for the development and standardization of most multiplexing techniques. In the early 1960s, the first multiplexing system used for telephony was FDM. Figure 1.4 illustrates this early multiplexing system. This system multiplexed 12 subscribers, each with a 4 kHz bandwidth signal in frequencies between 60 and 108 kHz. Figure 1.4 shows the original bandwidths for each subscriber, the bandwidths raised in frequency, and the multiplexed channel. In PSTN connection to the home, as shown in Figure 1.5, FDM is also used to multiplex DSL data services with POTS and the so called Home Phone Network Alliance (HPNA) signal for Ethernet-like networking over home telephone wires. The POTSs use the frequency range between 20 Hz and 3.4 kHz, DSL uses

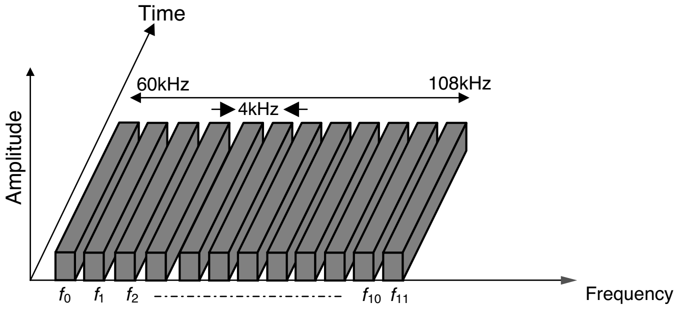


FIGURE 1.4 FDM used in telephone networks multiplexing 12 subscribers each with 4 kHz bandwidth in frequencies between 60–108 kHz.

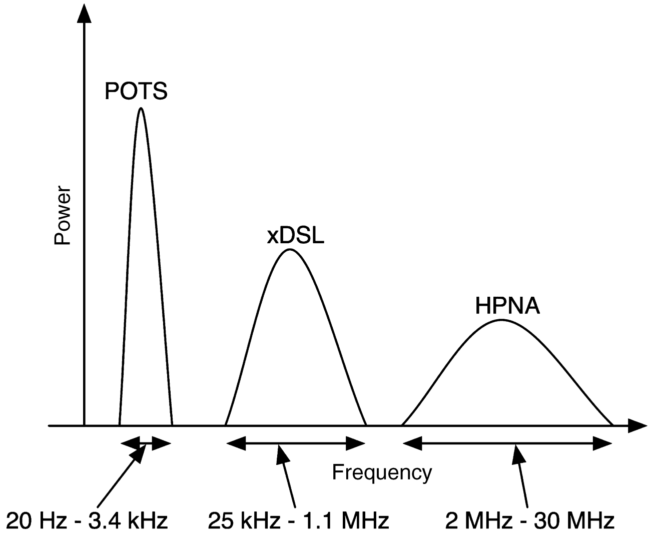


FIGURE 1.5 Phone line wirings shared among three technologies using FDM.

between 25 kHz and 1.1 MHz, and HPNA uses between 2 and 30 MHz. The TV signals for black and white content, color, and audio are also frequency-division multiplexed. The analog cable TV channels are also separated using FDM.

The first TDM system was developed for telephony application as well. This system carried 24 PCM-encoded digitized voice calls, each 64 kb/s, over four-wire copper lines at a rate of 1.544 Mb/s. Figure 1.6 shows the 193-bit frame used in a T1-carrier: there are 8 bits per channel for 24 × 8 = 192 bits and a single bit known as the frame bit. Each 8 bits consists of 7 bits representing a sample of the voice and 1 bit for control signaling. Each frame is transmitted in 125 ms to support a bit rate of 8 (bits)/125 (ms) = 64 kb/s per channel and an effective data rate of 7 (bits)/125 (ms) = 56 kb/s. The remaining 8 kb/s is used to carry the signaling information to set up telephone calls. A T1-carrier has a higher layer hierarchy to support higher data rates. Figure 1.7 shows the T-carrier hierarchy: every four T1 carriers result in a T2 carrier with the rate of 6.312 Mb/s, each seven T2 carriers form a T3 carrier which carries 44.736 Mb/s, and six T3 carriers form a T4 carrier at 274.176 Mb/s. Although the Japanese followed the American carrier system, the Europeans later developed their

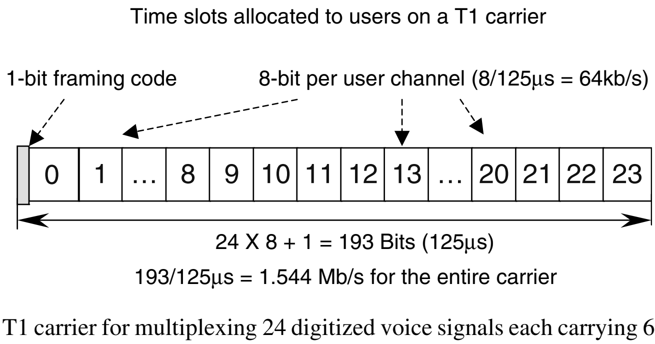

Time slots allocated to users on a T1 carrier
1-bit framing code
8-bit per user channel (8/125μs = 64kb/s)
0
1
...
8
9
10
11
12
13
...
20
21
22
23
24 X 8 + 1 = 193 Bits (125μs)
193/125μs = 1.544 Mb/s for the entire carrier

FIGURE 1.6 T1 carrier for multiplexing 24 digitized voice signals each carrying 64 kb/s for a total 1.544 Mb/s.

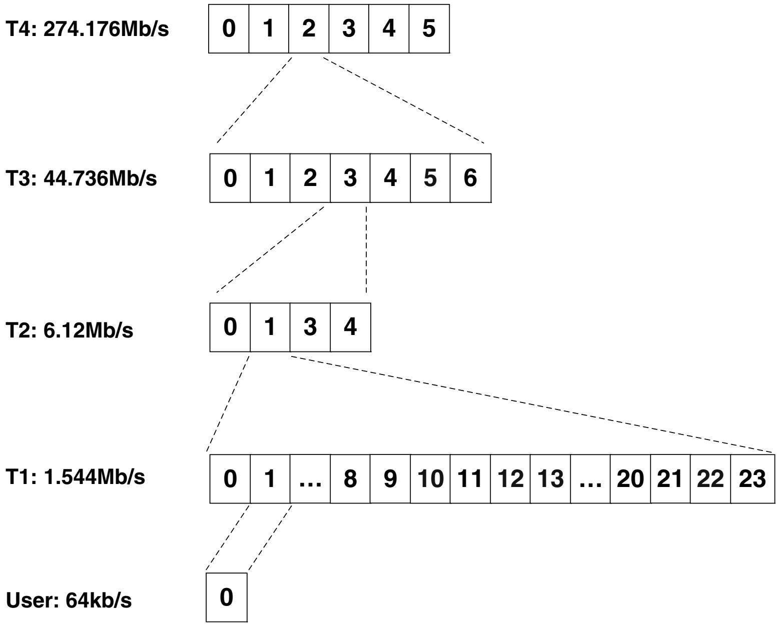


FIGURE 1.7 The legacy T-carrier hierarchy for long-haul TDM transmission of PSTN voice streams in the USA and Japan.

own TDM hierarchy for telephony which combines 30 channels rather than 24 channels to form the basic stream. The next three layers of hierarchy in the European system each are four times larger than the previous level.

With the popularity of optical communications in the 1980s and their ability to support much higher transmission rates, there was a need to extend the existing TDM hierarchies. At the same time, AT&T was broken up and there was a need to develop a standard TDM hierarchy so that the multivendor manufactured devices for different companies could cooperate with one another. In this setting, the synchronous optical network (SONET) hierarchy started in North America. Later on the ITU became involved and the name synchronous digital hierarchy (SDH) was selected as the name of the standard (which has only minor differences with SONET). The SONET/SDH hierarchy defines a common signaling standard for multivendor operation (for wavelength, timing, and frame structure), unifies the US, the EU, and Japanese standards, defines data rates higher than T-carriers, and provides for support of operation, administration, and maintenance (OAM) of the network. SONET/SDH can be used for encapsulation of earlier digital transmission standards and they can be directly used to support ATM or SONET/SDH packet mode of operation providing for generic and all-purpose transport containers for multimedia information.

In a manner similar to the T1 carrier, the overhead and the payload of the SONET streams are interleaved and all frames take 125 ms. Here, the basic frame, called the synchronous transport signal1 (STS-1)[3] or optical carrier1 (OC-1), is 810 octets in size and the frame is transmitted with three octets of delimiter indicating the start of the frame followed by nine 87 octets of payload each having three octets of overhead, followed by 84 octets of user data.

[3] This is the name for the electrical rather than optical signal.

TABLE 1.3 SONET/SDH Hierarchy of Data Rates and Frame Format

SONET optical	SONET electrical	SDH level electrical	Payload bandwidth (kb/s)	Line rate (kb/s)
OC-1	STS-1	STM-0	48 960	51 840
OC-3	STS-3	STM-1	150 336	155 520
OC-12	STS-12	STM-4	601 344	622 080
OC-24	STS-24	STM-8	1 202 688	1 244 160
OC-48	STS-48	STM-16	2 405 376	2 488 320
OC-96	STS-96	STM-32	4 810 752	4 976 640
OC-192	STS-192	STM-64	9 621 504	9 953 280
OC-768	STS-768	STM-256	38 486 016	39 813 120
OC-1536	STS-1536	STM-512	76 972 032	79 626 120
OC-3072	STS-3072	STM-1024	153 944 064	159 252 240

If we align all 87 rows then we have a block of 9 × 87 octets of payload in which the first three overhead columns appear as a contiguous block, as do the remaining 84 columns of user data. With this format, if we extract one octet from the bitstream every 125 ms duration of a frame, this gives a data rate of 8 (bits)/125 (ms) = 64 kb/s representing a telephone user. This is very useful for fast switching of low-speed streams embedded in extremely high data streams without any need to understand or decode the entire frame. The basic frame in SDH is called synchronous transport mode1 (STM-1) which has similar properties to OC-1/STS-1 with minor differences. These simple formats allow design of relatively simple devices to take a SDH data for a specific user and insert that in a SONET stream. This simple structure allows implementation of faster switches operating at higher data rates. Again, like the T1 hierarchy, SONET and SDH have their own TDM hierarchy, which is shown in Table 1.3.

1.2 EVOLUTION OF WIDE-AREA NETWORKS

The information sources carried through the information networks are divided into voice, data, and video. The most commonly used wired infrastructures to carry voice, data, and video are the PSTN, Internet, and HFC respectively. The dominant voice application is telephony that is a bidirectional symmetric real-time conversation. To support telephony, telephone service providers have developed a worldwide network infrastructure that establishes a connection for a telephone call during the dialing process and disconnects it after completion of the conversation. This network, commonly referred to as PSTN, is connection based using circuit switching and it has the capability of assuring a certain quality of service (QoS) to the user during the connection time. The core of the PSTN is a huge digital transmission system that allocates a 64 kb/s circuit for each direction of a telephone conversation. Other network providers, to interconnect their nodes, often lease the PSTN transmission facilities (e.g. making use of the long-haul standards described in the previous section). The Internet is a connectionless worldwide public data network (PDN) using packet switching and the standard IP. The Internet is a "network of networks" that consists of numerous academic, business, and government networks, which together carry various information and services, such as e-mail, web access, file transfer, and many other emerging applications. HFC is the infrastructure developed for video applications in cable TV. This network broadcasts wideband video signals to residential buildings. A cable goes from an end office to a residential area and all users tap their signals from the same cable. The set-top boxes leased by cable companies provide selectivity of channels depending on the rates charged. The end offices, where a group of distribution cables arrive, are connected to one another through a fiber line. For this reason, the cable TV network is also called HFC. Cable distribution is also now used for broadband home access to the Internet.

1.2.1 Evolution of the Public Switched Telephone Network

The invention of the telephone in 1876 was the start of analog telephone networking and the PSTN. In the early days, networks used operators to switch or route a session from one terminal to another manually. At the beginning of the twentieth century the telecommunications industry had already been exposed to a number important issues that played different roles during the entire course of the past century, culminating in the emergence of modern wireless networks. Among these important issues were analog versus digital, voice versus data, wireless versus wired, local versus long-haul communications, and personal versus group services. By the 1950s, the PSTN had more than 10 million customers in the USA, and those interested in long-haul communication issues also needed their services to solve their problems. Although end users are still mostly connected to the PSTN with twisted-pair analog lines, to provide flexibility and ease of maintenance and operation of the PSTN, the core network gradually changed to digital switches and digital wired lines connecting switches together. A hierarchy of digital lines (the T-carriers in the USA described previously) evolved as trunks to connect switches of different sizes together.

Another advancement in the PSTN was the development of private branch exchanges (PBXs) as privately owned local telephone networks for large offices. A PBX is a voice-oriented LAN owned by the end organization itself, rather than the telephone service provider. This small switch allows the telephone company to reduce the number of wires that are needed to connect all the lines in an office to the local office of the PSTN. In this way, the service provider reduces the number of wires to be laid to a small area where large offices with many subscribers are located. The end user also pays less to the telephone company. The organization thus has an opportunity to enhance services to the end users connected to the PBX.

In the 1920s, Bell Laboratories conducted studies to use the PSTN facilities for data communications. In this experiment, the possibility of using analog telephone lines for transferring transoceanic telegrams was examined. Researchers involved in this project discovered several key issues, including the sampling theorem and the effects of phase distortion on digital communications. However, these discoveries did not affect applications until after the Second World War, when Bell Laboratories developed voice-band modems for communication among air force computers in air bases that were geographically separated by large distances [Pah88]. These modems soon found their way into commercial airlines and banking industries, resulting in associated private long-haul data networks. These pioneering computer communications networks consisted of a central computer and a bank of modems operating over four-wire commercial-grade leased telephone lines to connect several terminals to the computer. In the late 1960s, the highest data rate for commercial modems was 4800 bits/s. By the early 1970s, with the invention of quadrature amplitude modulation (QAM), the data rate of four-wire voice-band modems reached 9600 bits/s. In the early 1980s, trellis-coded modulation (TCM) was invented, which increased data rates to 19.2 kb/s and beyond.

In parallel with the commercial four-wire modems used in early long haul computer networks, two-wire modems emerged for distance connection of computer terminals. The two-wire modems operated over standard two-wire telephone lines and they were equipped with dialing procedures to initiate a call and establish a POTS line during the session. These modems started at data rates of 300 bits/s. By the early 1970s, they had reached 1200 bits/s, and by the mid 1980s they were running at 9600 bits/s. These two-wire voice-band modems allowed users in the home and office to have access to regular telephones to develop a data link connection with a distant modem also having access to the PSTN. Voice-band modems using two-wire telephone connections soon found a large market in the residential and small office remote computer access (Telnet), and the technology soon spread to a number of popular applications, such as operating a facsimile machine or credit card verification device. With the popularity of Internet access, a new gold rush for higher speed modems began, resulting in 33.6 kb/s full-duplex modems in 1995 and 56 kb/s asymmetric modems by 1998. The 56 kb/s modems use dialing procedures and operate within the 4 kHz voice-band, but they connect directly to the core PCM digital network of the PSTN that is similar to DSLs. DSLs use the frequency band between 2.4 kHz and 1.1 MHz to support data rates up to 10 Mb/s over two-wire telephone lines.

More recently, cellular telephone services have evolved. To connect a cellular telephone to the PSTN, the cellular operators developed their own infrastructure to support mobility. This infrastructure was connected to the PSTN to allow mobile to fixed telephone conversations. Addition of new services to the PSTN demanded increases in the intelligence of the core network to support these services. As this intelligence advanced, the telephone service provider added value features such as voice mail, autodialing through network operators, call forwarding, and caller identification to the basic POTS traditionally supported.

1.2.2 Emergence of the Internet

Public data networks (PDNs) that evolved around voice-band modems connected a variety of applications in a semi-private manner. The core of the network was still the PSTN, but the application was for specific corporate use and was not offered privately to individual users. These networks were private data networks designed for specific applications and they did not have standard transport protocols to allow them to interconnect with one another. Another irony of this operation was that the digital data was first converted to analog to be transmitted over the telephone network and then, within the telephone network, it was again converted to digital format for transmission over long distances using the digital subcarrier system. To avoid this situation, starting in mid 1970s, telephone companies started to introduce digital data services (DDSs) that provided a 56 kb/s digital service directly delivered to the end user. The idea was great because, at that time, the maximum data rate for voice-band modems was 9600 bits/s. However, like many other good and new ideas in telecommunications, this idea did not become popular. A large amount of capital was already invested in the existing voice-band-based data networks and it was not practical to replace them at once and DDSs were not interoperable with the analog modems. The DDSs later emerged as ISDN services providing 2×64 kb/s voice channels and 1 × 16 kb/s data channel to individual users. Penetration rates of ISDN services were not as expected, but laid a foundation for digital cellular services. Digital cellular systems can be viewed as a sort of wireless ISDN technology that integrates basic digital voice with a number of data services at the terminal.

The major cost for operation of a computer network over the four-wire lines was the cost of leasing lines from the telephone company. To reduce the operation cost, multiplexers were used to connect several lower speed modems and carry all of them at once over a higher speed modem operating over a long-distance line. The next generation of multiplexers consisted of statistical multiplexers that multiplexed flows of data rather than multiplexing individual modem connections. Statistical multiplexing technology later evolved into router technology that is generalized packet data switching.

In the early 1970s, the rapid increase in the number of terminals in offices and manufacturing floors was the force behind the emergence of LANs. These LANs provided high-speed connections (greater than 1 Mb/s) among terminals, facilitating sharing printers or mainframes from different locations. LANs were providing a local medium specifically designed for data communications that was completely independent from the PSTN. By the mid 1980s, several successful LAN topologies and protocols were standardized and LANs were installed in most large offices and manufacturing floors connecting their computing facilities. However, the income of the data communication industry, both LANs and PDNs, was far below that of the PSTN, still leaving the PSTN as the dominant economical force in the information networking industry.

Another important and innovative event in 1970s was the implementation of ARPANET, the first packet-switched data network connecting 50 cities in the USA. This experimental network used routers rather than the PSTN switches to interconnect data terminals. The routers were originally connected via 56 kb/s digital leased lines from the telephone company and the ARPANET interconnected several universities and government computers around a large geographical area. This network was the first packet-switched network supporting end-to-end digital services. This basic network later on upgraded to higher speed lines and numerous additional networks. To facilitate a uniform communication protocol to interconnect these disparate networks, TCP/IP protocols evolved that allowed LANs and a number of other PDNs to interconnect with one another and form the Internet. In the mid 1990s, with the introduction of popular applications such as Telnet, file transfer protocol (FTP), e-mail, and web browsing, the Internet industry was created. Soon, the Internet penetrated the home market and the number of Internet users became comparable with that of the PSTN, creating another economical power, namely computer communications applications, that compete with the traditional PSTN. The IP-based Internet provides a cheaper solution than circuit-switched operations, and today people are thinking of employing IP to capture a large share of the traditional telephony market served by the PSTN. The Internet provides a much lower cost alternative to PSTN for support of multimedia applications. With the growth of the wireless industry in the past two decades (WLANs and wireless wide-area data networks), wireless access to the Internet has become very popular.

In a manner similar to cellular telephony, wireless data network infrastructures have evolved around the existing wired data network infrastructures. WLANs are designed mostly for in-building applications to cover a small area and the network has a minimal infrastructure. WLANs are usually connected to the existing wired LANs as an extension. Mobile data services are designed for lower speed wireless data applications with metropolitan, national, and global coverage.

1.2.3 HFC Infrastructure for Cable TV

Another competing wired infrastructure that evolved in the last few decades of the twentieth century was the cable TV network. Installation of cable TV distribution networks in the USA started in 1968 and has penetrated more than 60% of the residential homes in the USA. This penetration rate is getting close to that of the PSTN. The cable TV network consists of three basic elements: a regional hub, a distribution cable bus, and a fiber ring to connect the hubs to one another. Because of the hybrid usage of fiber and cable, this network is also referred to as an HFC network, as mentioned earlier. The signals containing all channels at the hub are distributed through the cable bus in a residential area and each home taps the signal off the bus. This is radically different from home access through twisted-pair wires provided by the PSTN in many ways. The bandwidth of the coaxial cable supports about 100 TV channels, each around 6 MHz, while the telephone basic channel is around 4 kHz. The extended telephone channel using DSL uses about 1 MHz of bandwidth. The cable access is via a long bus originally designed for one-way multicast that has a number of taps (up to 500) creating a less controllable medium. The twisted-pair star access for the PSTN is designed for two-way operation and is easier to control. The HFC channel is noisier than the telephone channel, and in spite of its wider bandwidth its current supported broadband data rates are of the same range as the xDSL services operating on telephone wirings.

The cable TV network was also considered as a backbone for cellular telephone systems and it is considered as the leading method for broadband home access to support evolving home networks. Some of the cable TV providers in the USA also offer telephone services over this medium. In the late 1990s, the success of cable in broadband access encouraged some of the PSTN providers, such as AT&T, to acquire cable companies, such as MediaOne. With the introduction of fiber to home services with integrated high-speed internet access, telephony, and TV services, the cable TV industry is experiencing challenges from traditional PSTN companies that are providing fiber to home services.

1.2.4 Evolution of Cellular Telephone Networks

The technology for the first-generation (1G) FDMA analog cellular systems was developed at AT&T Bell Laboratories in the early 1970s. However, the first deployment of these systems took place in the Nordic countries under the name Nordic Mobile Telephone (NMT) about a year earlier than the deployment of the Advanced Mobile Phone Services (AMPS) in the USA. Since the USA is a large country, the frequency administration process was slower and so it took a longer time for deployment. The second-generation (2G) digital cellular networks started in the Nordic countries with the formation of the GSM standardization group. The GSM standard group was originally formed to address international roaming, a serious problem for cellular operation in the EU countries. The standardization group quickly decided to go for a new digital TDMA technology because it could allow integration of other services to expand the horizon of wireless applications [HAU94]. In the USA, however, the reason for migration to digital cellular was that the capacity of the analog systems in major metropolitan areas such as New York City and Los Angles had reached its peak value and there was a need for increasing the capacity in the existing allocated bands. Although the Nordic countries, led by Finland, have always maintained the highest rate of cellular penetration, in the early days of this industry the USA was by far the largest market. By 1994, there were 41 million subscribers worldwide, 25 million of them in the USA. The need for higher capacity motivated the study of CDMA that was originally perceived to provide a capacity that was orders of magnitude higher than other alternatives, such as analog band splitting or digital TDMA.

While the debate between TDMA and CDMA for 2G was in progress in the USA, deployment of GSM technology started in the EU in the early 1990s. At the same time, developing countries started their planning for cellular telephone networks and most of them adopted the GSM digital cellular technology over legacy analog cellular systems. Soon afterwards, GSM had penetrated into more than 100 different countries. An interesting phenomenon in the evolution of the cellular telephone industry was the unexpected rapid expansion of this industry in developing countries. In these countries, the growth of the infrastructure for wired POTS was slower than the growth of the demand for the new subscriptions, and always there were long waiting times to acquire a telephone line. As a result, in most of these countries, telephone subscriptions were sold in the black market at a price several times their actual value. Penetration of the cellular telephone in these counties was much easier because people were already prepared for higher prices for telephone subscription. Also, deployment and expansion of cellular networks could be done much faster than POTS.

In the beginning of the race between TDMA and CDMA, CDMA technology was deployed only in a few countries. Moreover, experimentation had shown that the capacity improvement factor of CDMA was smaller than expected. In the mid 1990s, when the first deployments of CDMA technology started in the USA, most companies were subsidizing the cost to stay in race with the TDMA and analog alternatives. However, from day one, the quality of voice using CDMA was superior to that of TDMA systems installed in the USA. As a result, CDMA service providers under banners like "you cannot believe your ears" started marketing this technology in the USA that soon become very popular with the users. Meanwhile, with the huge success of digital cellular, all manufacturers worldwide started working on the 3G international mobile telephone (IMT-2000) wireless networks.

The purpose of migration to 3G networks was to develop an international standard that combines and gradually replaces 2G digital cellular. At the same time, 3G systems were expected to increase the quality of the voice, capacity of the network, and data rate of mobile data services. Among several radio transmission technology proposals submitted to the ITU, the dominant technology for 3G systems was wideband CDMA (W-CDMA) that is discussed in Chapter 7.

1.3 EVOLUTION OF LOCAL NETWORKS

Outside of the of the PSTN, Internet, HFC, and cellular WANs, the PBX, LAN, WLAN and WPAN technologies are used as local networks to distribute the network services inside an office or at home. The cost of infrastructure in WANs is very high and the coverage is very wide. As a result, WANs are offered as a charged *service* to the user. The service provider invests a large capital for the installation of the infrastructure and generates revenue through monthly service charges. Local networks are often sold as end products to the user and there is no service payment for local communications. PBX networks are owned by the companies for their local communications. The only time that a company owning a PBX pays the PSTN service provider is when a call goes out of the local area using the service provider's infrastructure. Operations of LANs, WLANs, and WPANs are very similar to a PBX, in that the user owns them and pays monthly charges to the wide area Internet service providers for wide area communications.

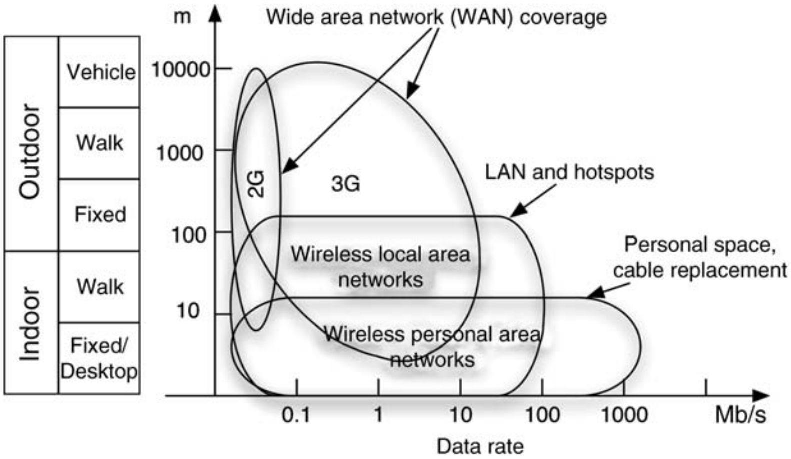


FIGURE 1.8 Relative coverage and data rates of different technologies.

Numerous local network standards and products have evolved around the PSTN and the Internet. To connect to the PSTN, first we had PBX systems and then cordless telephone and personal communication services (PCSs). Standards and products developed in these domains were either simple or they did not gain vast commercial success. As a result, we do not address their technical details in this book; details of more technically challenging wireless PCS standards and products are available in [Pah95, Pah02]. In wired data networking, users connect to the Internet through a LAN. In the past four decades, a number of wired LAN technologies have emerged in the market and the IEEE 802.3 Ethernet has survived as the technology of choice. Again, technologies such as token ring and fiber distributed data interface are not considered in this book; the reader can refer to books such as *Local and Metropolitan Area Networks* [Sta00] for details of these standards. In wireless data networks, the user can either connect directly to the Internet through a lower speed cell phone with universal coverage or can connect through a higher speed WLAN or WPAN that has a limited coverage. Figure 1.8 illustrates the relative coverage and data rates of 2G, 3G, WLAN, and WPAN technologies. In the remainder of this section we provide an overview of the evolution of the local access to the PSTN, IEEE 802.3 Ethernet, IEEE 802.11 WLAN, and the IEEE 802.11 WLANs.

1.3.1 Evolution of Local Access to Public Switched Telephone Network

Connection-based wired local networks were PBX systems which became popular after the Second World War for office applications. Every office in a large organization still has a PBX telephone branch, but it is not used as often as before. Cell phones are more common for voice applications and e-mails/instant messaging are replacing a number of office telephone applications. In large homes, people sometimes use intercoms for connection-based local communications. In the mid 1990s, when most people in the telecommunication industry believed that ATM would take over the entire emerging multimedia communications industry, ATM LAN emulation (LANE) was initiated. The purpose of ATM-LANE was to adapt the existing legacy LAN infrastructures and applications to the then perceived end-to-end ATM network. The main technical challenge in implementing LANE was the adaptation of a connectionless legacy LAN into a connection-oriented ATM network. This technology never gained commercial importance or acceptance.

Local wireless telephone applications started with the introduction of the cordless telephone that appeared in the market in the late 1970s. A cordless telephone provides a wireless connection to replace the wire between the handset and the telephone set. The technology for implementation of a cordless telephone was similar to the technology used in walkie-talkies that had existed since the Second World War. The important feature of the cordless telephone was that as soon as it was introduced in the market it became a major commercial success, selling on the order of tens of millions and generating an income exceeding several billions of dollars.

The success of the cordless telephone encouraged further developments in this field. The first digital cordless telephone was cordless telephone-2 (CT-2), a standard developed in the UK in the early 1980s. The next generation of cordless telephones was wireless PBX using the digital European cordless telephone (DECT) standard. Both CT-2 and DECT had minimal network infrastructures to go beyond the simple cordless telephone paradigm and perhaps cover a larger area and multiple applications. However, in spite of the huge success of the cordless telephone, neither CT-2 nor DECT are considered commercially successful systems. These local systems soon evolved into PCSs that consisted of a complete system with their own infrastructure, very similar to cellular mobile telephone systems.

In the technical communities of the early 1990s, PCSs were differentiated from the cellular systems. A PCS service was considered "personal" and as the next-generation cordless telephone designed for residential areas, providing a variety of services beyond the cordless telephone. The first real deployment of a PCS system was the personal handy phone (PHP), later renamed the personal handy system (PHS), introduced in Japan in 1993. At that time, the technical difference between PCS services and cellular systems was perceived to be smaller cell size, better quality of speech, lower tariff, less power consumption, and lower mobility. However, from the user's point of view, the terminals and services for PCS and cellular looked very similar and the only significant difference was marketing strategy and the way that they were introduced to the market. For instance, around the same time, in the UK, DEC-1800 services were introduced as a PCS service. DEC-1800 was using GSM technology at a higher frequency of 1800 MHz, but it employed a different marketing strategy. The last PCS standard was PACS in the USA, finalized in 1995. All together, none of the PCS standards became a major commercial success or a competitor to cellular services.

In 1995, the Federal Communication Commission (FCC) in the USA auctioned the frequency bands around 2 GHz as the PCS bands, but PCS-specific standards were not adopted for these frequencies. Eventually, the name PCS started to appear only as a marketing pitch by some service providers for digital cellular services, in some cases not even operating in the PCS bands. While the more advanced and complex PCS services evolving from simple cordless telephone applications did not succeed and merged into the cellular telephone industry, the simple cordless telephone industry itself still remains active. In more recent years, the frequency of operation of cordless telephone products has shifted into unlicensed industrial, scientific, and medical (ISM) bands rather than the licensed PCS bands. Cordless telephones in the ISM bands can provide a more reliable link using spread spectrum technology. More recently, Voice-over-IP (VoIP) phones have been introduced in the market which use WLANs in the home. Femtocell technology is emerging for local home applications to integrate wireless local data and cellular telephone applications.

1.3.2 Evolution of the IEEE 802.3 Ethernet

The LAN industry emerged during the 1970s to enable sharing of expensive resources like printers and to manage the wiring problem caused by the increasing number of terminals in offices. By the early 1980s, three standards were developed: Ethernet (IEEE 802.3), token bus (IEEE 802.4) and token ring (IEEE 802.5), specifying three distinct PHY and MAC layers and different topologies for networking over the thick cable medium but sharing the same management and bridging (IEEE 802.1) and logical link control (LLC) layers. With the growing popularity of LANs in the mid 1980s, the high installation costs of thick cable in office buildings moved the LAN industry toward using thin cables, which is also referred to as "cheapernet." Cheapernet covered shorter distances of up to 185 m, compared with the 500 m coverage of thick cables. In the early 1990s, the star topology (often referred to as hub-and-spoke LANs) using easy-to-wire twisted-pair wiring with coverage of 100 m was introduced. Figure 1.9 depicts the evolution of the Ethernet wired LAN from thick to thin and finally twisted-pair networks. The interesting observation is that this industry has made a compromise on the coverage to obtain a more structured solution that is also easier to install. Twisted-pair wiring, also used by PSTN service providers for telephone wiring distribution in homes and offices for over 100 years, is much easier to install. The star network topology opened an avenue for structured hierarchical wiring, also similar to the telephone network topology. Today, IEEE 802.3 (Ethernet) using twisted-pair wiring is the dominant wired LAN technology.

The data rates of legacy LANs (thick, thin, and twisted-pair) were all 10 Mb/s. The need for higher data rates emerged from two directions: (1) there was a need to interconnect LANs located in different buildings oo a campus to share high-speed servers; (2) computer terminals became faster and capable of running high-speed multimedia applications. To address these needs, several standards for higher data rate operations were introduced. The first fast LAN operating at 100 Mb/s was the fiber distributed data interface (FDDI) that emerged in the mid 1980s as a backbone medium for interconnecting LANs. The ANSI

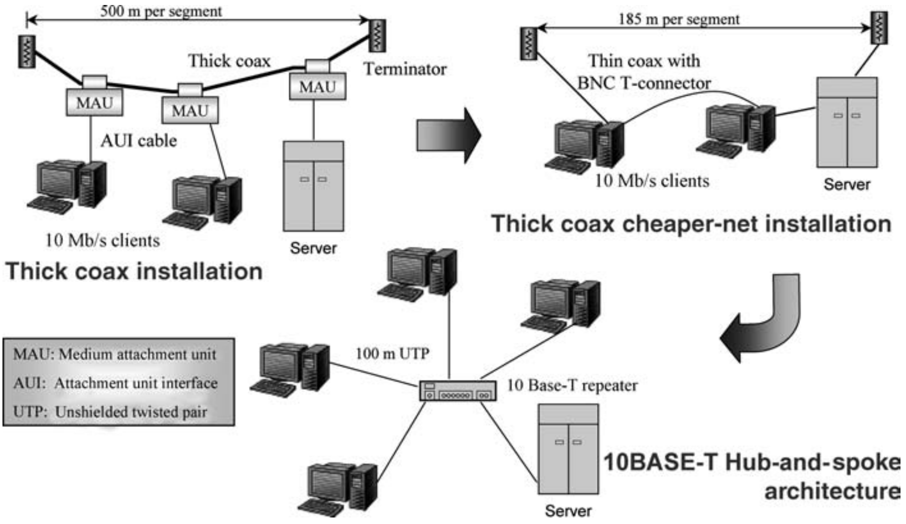


FIGURE 1.9 Evolution of Ethernet topologies.

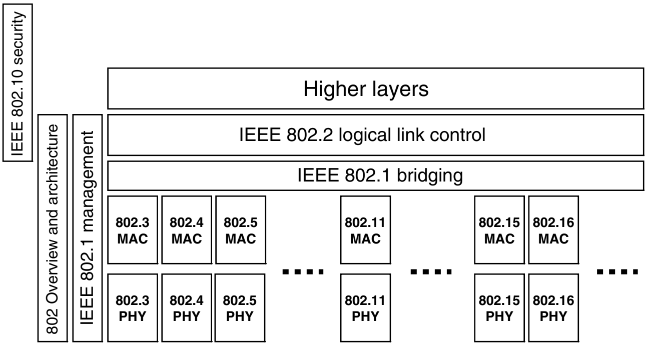


FIGURE 1.10 Organization of IEEE 802 standard series.

published this standard directly. In the mid 1990s, 100 Mb/s fast Ethernet was developed under IEEE 802.3 and 100VG-AnyLAN under IEEE 802.12. In the late 1990s, IEEE 802.3 approved the gigabit Ethernet and more recently 10 Gb/s Ethernet was introduced, mostly for backbone networking in metropolitan areas. Currently, 100 Gb/s Ethernet [Cvi08] is under development in the IEEE 802.3 community. These advances have materialized based on design of new physical layers for more efficient transmission over a variety of wired media.

In general, the LAN industry has developed a number of standards, mostly under the IEEE 802 community. Figure 1.10 shows an overview of the important IEEE 802 community standards. The 802.1 and 801.2 parts are common for all the standards, 802.3, 802.4 and 802.5 are wired LANs, and 802.9 is the so called ISO-Ethernet that supports voice and data over the traditional Ethernet mediums. IEEE 802.6 corresponds to metropolitan area networking and IEEE 802.11, 15, and 16 are related to WLAN, WPAN, and WMAN. IEEE 802.14 is devoted to cable-modem-based networks providing Internet access through cable TV distribution networks operating over coaxial cable wiring and fiber originally installed for TV distribution. IEEE 802.10 is concerned with security issues and operates at higher layers of the protocols. Newer standards beyond IEEE 802.16 are being developed, but these are not shown in Figure 1.10.

1.3.3 Evolution of the IEEE 802.11 Wireless Local-Area Network

Gfeller, at the IBM Rüschlikon Laboratories in Switzerland, first introduced the idea of a WLAN in late 1970s [Gfe80] as a method for local area networking in manufacturing areas. Diffuse infrared (IR) technology was selected for the implementation to avoid interference with electromagnetic signals radiating from machinery and to avoid dealing with long-lasting administrative procedures with frequency administration agencies. Ferrert at HP's Pal Alto Research Laboratories in California performed the second project on WLANs around the same time [Fer80]. In this project, a 100 kb/s direct sequence spread spectrum (DSSS) WLAN operating at around 900 MHz was developed for office areas.

WLANs need a bandwidth of at least several tens of megahertz, but they did not have a market compatible in strength with the cellular voice industry that originally started with two pieces of 25 MHz bands that produced a huge market. The dilemma for the frequency administration agencies was to justify a large frequency allocation for a product with a weak market. In the mid 1980s, the FCC found two solutions for this problem. The first and the simplest solution was to avoid the 1–2 GHz bands used for the cellular telephone and PCS applications and approve higher frequencies at several tens of gigahertz where plenty of unused bands were available. This solution was first negotiated between Motorola and the FCC and resulted in Motorola's Altair, the first WLAN product operating in licensed 18–19 GHz bands. Motorola had actually established a headquarters to facilitate user negotiation with the FCC for the usage of WLANs in different areas. A user who changed the location of operation of their WLAN substantially (from one town to another) contacted Motorola and they would manage the necessary frequency administration issues with the FCC.

The second and more innovative approach was resorting to unlicensed frequency bands as the solution. In response to the applications for bands for WLAN projects mentioned in the previous section and motivated by studies for various implementations of WLANs [Pah85], Mike Marcus of the FCC initiated the release of the unlicensed ISM bands in the May of 1985 [Mar85]. The ISM bands were the first unlicensed bands for consumer product development and played a major role in the development of the WLAN and later on the WPAN industry. Encouraged by the FCC ruling [Mar85] and some visionary publications in wireless office information networks summarizing previous works and addressing the future directions in this field [Pah85,Pah88, Kav87], a number of WLAN product development projects mushroomed almost exclusively over the North American continent. By the late 1980s the first generation of WLAN products using three different technologies, licensed bands at 18–19 GHz, spread spectrum in the ISM bands around 900 MHz, and IR appeared in the market. At around the same time, a standardization activity for WLANs under IEEE 802.4L was initiated that was soon converted into an independent group – IEEE 802.11 that was finalized in 1997.

Another WLAN standardization activity that started in 1992 was the high-performance radio LAN (HIPERLAN). This ETSI-based standard aimed at high performance LANs with data rates of up to 23 Mb/s that was an order of magnitude higher than the original 802.11 data rates of 2 Mb/s. To support these data rates, the HIPERLAN community was able to secure two unlicensed 200 MHz bands: 5.15–5.35 GHz and 17.1–17.3 GHz for WLAN operation. This encouraged the FCC to release the so-called Unlicensed National Information Infrastructure (U-NII) bands in 1997 when the original HIPERLAN standard (now called HIPERLAN/1) was completed. The U-NII bands were used by IEEE 802.11a and HIPERLAN/2 projects for the implementation of 54 Mb/s orthogonal frequency division multiplexing (OFDM)-based WLANs which was completed in 1999.

In the first half of the 1990s, WLAN products were expecting a sizable market of around a few billion dollars per year for shoebox-sized products used for LAN extension in indoor areas and this did not materialize. Under this situation, two new directions for product development emerged. The first and simplest approach was to take the existing shoebox-type WLANs, boost up their transmitted power to the maximum allowed under regulations, and equip them with directional antennas for outdoor interbuilding LAN interconnects. These technically simple solutions would allow coverage of up to a few tens of kilometers with suitable rooftop antennas. The new inter-LAN wireless bridges could connect corporate LANs that were within range. The cost of the inter-LAN wireless solution was much cheaper than the wired alternative, T1-carrier lines, leased from the PSTN service providers. This technology later evolved into point-to-multipoint WiMAX metropolitan area networks to solve the "last mile" problem for Internet service providers, which is discussed later. The second alternative was to reduce the size and cost of the design to a Personal Computer Memory Card International Association (PCMCIA) WLAN card to be

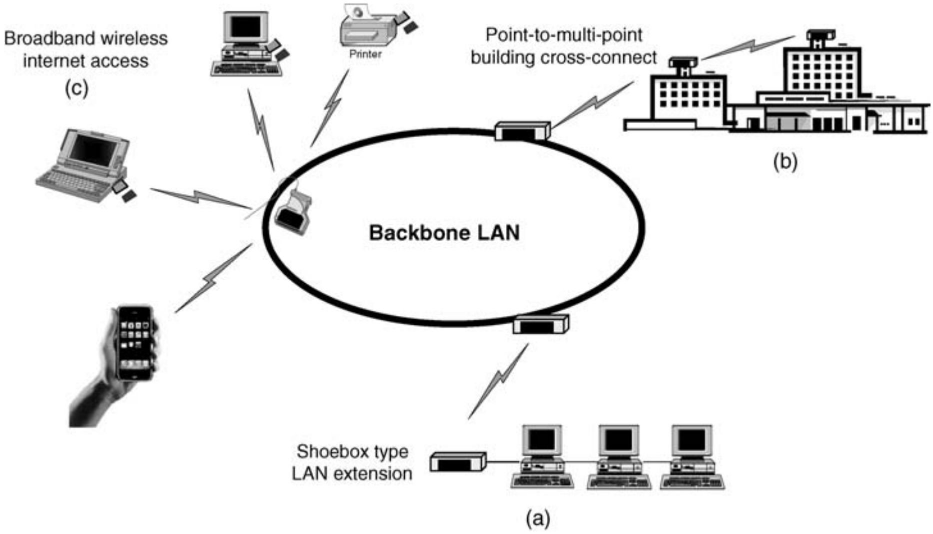

Broadband wireless internet access
(c)
Printer
Point-to-multi-point building cross-connect
(b)
Backbone LAN
Shoebox type LAN extension
(a)

FIGURE 1.11 Evolution of WLAN products and applications: (a) legacy LAN-extension; (b) inter-LAN bridge; (c) broadband wireless Internet access.

used with laptops that were enjoying a sizable growth and demanded mobility for LAN connectivity. These products opened a new horizon for massive market development for WLANs in small office and home office (SOHO) networking and hot-spot Internet access which resulted in an exponential growth of the WLAN industry in the past decade. Most recently, WLANs are being integrated in most mobile phones and smartphones. To complement the home and office networking products, there are also low-cost products for LAN extension that can convert a serial port or Ethernet connector to a WLAN interface for desktop PC or printer connections. Figure 1.11 illustrates the evolution of applications for the WLANs.

Figure 1.12 illustrates the chronology of the evolution of the WLAN industry. The first small hump after the emergence of the technology was halted by the lack of frequency bands. The second and larger hump evolved after the release of unlicensed ISM bands which resulted in the shoebox products for LAN extension in the early 1990s. The third period of continual growth started after completion of the first IEEE 802.11 standard at 2 Mb/s followed by release of IEEE 802.11b at 11 Mb/s and IEEE 802.11a/g and “n” to support 54 Mb/s and over 100 Mb/s consecutively. These technologies supported successful expansion of the market to home networking, hotspots, corporate networking, and home access. More recently, WLAN signals are being exploited for localization applications as a softer product complementing global positioning system (GPS) technology which does not work properly in indoor and urban areas.

1.3.4 Internet Access to Home and IEEE 802.16

The connection between a network provider and the customer which provides the access to the network is very important for network providers. In the PSTN and cable TV industries this connection is referred to as the “last mile”. The last mile connection is important because it is the route to get any information deliverable to the customer for generating

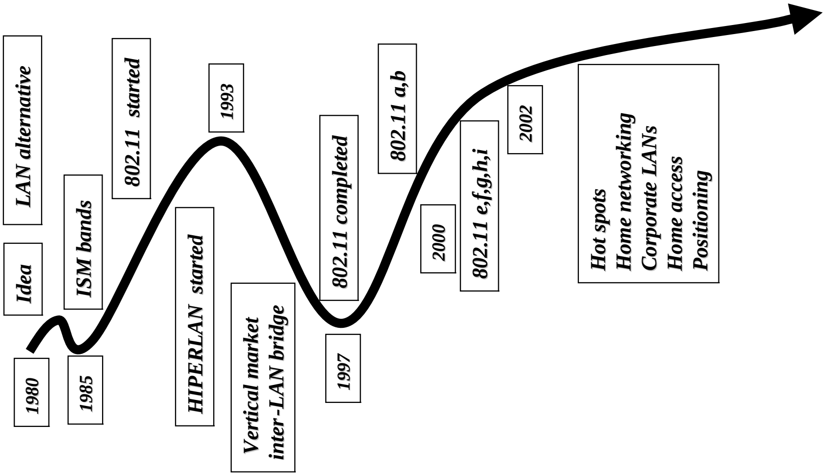


FIGURE 1.12 Chronology of the evolution of the WLAN industry.

income. However, it is an expensive challenge, because laying wires or cables is a considerably laborious undertaking. For that reason, a wireless solution to the last-mile problem has always attracted attention. The traditional PSTN service has been the analog POTS over the twisted-pair wirings and the cable TV distributors were using multichannel analog TV over coaxial cable. In modern times, the need for broadband Internet access has stimulated a number of innovative approaches in this industry. In office environments, the cost of the installation of the access leg of the network can be better justified by a larger size of the income from customer service. In the home environment, where usually income is generated from only one private user, the justification of the cost of making the connection becomes more challenging. In addition, the diversity of terminals used in the home environment is very wide. Most of the evolving terminals shown in Figure 1.2 are used at home. Figure 1.13 provides a classification of the home equipment demanding networking. Different classes need different technologies to connect to one another and they use different home access techniques to connect to WANs. We discuss local networking of this equipment in the next subsection, and in the rest of this section we focus on the access technologies.

Figure 1.14 illustrates the traditional networking connections in most residences. A typical traditional residence in the USA is connected to the PSTN for telephone services, the Internet for web access, and a cable network for multichannel TV services. Within the home, computers and printers are connected to the Internet through voice-band modems, DSL services, or cable modems. The telephone services and security systems are connected through PSTN wiring. The TV is connected to multichannel services through HFC cables or satellite dishes. Audio and video entertainment equipment, such as video cameras and stereo systems, and other computing systems, such as laptops, are either isolated or have proprietary wired connections. This fragmented networking environment has prompted a number of recent initiatives to create a unified home network. The home networking industry started in the early 2000s by the design of the so-called home or residential gateways to connect the increasing number of information appliances at home through a single Internet connection.

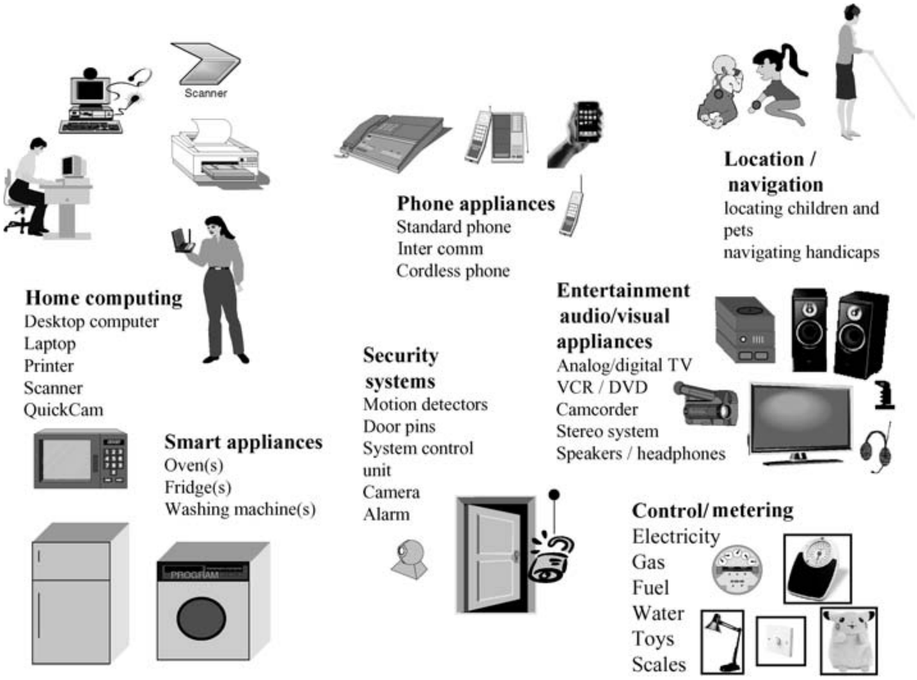


FIGURE 1.13 Classification of home equipment demanding networked operation.

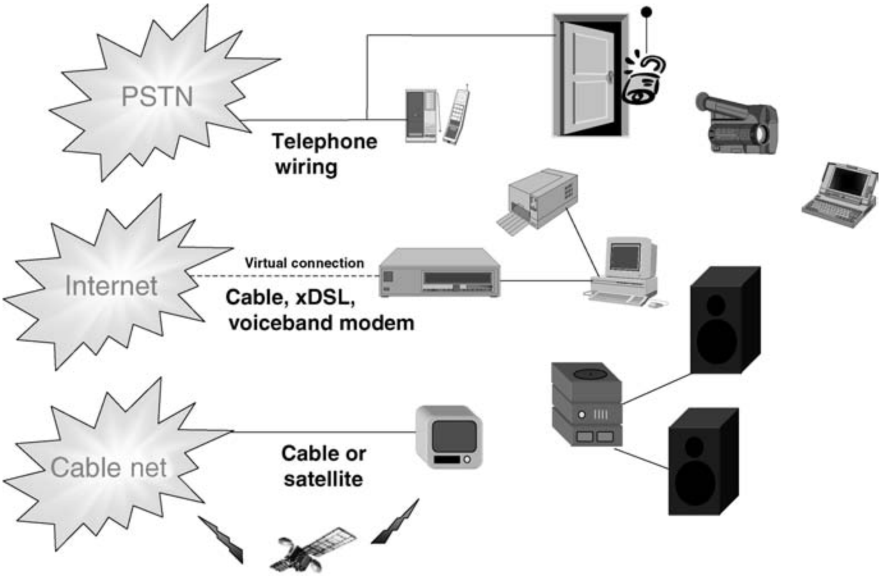


FIGURE 1.14 Traditional fragmented home access and distribution networks.

The early home access technology was based on voice-band modems. Today, broadband home access (with data rates on the order of 10 Mb/s) is provided through cable modems and DSL services over the telephone lines. Cable modems operate on the cable TV wiring. Some of the TV channel frequency bands are allocated to cable modems to provide broadband access to the Internet. The cable distribution in the residential areas has a bus topology that is optimally designed for one-way TV signal distribution. The bus carries all the stations in the neighborhood and the cable-TV box selects the station for the TV and if it is a scrambled paid channel it also de-scrambles the signal. To control the set-top boxes, a reverse channel is available in all modern cable wiring. Broadband cable services use one of the video channels and the reverse channel to establish a two-way communication and access to Internet. The DSL services use the 25 kHz–1.1 MHz bands on the telephone wirings to support broadband Internet access to the users. The topology of the telephone line is a star topology that connects every user directly to the end office where the DSL data is directed to the Internet through a router. The latest services for selected areas in the USA are the fiber to home solutions which are more popular in some other countries, such as South Korea or Japan.

Fiber is an excellent medium with respect to information capacity but is not readily available to most end users worldwide. Fiber optic lines are generally laid underground in conduits requiring a relatively expensive installation which poses an economical challenge for their rapid deployments worldwide.

For places with wiring problems, a wireless solution to the last mile problem has attracted considerable attention in the past decade. These services are sometimes referred to as fixed-wireless services. The advantage of using fixed-wireless solutions is that it does not involve wiring the streets. If there is no available wiring in the neighborhood, then a wireless solution is certainly the main economical solution, and for that reason it has become very popular in countries where networking infrastructure is limited. WLAN technologies can be used for this purpose, as we discussed earlier, by changing the antenna and boosting the transmitting power they can cover up to a few miles. Without changing the power, if the antennas are posted in heights outside, WLANs should easily cover a mile in line-of-sight (LOS) conditions as an inter-LAN bridge. Another alternative for broadband wireless home access is to use the IEEE 802.16 technologies. IEEE 802.16 has been working on this solution since the late 1990s. Other wireless alternatives are direct satellite TV broadcasting and 3G wireless networks. Direct broadcast suffers from the lack of a reverse channel and high delay that challenges the implementation of broadband services on this medium. The high-speed 3G wireless packet data services are expected to provide up to 2 Mb/s that is very suitable for Internet access. The data rates on these systems are lower and they are using licensed bands that ultimately may be expensive. Figure 1.15 summarizes the existing solutions for the home access technologies.

The most popular of all wireless solutions in recent years has been the IEEE 802.16 solution. This standard committee officially started in 1999 for the so-called local multipoint distribution system (LMDS). But it was a short-term hype attracting attention from cellular network equipment manufacturers. In 2001, the WiMAX forum industrial group was formed to improve 802.16 as an alternative to the popular DSL and cable modem technologies. This initiative finally introduced the IEEE 802.16d in 2004 as a fixed wireless solution and IEEE 802.16e, also known as mobile WiMax, in 2005. In comparison with WLAN solutions, WiMAX defines a more complex architecture to support QoS and mobility using outdoor antennas. Compared with cellular networks, it has smaller coverage and uses OFDM technology rather than CDMA and it has options for operation in

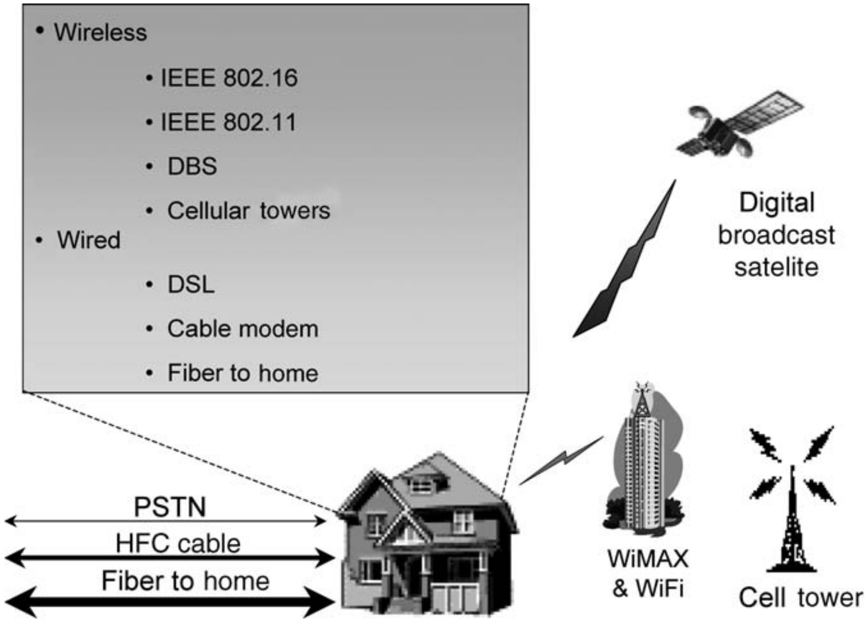


FIGURE 1.15 Broadband home access alternatives.

unlicensed as well as in licensed frequency bands. Therefore, WiMAX can be considered a technology between the WLAN and cellular networks with more similarities to WLANs. As a result, we discuss this technology as a part of Chapter 9.

1.3.5 Evolution of IEEE 802.15 Wireless Personal-Area Networks

Figure 1.16 illustrates the basic concept behind different classes of applications for local networks inside a home. The data rate, power consumption, and coverage for the networking technology that interconnects these devices are quite diversified, demanding different technologies. In response to that diversity of requirements, a number of local networking standards have evolved in the past decade. Figure 1.16 provides a simplified overview of the popular IEEE 802 standards on local networking as they relate to a variety of applications demanding data rates from a few kilobits per second up to a gigabit per second. For low data rate applications, such as utility metering, smart appliances, and security systems, the IEEE 802.15.4 and, for audio and short-range phones, IEEE 802.15.1 provide good solutions. For mid-range data rate applications for home computing and Internet access, IEEE 802.11 suits well, and IEEE 802.15.3 is a better choice for high-speed applications involving good quality video or large file transfers. We have already discussed IEEE 802.11WLANs in Section 1.3.3. Here, we give an overview of the IEEE 802.15 WPANs.

The very first personal area network (PAN) to be announced was the BodyLAN that emerged from a DARPA project in the mid 1990s. This was a low-power, small size, inexpensive, wireless PAN with modest bandwidth that could connect personal devices in many collocated systems with a range of around 5 feet [Den96]. Motivated by the BodyLAN

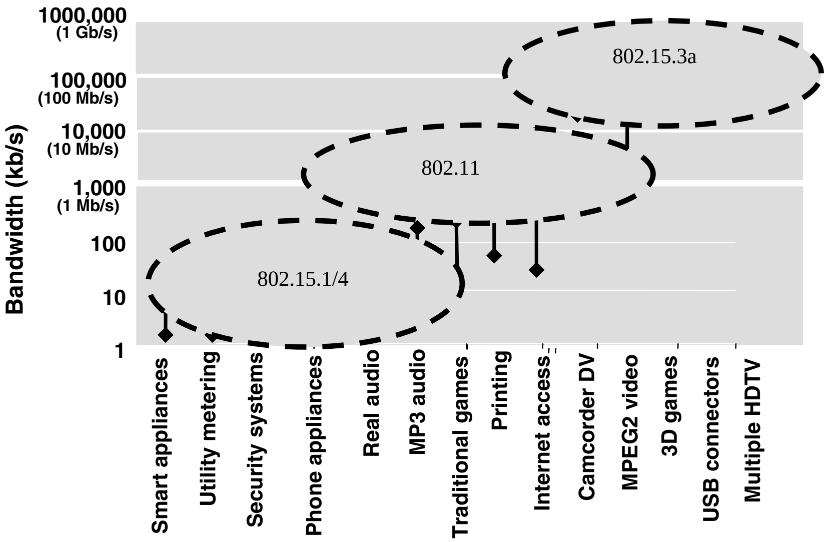


FIGURE 1.16 Applications bandwidth requirements and IEEE standards.

project, a WPAN group originally started in June 1997 as a part of the IEEE 802.11 standardization activity. In January 1998, the WPAN group published the original functionality requirement. In May 1998 the Bluetooth development was announced and a Bluetooth special group was formed within the WPAN group [Sie00]. In March 1999, IEEE 802.15 was approved as a separate group in the 802 community to handle WPAN standardization. This standardization committee adopted Bluetooth technology as its first standard, IEEE 802.15.1. Like legacy IEEE 802.11WLANs, the IEEE 802.15.1 Bluetooth technology operates in ISM unlicensed bands. Bluetooth operates at lower data rates than WLANs but uses a voice-oriented TDMA wireless access method that provides a better environment for supporting QoS needed for better quality of voice in an integrated voice and data service. The weakness of Bluetooth is its limitation on the number of simultaneous users and a slow network connection start up. IEEE 802.15.4 defined the first version of a low-power, low-cost, and low data rate WPAN in 2003. This standard is also known as ZigBee after the name of the industrial alliance group defining higher layer specifications for application development over MAC and PHY defined by IEEE 802.15.4. The MAC layer of the ZigBee is contention-based data-oriented CSMA with collision avoidance (CSMA/CA), which is a lighter version of the MAC used in IEEE 802.11 allowing larger numbers of users and faster network set up.

In 2003, a new wave of interest for ultra wideband (UWB) technology initiated the IEEE 802.15.3 group for standardization within 802.15. This group aimed at WPANs with extremely high data rates of up to 1 Gb/s to support short-range wireless connection for cable replacement and video-based applications. This group has defined two options for UWB WPANs in the unlicensed band in 3.4–10.6 GHz and for the time being has delayed their meetings until the market in this area evolves. Figure 1.17 shows a chronology of the evolution of the WPAN industry and the IEEE 802.15 standardization group.

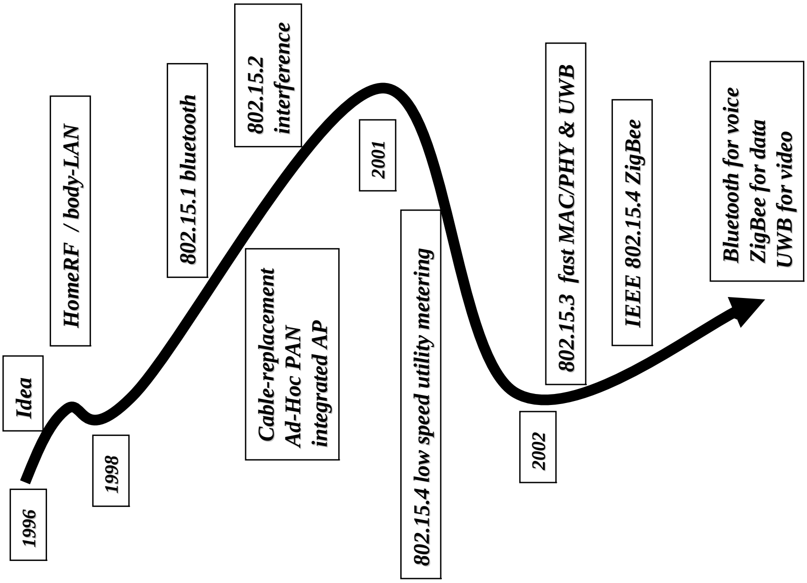


FIGURE 1.17 Chronology of evolution for WPAN industry.

1.4 STRUCTURE OF THE BOOK

This book is organized to start with a chapter providing an overview of information networks followed by four parts each including several chapters. The first chapter provides an overview of evolution of wired and wireless information networks from 1834 when the telegraph was introduced up to the modern time and the introduction of the iPhone and the Internet of Things.

Part One of the book is devoted to the fundamentals of transmission and access. This part of the book consists of four chapters describing the fundamentals of channel behavior, data transmission, coding, and MAC techniques. These chapters prepare students to understand the technical aspects related to emergence of new wired and wireless networking technologies. To the students with extensive background in signal analysis, these chapters provide a comprehensive summary of applied techniques that they may have seen in part in other courses. For students without a background in signal analysis these chapters provide an intuitive understanding of the lower layer issues, and with some extra effort they can learn the fundamentals of operation in these layers of networking. Chapter 2 is devoted to the characteristics of the wired and wireless medium. It provides the details of how the different guided and wireless media operate and how we can differentiate among cable, wire, fiber, and wireless networking in different frequencies. We discuss how bandwidth relates to length of a guided medium and how one can calculate coverage of a wireless network in different environments. Chapter 3 describes the fundamentals of data transmission techniques applied to wired and wireless networks. This chapter provides the details of how electrical signals are used as mathematical symbols to carry a bit stream generated by an application. Chapter 4 studies reliable transmission and coding techniques. Here, we provide a survey of coding techniques and the congestion control protocol applied to popular networking applications using simple examples explaining how we can implement them. Chapter 5 is devoted to MAC techniques used in a variety of LANs and wireless networks. This includes assigned access techniques such as FDMA, TDMA, and CDMA used in cellular telephone networks and different versions of CSMA techniques used in wired and wireless local networks connecting to the Internet.

Part Two of the book provides the details of technologies used in dominant WANs of the modern time, the Internet and cellular networks. Chapter 6 describes the interconnecting technologies used to implement the Internet. Chapter 6 describes issues related to addressing and QoS before a detailed description of bridging, switching, and routing technologies. These are issues and technologies used for the implementation of the Internet. Chapter 7 describes TDMA and CDMA cellular network technologies used in the GSM and 3G cellular networks and explains how these networks are deployed in the field. Here, we describe how packets are formed, how connections are made, and how the information transfers in connection based networks.

Part Three of the book is devoted to LANs and PANs. This part includes three chapters. Chapter 8 describes details of Ethernet. Here, we describe how an Ethernet packet is formed, how the medium is accessed by a packet, and how the bits in the packets are transmitted over different media. Chapter 9 describes the same feature for the IEEE 802.11 WLAN technologies known as WiFi and its extension to the WiMAX technology. Chapter 10 describes the different WPAN technologies used in lower speed Bluetooth and ZigBee technologies, as well as high-speed UWB systems. These chapters provide a wide variety of popular applications of the fundamental material presented in Part One of the book.

Part Four is the last part of the book, describing three important system aspects of networks: security, localization, and sensor networking. Chapter 11 is devoted to the security aspects of information networks. This chapter describes how the security of a network can be attacked and what can be done to protect a network and prevent these attacks. Chapter 12 addresses location sensing and geolocation systems. It describes how we can sense the radio-frequency (RF) signals in a way that it provides an indicator of the location and how we can use these indicators to localize a terminal in cellular or other wireless networks. Chapter 13 describes issues related to sensor networks and how we use WPAN technologies to implement them.

QUESTIONS

1. How is a wireless network different from a wired network? Explain at least five differences.
2. What is the difference between the 3G cellular networks, WLANs, and WPANs in terms of frequency of operation, orientation of the application over the network, and supported rates for data services?
3. Name the three major telecommunication services that dominate today's commercial services and give the approximate time when they were first introduced.
4. What is a WPAN? What is the difference between WPANs and WLANs? Name two example technologies for WPANs.
5. What is the difference between connectionless and connection-oriented backbone networks? Name the two major networks supporting these techniques.
6. What was the first transmission technique used for telecommunications? Was it voice or data? Analog or digital?

7. How many years did it take to go from the invention of the telegraph to transoceanic telegraph? How does it compare with the same thing for wireless telegraph? Explain why.
8. When and how were the first computer communication networks started?
9. How has the Internet evolved?
10. Name the three major existing infrastructures that support home networking.
11. What are the available access methods at homes?
12. What are the differences between a LAN and a WAN in terms of data rate, geographical coverage, and ownership?
13. What are the four generations of wired LANs?
14. What are the major home distribution technologies?
15. Name four home access technologies.
16. What is the difference between a MAN and a WAN?
17. Which standardization bodies regulate 3G cellular, WLAN, WPAN, and WiMax and what are the differences among these technologies in terms of frequency regulations, supporting data rates, and the coverage of the service?
18. What do you see as the next generation of the Internet and how does that relate to emergence of RFID applications?
19. Name three low speed (less than 1 Mbps), three medium speed (1–10 Mbps), and three high speed (10–1000 Mbps) applications. Give three IEEE standards which address the needs of these three classes of applications.
20. What is Ethernet and how does it relate to the IEEE 802 standardization activities? What was the speed of legacy Ethernet and what is the speed of the latest recommendations by the standardization committee?
21. What were the most popular TDMA and CDMA standards used in 2G cellular networks, where were these standards developed, and how did they evolve into the 3G networks?
22. What are the advantages of SONET over T1 transmission systems?
23. What are the differences between the different 802.15 standards in terms of applications that they target and the data rates they support?

PROJECT 1

Search Chapter 1 and the Internet (IEEE Explore, Wikipedia, Google Scholar, ACM Digital Library) to identify one area of research and one area in business development which you think are the most important for the future of the information networking industry. Give your reasoning as to why you think the area is important and cite at least one paper or a website to support your statement.

PART ONE

FUNDAMENTALS OF TRANSMISSION AND ACCESS

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

2

CHARACTERISTICS OF THE MEDIUM

2.1 INTRODUCTION

A wired medium provides a reliable *guided* link that conducts an electrical signal associated with the transmission of information from one fixed terminal to another. There are a number of alternatives for wired connections, including the twisted pair used in telephone wiring and high-speed local area networking, coaxial cables used for television distribution, and fiber optics used in the backbone of long-haul networks. Guided media act as *filters* that limit the maximum transmitted data rate of the channel because of band-limiting frequency response characteristics. The signal passing through guided media may also radiate outside the wire to some extent, which can cause interference to nearby radio or other wired transmissions. These characteristics differ from one wired medium to another. Laying additional cables can, in general, duplicate the wired medium and thereby we can increase the bandwidth.

Compared with wired media, the wireless medium is unreliable, has low bandwidth, and is of broadcast nature. However, it supports mobility due to its tetherless nature and it is free of wiring installation costs. Different signals through wired media are physically conducted through different wires, but all wireless transmissions share the same medium, namely air. Thus, it is the frequency of operation and the legality of access to the band that differentiates the various alternatives for wireless networking. Wireless information

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

networks operate around 1 GHz (cellular), 2 GHz (PCS and WLANs), 5 GHz (WLANs), 30–60 GHz (point-to-point BS connections), and IR frequencies for optical communications. These bands are either licensed (similar to cellular and PCS bands) or unlicensed (similar to the ISM bands). As the frequency of operation and data rates increase, the implementation cost (in hardware) increases and the ability of a radio signal to penetrate walls decreases. The electronic cost difference has become insignificant with time, but in-building penetration and licensed versus unlicensed frequency bands have become important differentiations. For frequencies up to a few gigahertz the signal penetrates through the walls, allowing indoor applications with minimal wireless infrastructure inside a building. At higher frequencies, a signal that is generated outdoors does not penetrate into buildings and the signal generated indoors stays confined to a room. This phenomenon imposes restrictions on selection of a suitable band for a wireless application.

For example, if it is intended to bring a wireless Internet service to the rooftop of a residence and distribute that service inside the house using other alternatives, such as existing cable or twisted-pair wiring, then one may select WMAN equipment operating in licensed bands at several tens of gigahertz. If the intention is to penetrate the signal into the building for direct wireless connection to a computer terminal, then one may prefer equipment operating in the unlicensed ISM bands at 900 MHz or 2.4 GHz. The first approach is more expensive, because it operates at licensed higher frequencies where implementation and the electronics are more expensive and the service provider has paid to obtain the frequency bands. The second solution does not have any interference control mechanism, because it operates in unlicensed bands.

Wired media provide us with an easy means to increase capacity: we can lay more wires where required if it is affordable. With the wireless medium, we are restricted to a limited available band for operation and we cannot obtain new bands or easily duplicate the medium to accommodate more users. As a result, researchers have developed a number of techniques to increase the capacity of wireless networks, to support more users with a fixed bandwidth. The simplest method, comparable to laying new wires in wired networks, is to use a cellular architecture that reuses the frequency of operation when two cells are adequately far from one another. Then, to increase the capacity of the cellular network further, as will be explained in Chapter 7, one many reduce the size of the cells. In a wired network, doubling the number of wired connections will allow twice the number of users at the expense of twice the number of wired connections to the terminals. In a wireless network, reducing the size of the cells to a half will allow twice as many users in one cell. Reduction of the size of the cell increases the cost and complexity of the infrastructure that interconnects the cells. Recently, expanding the capacity of wireless links has become possible by exploiting the rich multipath environment with multiple-input–multiple-output (MIMO) techniques.

In this chapter, we first present the characteristics of guided media. Twisted-pair and coaxial cables are presented in Section 2.2 and then the characteristics of the wireless medium are discussed in Section 2.3. Propagation of signals in the wireless medium is fairly complicated and, hence, this section is more detailed than Section 2.2. The reader is referred to Thompson *et al.* [Tho06] for more details on the properties of copper and fiber media and to Pahlavan and Levesque [Pah05] for a rigorous treatment of wireless channels.

2.2 GUIDED MEDIA

The most common guided media for communications are twisted-pair wires, coaxial cables, optical fibers, and power line wires. Twisted-pair wires, in shielded twisted-pair (STP) and unshielded twisted-pair (UTP) forms, are available in a variety of categories. They are commonly used for local voice and data communications in homes and offices. The telephone companies use category 3 UTPs to bring POTS to customer premises and distribute the service in the premises. This wiring is also used for voice-band modem data communications, ISDN, DSL, and home phone networking (HPN). Wired media in LANs are dominated by a variety of UTP and STP wires to support a range of local data services from several megabits per second up to several gigabits per second within a distance of 100 m.

Coaxial cable provides a wider band useful for multichannel FDM operation, lower radiation, and longer coverage than twisted pair, but it is less flexible, particularly for installations in indoor areas where the cables need to run inside the walls and ceilings. The early LANs were operating on so-called thick cable to cover up to 500 m per segment. To reduce the cost of wiring, thin cable (sometimes referred to as cheaper-LAN), that covers up to 200 m per segment, replaced thick cables. Cabled LANs are not popular anymore and twisted-pair wiring is taking over the LAN market. Another important application for cable is cable television, which has a huge network to connect homes to the cable TV network. This network is also used for broadband access to the Internet. The cable LANs were using baseband technology with data rates of 10 Mb/s, while broadband cable-modems provide comparable data rates over each of around 100 cable TV channels. Today, in addition to traditional cable TV, cable modems are becoming a popular access method to the Internet, and some cable service providers are offering voice services over the cable connections.

Fiber-optic lines provide extremely wide bandwidth, smaller size, lighter weight, less interference, and very long coverage (low attenuation). However, fiber lines are less flexible and more expensive to install, not suitable for FDM (though WDM is becoming common), and have expensive electronics for TDM operation. Because of the wide bandwidth and low attenuation, optical fiber lines are becoming the dominant wired medium to interconnect switches in all long-haul networks. In LAN applications, fiber lines mostly serve as the backbone to interconnect servers and other high-speed elements of the local networks. Optical fibers have not found a considerable market for distributing any service to the home or office desktop.

Power-line wiring has also attracted some attention for low-speed and high-speed home networks. The bandwidth of power lines is more restricted, the interference caused by appliances is significant, and the radiated interference in the frequency of operation of AM radios is high enough that frequency assignment agencies do not allow power-line data networking at these frequencies. However, power lines have good distribution in existing homes because power plugs are available in all rooms. In addition, almost all appliances are connected to a power line and with only one connection a terminal can connect to the network and the power supply. Existing power-line networks either operate at low data rates at low frequencies, below the frequency of operation of AM radios, to interconnect evolving smart appliances, or they operate at high data rates of up to 10 Mb/s in frequencies above AM radio in support of home computing.

Wireless channels considered in this book are all for relatively short connections. For long-haul transportations these networks are complemented by wired connections. As we have discussed in this chapter, the average path loss in wireless channels is exponentially related to the distance, and we often describe the path loss in terms of a fixed decibel value per decade of distance (see Appendix A for a definition of the decibel). The path loss in wired environments is linearly related to the distance. Therefore, wired connections are more power efficient for shorter distances and wireless for longer distances.

The category 3 UTP wires used for Ethernet lose around 13 dB per 100 m distance [Sta00]. Therefore, the path loss for 1 km is 130 dB. In free space, the path loss of a 1 GHz radio in the first meter is around 30 dB (with an omnidirectional dipole antenna) and after that the path loss is 20 dB per decade of distance. Therefore, the path loss at 100 m is 70 dB and at 1 km it is 90 dB. Indeed, for 130 dB path loss the 1 GHz radio can cover 1000 km. Therefore, for short distances, the path loss with wired transmission is smaller, but for long distances the path loss with radio systems is smaller. This is the reason why repeaters are commonly used in long-haul wired transmissions. Also, for the same reason, radio signals are able to provide very long-distance satellite communications.

2.2.1 Twisted Pair

Twisted-pair wiring is the most popular wiring used in distribution and access networks for end users. At home, twisted pairs are used for access and distribution of telephone services and Internet access using DSL. In offices, the original twisted pairs were used for telephone distribution and a separate better graded twisted pair is used for LAN connections.

In twisted-pair wiring, as the name indicates, two insulated copper conductors are twisted together for the purposes of canceling out electromagnetic interference from external sources and crosstalk from neighboring wires. The greater the number of twists, the greater the attenuation of crosstalk. Twisted-pair wires are normally bundled in a sheath as a group of pairs differentiated by a standard color code. The original twisted pairs first widely used by the telephone companies across the world did not have any shields between the pairs or around the bundle, and today they are referred to as UTP. Figure 2.1 shows a general picture showing details of twisted-pair wiring. Figure 2.1*a* shows how wires are twisted together,

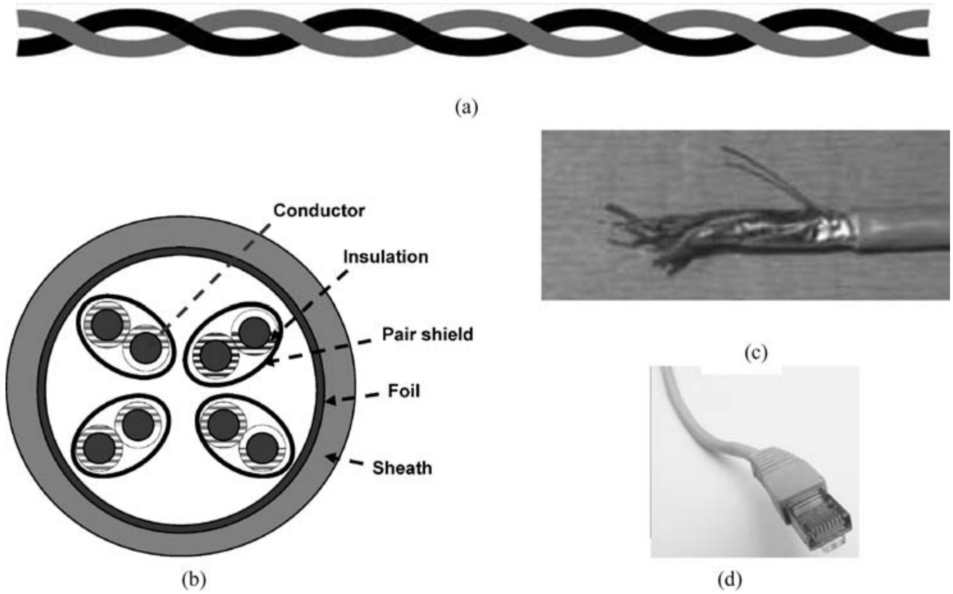


FIGURE 2.1 Twisted-pair wirings: (*a*) basic twisting concept; (*b*) details inside a bundle of four pairs; (*c*) photograph of a sample twisted-pair bundle; (*d*) RJ-45 twisted-pair connector.

Fig. 2.1*b* shows all possible details inside the sheath jacket of the wire, Fig. 2.1*c* shows a picture of a piece of UTP wiring used for Ethernet wiring, and Fig. 2.1*d* shows a registered jack (RJ-45) connector usually used with twisted-pair cables. If individual pairs in a bundle are shielded with metallic screens, then the bundle is referred to as an STP. If the bundle has a metallic shield, then it is referred to as foiled twisted pair (FTP). The most complicated twisted-pair wiring has both foil and shield and is referred to as S/FTP. Therefore, the characteristics of the twisted-pair wires depend on the number of twists, shielding complexity, and the length of the cable. For local distribution inside a home or office, the length of the twisted pairs is less than 100 m; for local access to the network it may go up to a couple of kilometers.

In industry, different grades of twisted-pair wiring are referred to in terms of categories. The original twisted-pair cables first used in the American telephone network around the 1900s are now referred to as category 1 (Cat-1) wiring. This wiring was previously used in POTS, ISDN, and also for doorbells. The Cat-1 twisted pair, however, is not recognized by a TIA/EIA standard. Twisted pairs dominated the twentieth century wiring of the telephone and computer communications networks, and a number of twisted-pair wiring options became standards for wiring for indoor and outdoor applications. The most popular twisted-pair wirings are Cat-3 and Cat-5. Billions of meters of these twisted-pair wirings around the world are installed, mostly by telephone companies, and are commonly used in telephone and Ethernet networks respectively. Recently, twisted-pair wiring has been complemented by cable and fiber media, which are described later in this chapter.

Telephone wiring had 25 color codes and the bundles of the telephone wiring could be as large as 25 pairs; but today, the most popular bundles carry four pairs, as shown in Fig. 2.1*b*. Today, the most popular twisted-pair wiring is the so-called voice-grade Cat-3 UTP wiring defined by the TIA/EIA 568-B standard, which has three or four twists per foot and can support up to 16 MHz of bandwidth for 100 m of the wires bundled in four pairs. The main question we arrive at at this point is how to define the bandwidth.

As we show in Chapter 3, increasing the symbol transmission rate or the "bandwidth" of the system will increase the error rate of the system. This error rate is a function of the signal-to-noise ratio (SNR) of the received signal. The noise is either caused by thermal noise of the receiver components or interference from other sources. The dominant source of noise in bundled twisted-pair wiring is the interference between the wires in the same bundle. In industry, two terminologies are used for these interferences: near end crosstalk (NEXT) and far end cross talk (FEXT), which refer to the electromagnetic interference caused by crosstalk among parallel pairs of wires in a bundle. Figure 2.2 illustrates the nature of these crosstalks in a four-pair bundle of twisted-pair wires. The dominant crosstalk occurs at the ends of the cable where all protection on the wire is removed to connect the wire to the connector jack. Standardization organizations, such as TIA/EIA, specify the maximum allowed NEXT and FEXT for 100 m twisted pairs for different categories, and manufacturers design their wires to comply with these specifications.

The path-gain function $G_p(f,l)$ of a Cat-3 twisted pair as a function of the frequency f (MHz) and length of the cable l (m) can be approximated as

$$G_p(f,l) = -0.0235(\sqrt{f} + 0.1f)l \tag{2.1}$$

The relation between the power or path-loss of a guided medium and the distance is linear, while the power loss in a conductor is related to the square root of the frequency. Deviation from the ideal square-root relation to frequency by adding a linear component is caused by

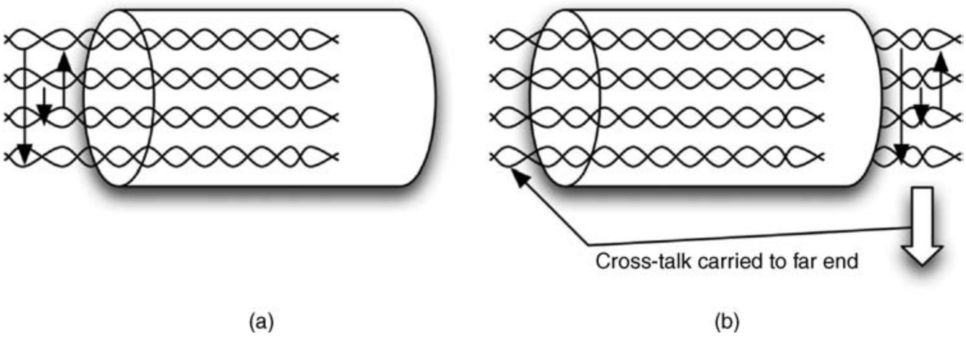


FIGURE 2.2 Cross-pair interference in twisted-pair bundles: (*a*) NEXT; (*b*) FEXT .

dielectric isolation loss and other practical issues [Che97]. The solid line in Fig. 2.3 shows this relation, Eq. (2.1), for frequencies up to 40 MHz for a Cat-3 cable with a length of 100 m used as the limit of the length for twisted-pair LAN options in the IEEE 802 standardization committees. As the frequency increases, the signal becomes weaker; at 16 MHz the signal strength is approximately 13.2 dB weaker.

On the other hand, the NEXT interference for 100 m Cat-3 can be approximated by

$$\mathrm{NEXT}(f) = -23 + 15 \log\left(\frac{f}{16}\right) \tag{2.2}$$

The dashed line in Fig. 2.3 shows the NEXT interference, Eq. (2.2), for a Cat-3 twisted-pair wiring as a function of frequency. As the frequency increases, the interference is increased. At 16 MHz, the NEXT interference is approximately −23 dB. Therefore, the SNR, which is the difference between the signal and interference levels at 16 MHz, is 9.8 dB. As we show in Chapter 3, each transmission technique requires a different level of SNR to support a given data rate. Therefore, the cable path gain characteristics versus interference level govern the

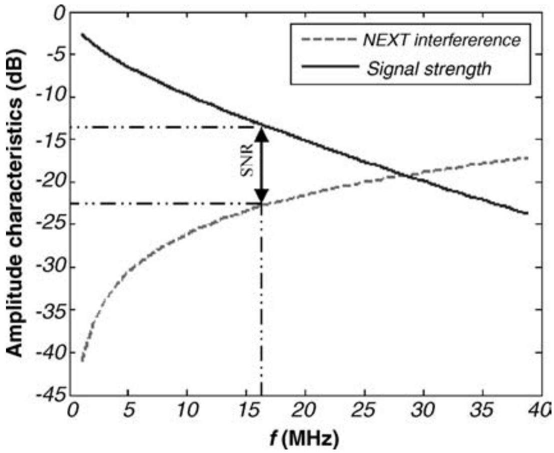


FIGURE 2.3 Amplitude characteristics of the Cat-3 twisted-pair wires; the solid line is the 100 m cable attenuation and the dashed line shows the interference.

supported data rate of a transmission technique. For example, the characteristics of the Cat-3 twisted pair described in Fig. 2.3 allow a 5×5 pulse amplitude modulation (PAM) technique (described in Chapter 3), which can carry up to 100 Mb/s over two pairs (each 50 Mb/s) of Cat-3 voice-grade twisted-pair wirings for 100 m [Che97].

To improve the performance of a twisted pair to support higher data rates, Cat-5 wiring was specified by the TIA/EIA standard 568-A. To implement this twisted pair we need three or four twists every inch, rather than for every foot in Cat-3. This arrangement reduces the power versus frequency and the amount of NEXT interference. The approximated equation for path gain is now given by

$$G_{\mathrm{p}}(f, l) = -0.02(\sqrt{f} + 0.01f)l \tag{2.3}$$

and the NEXT interference by

$$\mathrm{NEXT}(f) = -32 + 15 \log\left(\frac{f}{100}\right) \tag{2.4}$$

Figure 2.4 shows the plots of the signal strength, Eq. (2.3), versus NEXT interference, Eq. (2.4). At 100 MHz the signal strength is attenuated by 22 dB and the NEXT interference level is −33 dB, which results in an SNR = 10 dB and a performance at 100 MHz which is comparable to that of Cat-3 at 16 MHz. As we describe in Chapter 8, this characteristic of Cat-5 has been exploited by the IEEE 802.3 standardization committee to define Ethernet options operating at 100 Mb/s and above.

Based on the above discussion, in the computer communications literature, it is customary to state that the bandwidth of the Cat-3 UTP is 16 MHz and the bandwidth of Cat-5 UTP is 100 MHz. As the need for higher data rates at gigabits per second raised the bar, TIA/EIA standardization activities have defined higher quality twisted-pair options by specifying more details on FEXT as well as NEXT to guide manufacturers to design these cables. Today, the most popular of these twisted-pair options are enhanced Cat-5 (Cat-5e) with the same bandwidth as Cat-5 but more specified FEXT behavior, Cat-6 with 250 MHz

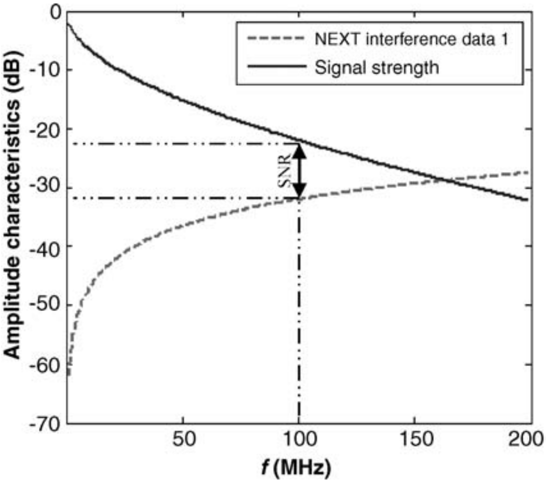


FIGURE 2.4 Amplitude characteristics of the Cat-5 twisted-pair wires; the solid line is the 100 m cable attenuation and the dashed line shows the interference.

bandwidth, and Cat-7 with up to 600 MHz of bandwidth. As we explain in Chapter 8, the IEEE 802.3 community specifies physical layer options which can be used with these media options to implement data rates of up to 100 Gb/s.

2.2.2 Coaxial Cables

Coaxial cables have been the most popular wiring used for home access for broadband multimedia services. The cable TV installations that started in the late 1960s were carrying multiple analog TV channels over the coaxial cables. More recently, the availability of such a broadband access to the home has motivated the invention of broadband modems for Internet access, and today cable TV providers are adding telephone services to their cable connections to the home. In the office environment, the so-called "thick cable" was used first in the 1970s to implement legacy LANs, such as Ethernet, token ring, and token bus recommended by the IEEE 802 standardization community. The second-generation wired LAN industry resorted to the more flexible lower bandwidth "thin cables" in the mid 1980s to improve the wiring difficulties facing the "thick cables," before they finally resorted to twisted-pair wirings in the early 1990s. Since the resistance in metallic conductors is inversely proportion to the diameter of the conductor, the thinner cables have more resistance, causing a higher attenuation rate. The IEEE 802.3 recommendation for the maximum length of the less-flexible thick cable for Ethernet was 500 m, while the more flexible thin cables were recommended to be used for distances up to 200 m. The recommended length of the most flexible twisted-pair wiring recommended by this standard is 100 m. In other words, in the early days the guided LAN networking community determined a practical compromise between a suitable length of coverage and a comfortable rigidity of the cable for ease of wiring. The cost of installation and relocation of wiring for LANs has been an important factor behind the growth of guided media LAN and the ultimate emergence of the WLAN industry.

Figure 2.5 provides the details of a coaxial cable construction. A coaxial cable consists of a round conducting wire, surrounded by insulation, embraced by a cylindrical braided conductor, and covered by a final plastic sheath as a jacket. In ideal conditions, this arrangement keeps the electromagnetic field carrying the signal only in the space between the inner and outer conductors, isolating the transmission of the signal from outside electromagnetic signals in the air from different sources. Therefore, an ideal cable does not interfere with or suffer interference from external electromagnetic fields. The general design concept of the coaxial cable which controls the interference provides for a medium for high-frequency and longer distance transmissions to distribute broadband signals such as multichannel TV stations in neighborhoods.

As we explained in the previous section with twisted-pair wiring operation, the main source of transmission impairment is interference caused by other transmission lines in the near and far end connectors. The connectors used in cables follow the same inner and outer connections, which controls the harmful effects of radiation around the connectors. Therefore, the general propagation characteristics of coaxial cables follows the same pattern as the path gain of twisted-pair wirings, while the NEXT and FEXT interference are almost eliminated.

The most commonly recognized cable in the industry is the so called RG-6, which is used for distribution of cable TV signals to homes and many other applications in commerce. RG means radio guide, and it was originally a military specification for cables carrying radio signals to the antenna or other parts of a radio system, though it is not used any more.

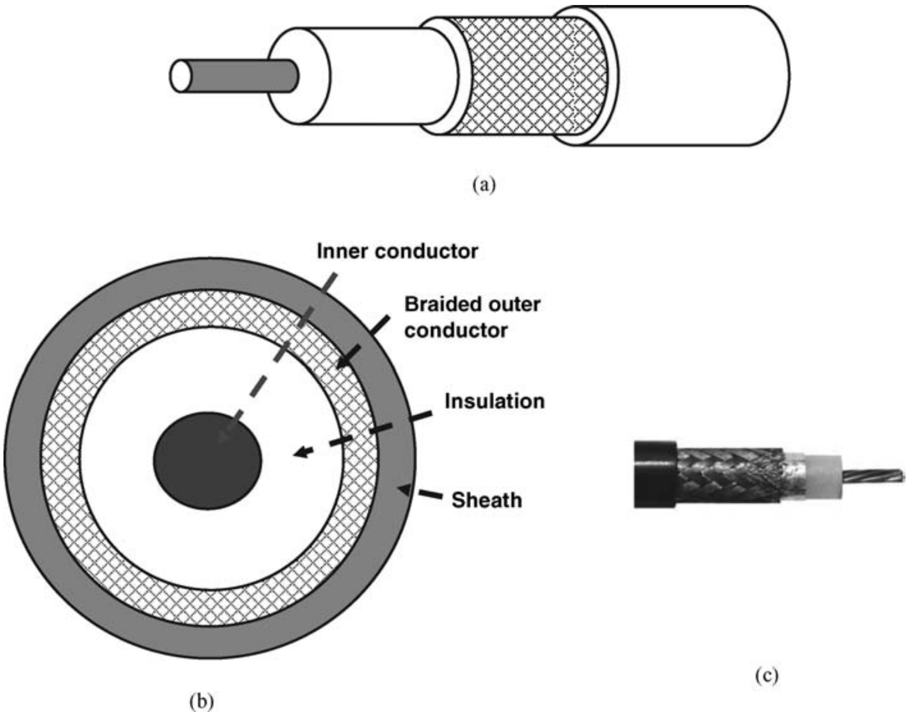


FIGURE 2.5 Coaxial cable wiring: (*a*) basic coaxial concept; (*b*) details inside a coaxial cable; (*c*) photograph of a sample coaxial cable.

Among the variety of RG coaxial cables, we use RG-6 as an example. RG-6 coaxial cables typically have a copper-coated steel conductor as the center metal wire and a combination of aluminum foil and aluminum braid shield as the outer conductor. Better grades of cables used in professional video applications have a more dense copper braid. The path gain of the RG-6 can be approximated by

$$G_{\mathrm{p}}(f, l) = -0.0067\sqrt{f}l \qquad (2.5)$$

Figure 2.6 compares the amplitude characteristics of the RG-6 coaxial cable, given by Eq. (2.5), with the Cat-5 twisted pair in Eq. (2.3) for a cable length of 100 m. For the same 22 dB loss observed at 100 MHz for the twisted pair we can have a coaxial cable with an order of magnitude wider bandwidth of more than 1 GHz. If we keep the bandwidth at 100 MHz, then the coaxial cable loss is 6.7 dB compared with the twisted-pair loss of 22 dB. If we increase the length of the cable three times, then we have $3 \times 6.7 = 20.1$ (db), which still provides a better performance than Cat-5 twisted pair. Figure 2.7 compares the performance of the RG-6 cables with the Cat-3 voice-grade twisted-pair wires. For the distance of 100 m used in IEEE 802.3 LANs and 10 MHz of bandwidth, Cat-3 has an attenuation of around 10 dB. For the same attenuation, an RG-6 coaxial cable can run for 500 m. This fact governs the assignment of 500 m for the original Ethernet using "thick cables" and the selection of 100 m as the length of the first twisted-pair star Ethernet.

Coaxial cables have different levels of thickness (between 1 and 2.5 cm) and flexibility. The more flexible cables have a braided sheath and are usually used with thin copper wires.

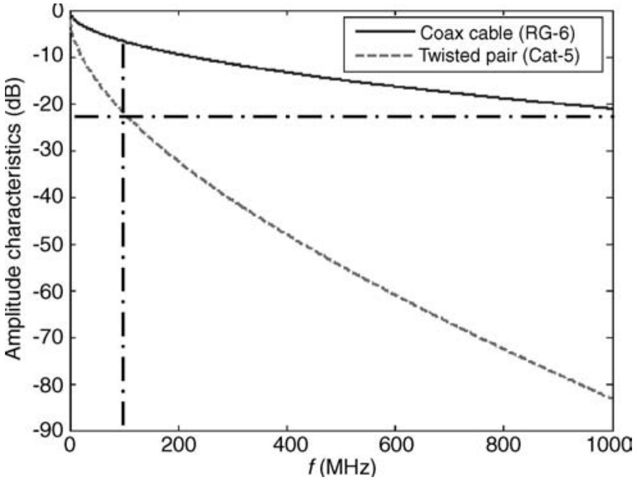


FIGURE 2.6 Amplitude characteristics of RG-6 coaxial cable and Cat-5 twisted-pair wiring for a cable with length of 100 m.

The thicker cables are less flexible and have a thicker inner copper. Since the resistance of the conductors is inversely proportional to the thickness of the inner wire, the more flexible thin cables have shorter coverage. For example, the "thick cable" used in the original legacy Ethernet can cover up to 500 m, while the "thin cable" adopted later on for the second generation of cabled Ethernet could cover up to 200 m. Connections to the ends of coaxial cables are usually made with RF connectors. The most popular RF connectors used with cables are the BNC connectors (derived from bayonet, Neill, and Concelman) and F-connectors, which are shown in Fig. 2.8.

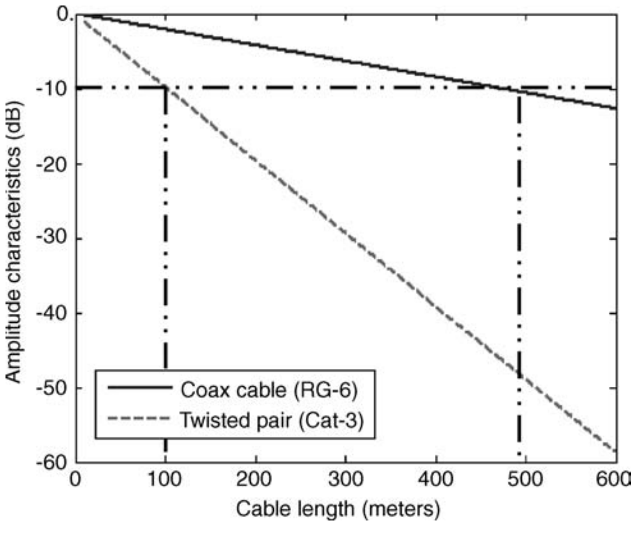


FIGURE 2.7 Amplitude characteristics of RG-6 coaxial cable and Cat-3 twisted-pair wiring at a frequency of 10 MHz.

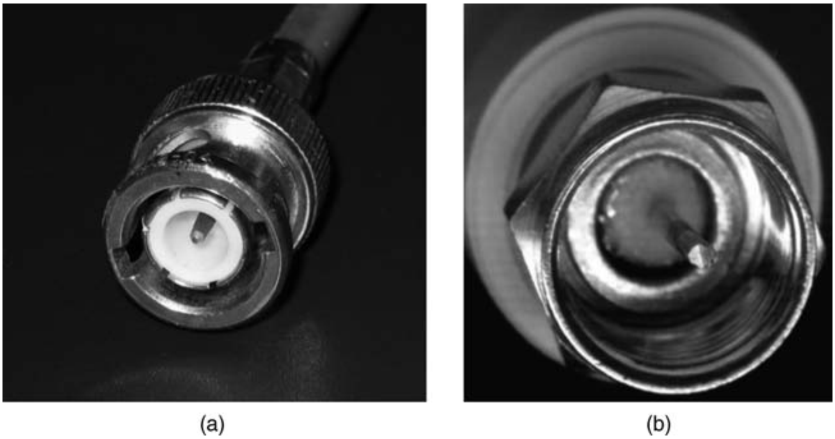


FIGURE 2.8 Cable connectors: (*a*) BNC connector; (*b*) F-connector.

2.2.3 Optical Fiber

Optical fiber is the medium of choice for the backbones of modern networks and it is rapidly penetrating the multimedia home access market in competition with the access by cable provided by the cable TV industry. In the office environment, optical fiber lines are connecting the backbone of LANs, and many modern buildings lay optical fiber lines in anticipation of fiber connections to the desktop computers which are currently connected by Cat-5 wirings. Fiber lines provide for extremely higher transmission rates, longer coverage, lighter weight, smaller size, and they are free from RF interference.

The optical fiber cable uses a glass or a plastic or a combination of the two fibers to confine and guide light along its length. The glass provides for lower optical attenuations and it is used for longer distance telecommunication applications. Figure 2.9 shows the general picture for optical fiber cables. The core of the fiber is used for guiding the light, the cladding is used to confine the light inside the fiber core, and a plastic jacket protects the cable. If the diameter of the core fiber is large, then the confinement is based on total internal reflection. In such a case, the fiber is referred to as a multimode fiber and it is used for shorter distance communications of up to a couple of hundred meters. The single area of contact is larger and the light is not directed. Inexpensive light emitting diodes (LEDs) are used with multimode fibers to carry the signal over multiple paths. Single-mode optical fibers using thinner cores and sharper laser diodes are used for longer distances. Figure 2.10 shows the basic concept behind single- and multi-mode optical fiber communications. To design multimode and single-mode optical fibers operating at different wavelengths, a number of different designing techniques, such as graded index or step index, are used.

Optical signals are similar to electromagnetic signals, but the frequency of operation is extremely high (on the order of 10^{14} Hz). It is customary in industry to use the wavelength $\lambda = c/f$, where f is the frequency of operation and c is the speed of light, to identify a specific band. Figure 2.11 illustrates the gain versus wavelength of an ensemble of fiber-optic cable material types. The bumps in the performance are caused by chemical reactions during the manufacturing process and cannot be avoided. This relationship does not follow the monotonic behavior of twisted-pair or coaxial cables and shows that certain wavelengths cause less attenuation, providing a better opportunity for information

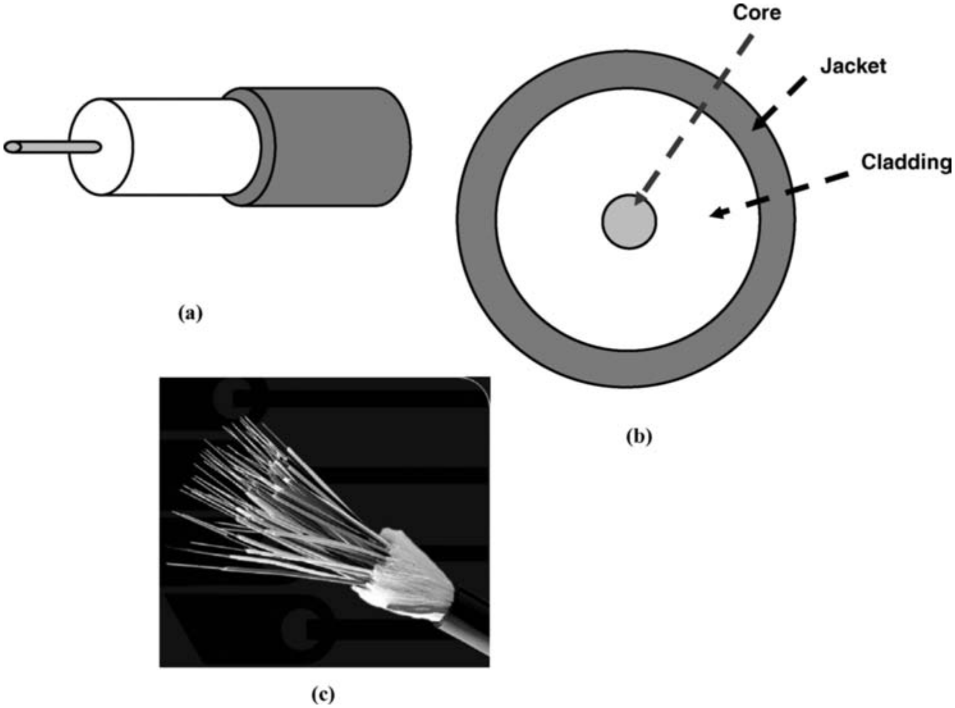


FIGURE 2.9 Optical-fiber cable: (*a*) basic concept; (*b*) details inside a fiber line; (*c*) photograph of a sample fiber bundle.

transmission. The four wavelengths of 850, 1300, 1310, and 1550 nm are the most popular in industrial applications. The lower wavelengths are used for multimode fibers and the upper wavelengths for single-mode fibers.

The core of the fiber-optic cable and its cladding have extremely small diameter sizes which is often expressed in micrometers (1 μm = 10^{-6} m). To appreciate the size of these diameters, it is helpful to point out that human hair has a diameter of around 100 μm. In industry, fiber-optic cable sizes are usually expressed by first giving the core size followed by the cladding size; for example, 50/125 μm indicates a core diameter of 50 μm and a cladding diameter of 125 μm.

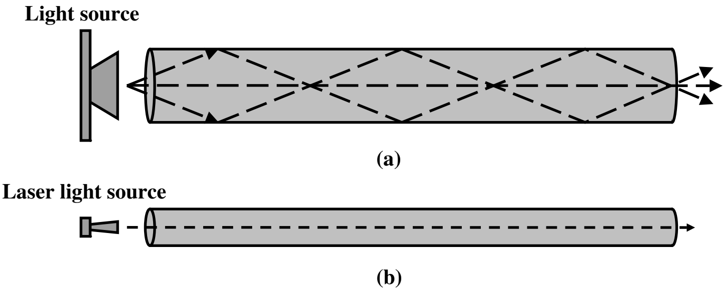


FIGURE 2.10 Light source and optical fiber transmission modes: (*a*) multimode with multiple reflecting lights; (*b*) single mode with narrow fiber and laser light.

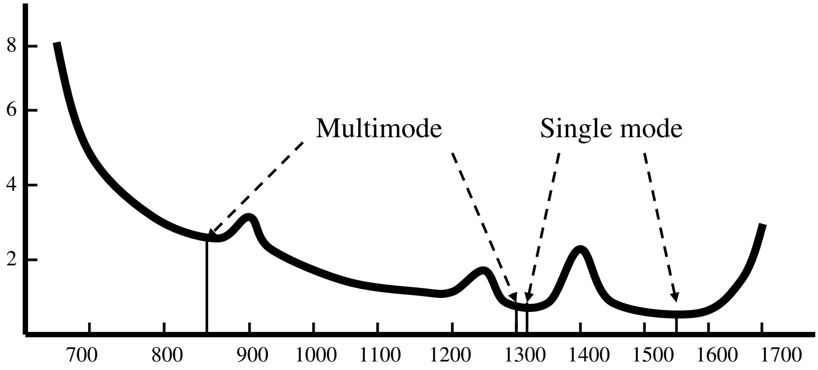


FIGURE 2.11 Attenuation versus wavelength for an ensemble of optical fiber lines.

The wire specification standardization organizations identify the mode, diameters, wavelength, and attenuation as a guideline to the manufacturers. For example, EIA/TIA specifies attenuations of 3.5 dB/km and 1.5 dB/km for 850 nm and 1300 nm multimode fibers respectively. This specification is defined for both 50/125 and 62.5/125 μm optical cables. The same standardization committee specifies single-mode fibers with 9 μm diameters at wavelengths 1310 nm and 1550 nm to have attenuations of 0.4 dB/km and 0.3 dB/km. The TIA/EIA also specifies the connector loss and splice loss of the fiber-optic cables to be 0.75 dB and 0.1 dB respectively. These numbers can be used to specify the length of the fiber-optic cable for a specific application.

Example 2.1: Optical Budget with an Optical Fiber in Gigabit Ethernet The IEEE 802.3 standardization committee for gigabit Ethernet recommends a maximum fiber-optic cable length of 700 km for single-mode 9/125 μm diameters at 1550 nm. Using the 0.3 dB/km loss recommended by the EIA/TIA, we have a cable loss of $700 \times 0.3 = 21$ dB for the cable. Adding connector losses of $0.75 \times 2 = 1.5$ for the two connectors at the ends of the cables results in a minimum power loss of 22.5 dB. A typical laser transmitter has -8 dBm transmitter power and its associated receiver has a minimum received signal power of -34 dBm for proper operation, allowing a maximum power loss of $-8 - (-34) = 26$ dB for the medium. This arrangement recommended by the IEEE standard allows $26 - 22.5 = 3.5$ dB as a safety margin. Usually, there is a loss of 0.1 dB per splice. If we have five splices in this long cable, then we will have a 3 dB safety margin. The difference between the transmitter power and the receiver sensitivity of 26 dB is referred to as the optical budget of the transmission system.

Example 2.2: Optical Fiber in 10 Mb/s Ethernet The IEEE 802.3 committee for legacy Ethernet operating at 10 Mb/s recommends using a maximum of 2 km of 50/125 μm multimode fiber-optic line at 850 nm. The path loss per kilometer for this fiber is 3.5 dB, resulting in a total of $3.5 \times 2 + 0.75 \times 2 = 8.5$ dB for the cable and connectors. Considering a 3.5 dB safety margin, a less expensive lower grade LED and photosensitive device with an optical budget of 12 dB can be used for the implementation of the system instead of the higher quality, more expensive laser beams used in the previous example with an optical budget of 26 dB.

2.3 WIRELESS MEDIA

In the previous section we analyzed the behavior of the guided media by relating the path loss to distance and frequency of operation. The basic relation between path loss, distance, and frequency of operation in the wireless medium and in free space is also relatively simple. However, wireless networks, such as cellular telephones or WLANs, operate in urban and indoor areas where signal is subject to multipath, which causes fading in the received signal that complicates the channel behavior. Indeed, this behavior is so complicated and environment dependent that we need to resort to statistical modeling. As a result, while we were describing the behavior of a guided medium with a simple path-loss equation that was deterministic and environment independent, here we need several complex statistical models to represent the behavior in different application environments. Since these models are mostly developed by telecommunication engineers with an electrical engineering background, a major question raised by a computer networking professional would be "Why should we study radio propagation?"

An understanding of radio propagation is essential for coming up with appropriate design, deployment, and management strategies for any wireless network. In effect, it is the nature of the *radio channel* that makes wireless networks far more complicated than their wired counterparts. Radio propagation is heavily site specific and can vary significantly depending on the terrain, frequency of operation, velocity of the mobile terminal, interference sources, and other dynamic factors. Accurate characterization of the radio channel through key parameters and a mathematical model is important for predicting signal coverage, achievable data rates, specific performance attributes of alternative signaling and reception schemes, analysis of interference from different systems, and for determining the optimum locations for installing transmitting and receiving antennas.

The three most important radio propagation characteristics used in the design, analysis, and installation of wireless information networks are the achievable signal coverage, the maximum data rate that can be supported by the channel, and the rate of fluctuations in the channel [Pah05]. The achievable signal coverage for a given transmission power determines the size of a cell in a cellular topology and the range of operation of a BS transmitter. This is usually obtained via empirical *path*-loss models obtained by measuring the received signal strength (RSS) as a function of distance. Most of the path-loss models are characterized by a distance–power or path-loss *gradient* and a random component that characterizes the fluctuations around the average path loss due to shadow fading and other reasons. For efficient data communications, the maximum data rate that can be supported over a channel becomes an important parameter. Data rate limitations are influenced by the multipath structure of the channel and the fading characteristics of the multipath components. This will also influence the signaling scheme and receiver design. Another factor that is intimately related to the design of the adaptive parts of the receiver, such as timing and carrier synchronization, phase recovery, and so on, is the rate of fluctuations in the channel, usually caused by movement of the transmitter, receiver, or objects in between. This is characterized by the *Doppler spread* of the channel. We consider path-loss models in detail in this section and provide a summary of the effects of multipath and Doppler spread in subsequent subsections of this chapter.

Depending on the data rates that need to be supported by an application and the nature of the environment, certain characteristics are much more important than others. For example, signal coverage and slow fading are more important for low data-rate narrowband systems, such as cordless telephones, low-speed data, and cellular voice telephony.

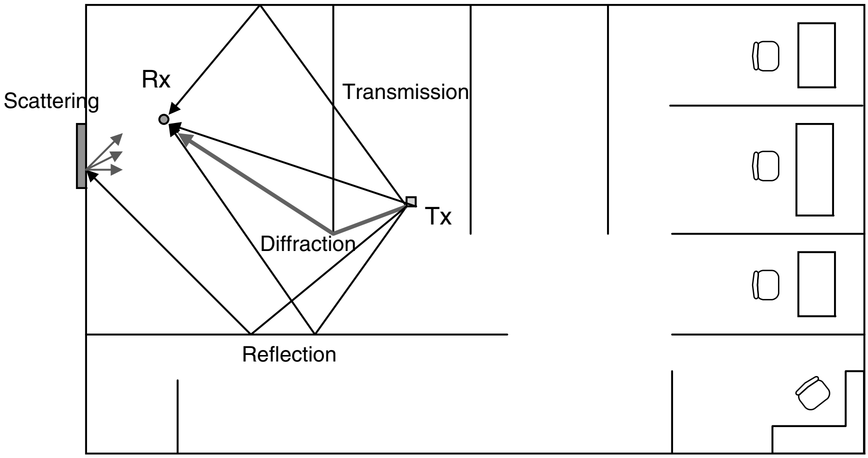


FIGURE 2.12 Radio propagation mechanisms in an indoor area.

2.3.1 Radio Propagation Mechanisms

Radio signals with frequencies used in wireless networks have small wavelengths compared with the dimensions of building features, so that electromagnetic waves can be treated simply as rays [Ber94]. This means that ray-optical methods can be used to describe the propagation within and even outside buildings by treating electromagnetic waves as traveling along localized ray paths. The fields associated with the ray paths change sequentially based upon the features of the medium that the ray encounters.

In order to describe radio propagation with ray optics, three basic mechanisms are generally considered, while ignoring other complex mechanisms. These mechanisms are illustrated in Figs 2.12 and 2.13 for indoor and outdoor applications respectively.

Reflection and Transmission. Specular reflections and transmissions occur when electromagnetic waves impinge on obstructions larger than the wavelength. Usually, rays incident upon the ground, walls of buildings, the ceiling, and the floor undergo specular reflection and transmission, with the amplitude coefficients usually determined by plane-wave analysis. Upon reflection or transmission, a ray attenuates by factors that depend on the

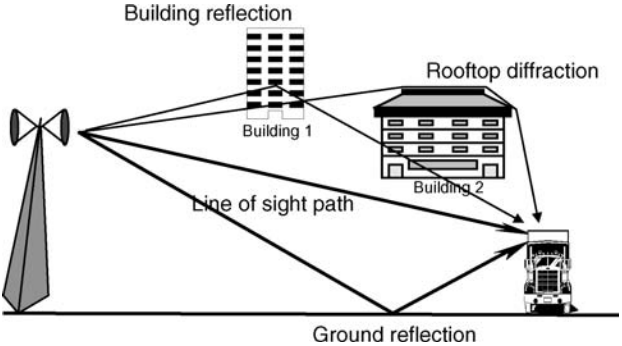


FIGURE 2.13 Radio propagation mechanisms in outdoor areas.

frequency, the angle of incidence, and nature of the medium (its material properties, thickness, homogeneity, etc.). These mechanisms often dominate radio propagation in indoor applications. In outdoor urban area applications, the transmission mechanism often loses its importance, because it involves multiple-wall transmissions that reduce the strength of the signal to negligible values.

Diffraction. Rays that are incident upon the edges of buildings, walls and other large objects can be viewed as exciting the edges to act as a secondary line source. Diffracted fields are generated by this secondary wave source and propagate away from the diffracting edge as cylindrical waves. In effect, this results in propagation into *shadowed* regions, since the diffracted field can reach a receiver, which is not in the line of sight of the transmitter. Since a secondary source is created, it suffers a loss much greater than that experienced via reflection or transmission. Consequently, diffraction is an important phenomenon outdoors (especially in microcellular areas), where signal transmission through buildings is virtually impossible. It is less consequential indoors, where a diffracted signal is extremely weak compared with a reflected signal or a signal that is transmitted through a relatively thin wall.

Scattering. Irregular objects, such as wall roughness and furniture (indoors) and vehicles, foliage and the like (outdoors), scatter rays in all directions in the form of spherical waves. This occurs especially when objects are of dimensions that are on the order of a wavelength or less of the electromagnetic wave. Propagation in many directions results in reduced power levels, especially far from the scatterer. As a result, this phenomenon is not that significant unless the receiver or transmitter is located in a highly cluttered environment. This mechanism dominates diffuse IR (DFIR) propagations when the wavelength of the signal is so high that the roughness of the wall results in extensive scattering. In satellite and mobile radio applications, tree leaves and foliage often cause scattering.

2.3.2 Path-Loss Modeling and Signal Coverage

Calculation of signal coverage is essential for design and deployment of wireless networks. In wired transmissions we needed models for loss of the signal strength in the media to find the maximum length of the cable for effective operation. In wireless networks, signal coverage is influenced by a variety of factors, prominently the radio frequency of operation and the terrain. Often, the region where a wireless network is providing service spans a variety of terrain. An operation scenario is defined by a set of operations for which a variety of distances and environments exist between a transmitter and the receiver. As a result, a unique channel model cannot describe radio propagation between the transmitter and the receiver, and so we need several models for a variety of environments to enable system design. The core of the signal coverage calculations for any environment is a path-loss model that relates the loss of signal strength to distance between two terminals at a given frequency. Using the path-loss models, radio engineers calculate the coverage area of wireless BSs and access points (APs), as well as maximum distance between two terminals in an adhoc network. In the following we consider path-loss models developed for several such environments that span different cell sizes and the terrain in the cellular hierarchy used for deployment of wireless networks.

Free-Space Propagation. In previous sections we discussed the relation between the guided media length l and the loss of the cable and provided equations to relate this relation to the frequency. We start our path-loss modeling for wireless networks with path loss in free space. Here, rather than length of the cable, we relate the loss to the distance d between the transmitter and the receiver. All of our models for the guided media were relating the power loss in decibels to the length of the cable l with a linear function. In the case of radio propagation, the relation between the path loss in decibels and the distance is no longer linear; instead, it follows a logarithmic function.

It is well known in the antenna propagation literature that, for an ideal isotropic antenna with no heat loss [Frii46] which radiates signal strength in all directions at the same rate, the radiation intensity is given by $\lambda^2/4\pi$. At a sphere of radius d the total radiated signal strength is divided by the area of the sphere, $4\pi d^2$, resulting in a path gain of

$$G_{\mathrm{p}}(\lambda, d) = \left(\frac{\lambda}{4\pi d}\right)^2 \tag{2.6}$$

where $\lambda = c/f$ is the wavelength of the carrier, c is the speed of light in vacuum (3×10^8 m/s), and f is the frequency of the radio carrier. Therefore, in free space, Eq. (2.6), signal strength falls with the square of the distance and the power path loss after a distance d in meters is proportional to d^2. The transmission delay as a function of distance is given by $\tau = d/c = 3d$ ns or 3 ns/m.

In the radio propagation literature, it is customary to express the channel attenuation behavior in terms of path loss, which is the inverse of the path gain. Taking 10 times the logarithm of quantities in Eq. (2.6), the path loss in free space, expressed in decibel form, is

$$L_{\mathrm{p}} = L_0 + 20\log_{10}(d) \tag{2.7}$$

where L_0 is the path loss in the first meter of distance and it is given by

$$L_0 = 20\log_{10}\left(\frac{4\pi}{\lambda}\right) \tag{2.8}$$

Equation (2.7) shows that the path loss in free space has a fixed component L_0 which increases as the frequency increases (wavelength decreases) and a second component which causes an attenuation of 20 dB/decade (or 6 dB/octave) of the distance.

Example 2.3: Path Loss in Wireless LANs The transmitted power of WLANs operating at 2.4 GHz is 100 mW (20 dBm). The wavelength of these devices is

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{2.4 \times 10^9} = 0.125\ \mathrm{m} = 12.5\ \mathrm{cm}$$

Therefore, the path loss in the first meter is

$$L_0 = 20\log_{10}\left(\frac{4\pi}{\lambda}\right) = 20\log_{10}\left(\frac{4\pi}{0.125}\right) = 40.05\ \mathrm{dB}$$

The overall equation for the path-loss calculation of these devices in free space is then given by

$$L_{\mathrm{p}} = L_0 + 10\alpha\log_{10}(d) = 40.05 + 20\log_{10}(d)$$

At 10 m we have 40.05 + 20 = 60.05 dB loss and at 100 m we have 80.05 dB path loss.

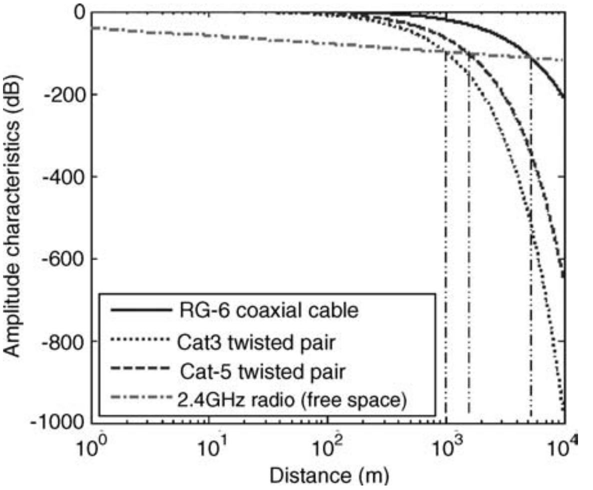


FIGURE 2.14 Attenuation versus distance for different guided media versus 2.4 GHz radio propagation in free space.

Figure 2.14 shows the attenuation versus distance for RG-6 coaxial cable and Cat-3 and Cat-5 twisted-pair wires versus 2.4 GHz radio propagation in free space. At shorter distances, radio transmissions lose significant power, but over a longer distance, since the path loss in decibels is exponentially related to distance, radio transmissions will perform better. Since loss of the power in a wireless medium is logarithmically related to the distance, when we show the distance in a logarithmic basis this relation is linear; this trend of presentation is common in the radio path-loss modeling literature. If we increase the distance exponentially, then the loss in any guided media will appear as a waterfall curve that ultimately crosses the loss of the radio at certain distance. At around 1 km, the path loss with a radio transmission becomes smaller than the loss over a Cat-3 wire, at 1.6 km smaller than the loss over a Cat-5 wire, and at 5 km better than the loss with an RG-6 coaxial cable. Therefore wireless media have more power loss than wired media at shorter distances from the transmitter, but form a more power-efficient solution over longer distances. But over longer distances we need to compensate the heavy power loss either by a high-power transmission or by designing more sensitive receivers. The difference between the transmitted power and the sensitivity of the receiver, which is sometimes referred to as the link budget for cable-based modems, is on the order of 20–30 dB, while in wireless media it can easily go over 100 dB. For example, IEEE 802.11 devices transmit at 20 dBm (100 mW) and the receiver sensitivity is around −80 to −95 dBm, providing for a link budget on the order of 100–115 dB. As shown in Fig. 2.14, this link budget can support operations up to several kilometers in free space. However, if we are around 100 m from an IEEE 802.11 AP, most of us have the experience that we do not have much of a chance of saving our connections. This is because of the extra attenuation caused by multipath arrivals from reflection off or loss through walls and objects in and around the transmitter and the receiver, which ultimately change the distance power gradient.

Distance–Power Relationship. In most environments, it is observed that the averaged received radio signal strength falls as some power of the distance α called the power–distance gradient or path-loss gradient. That is, the path loss after a distance d in meters is d^{α}.

For the free-space propagation described in the previous section the distance power gradient is $\alpha = 2$which indicates that the signal strength falls as the square of the distance between the transmitter and the receiver. Considering this general relation between the path loss and the distance, in the radio propagation literature the path loss in decibels is expressed by the general relation

$$L_p = L_0 + 10 \times \alpha \log_{10}(d) \tag{2.9}$$

which reduces to Eq. (2.7) for free space when we let $\alpha = 2$. In urban areas, typically $\alpha = 4$ is used for calculation of the coverage of cellular networks. In indoor areas α can take values from less than 2 up to 6 [Pah05]. Equation (2.9) again presents the total path loss as the path loss in the first meter L_0 plus the power loss relative to the power received at 1 m, i.e. $10 \times \alpha \log_{10}(d)$. For a one decade increase in distance the power loss is 10α dB, and for a one octave increase in distance it is 3α dB. For a free-space path, the power loss is 20 dB/decade or 6 dB/octave of distance as already discussed. In urban areas, with $\alpha = 4$ the attenuation is 40 dB/decade or 12 dB/octave. The received power in decibels is the transmitted power plus transmitter antenna gain in decibels minus the total path loss L_p plus receiver antenna gain.

Path-loss models of this form are extensively used for deployment of cellular networks. The coverage area of a radio transmitter depends on the power of the transmitted signal and the path loss. Each radio receiver has a particular power sensitivity, i.e. it can only detect and decode signals with strength larger than this sensitivity. Since the signal strength falls with distance, one can calculate the coverage using the transmitter power, the path-loss model, and the sensitivity of the receiver.

Example 2.4: Coverage of a WLAN in Free Space and Urban Areas The transmitted power and the receiver sensitivity of a WLAN operating at 2.4 GHz are 20 dBm and −80 dBm respectively. If the transmitter and the receiver antenna gains are 3 dB and 1 dB, then the maximum acceptable path loss or link budget is

$$L_p = 20\,\text{dBm} - (-80\,\text{dBm}) + 3\,\text{dB} + 1\,\text{dB} = 104\,\text{dB}$$

The path loss at 1 m for 2.4 GHz calculated in Example 2.3 was $L_0 = 40.05$ dB; using Eq. (2.9), the coverage of this system for free space, $\alpha = 2$, is

$$d = 10^{(104-40.05)/20} = 1576\,\text{m}$$

The coverage of the same system in urban areas with $\alpha = 4$ is

$$d = 10^{(104-40.05)/40} = 40\,\text{m}$$

The substantial difference in the coverage in different areas is caused by the change of distance–power gradient from 2 in free space to 4 in urban areas. Therefore, the distance–power gradient plays an important role in calculations of the coverage of wireless networks. The value of this gradient changes drastically in different environments and is often calculated empirically.

To measure the gradient of the distance–power relationship in a given area, the receiver is fixed at one location and the transmitter is placed at a number of locations with different distances between the transmitter and the receiver. The received power or the path loss in decibels is plotted against the distance on a logarithmic scale. The slope of the best-fit line through the measurements is taken as the gradient of the distance–power relationship.

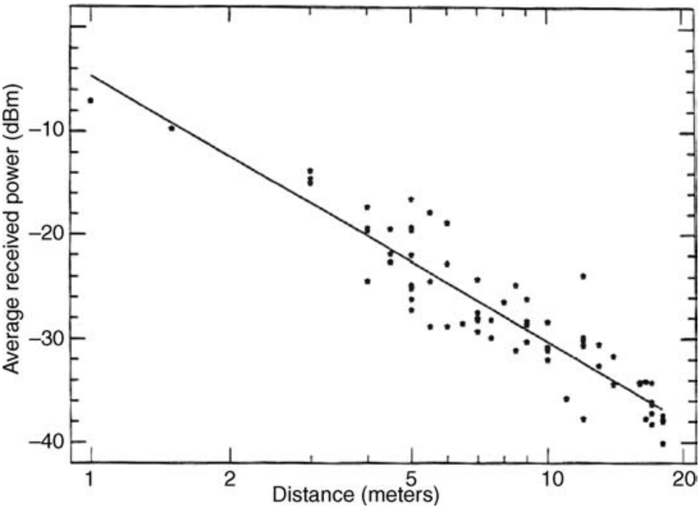


FIGURE 2.15 Measured received power and a linear regression fit to the data.

Figure 2.15 shows a set of measured data taken in an indoor area at distances from 1 to 20 m, together with the best-fit line through the measurements.

Shadow Fading. Depending on the environment and the surroundings, and the location of objects, the RSS *for the same distance* from the transmitter will be different. In effect, Eq. (2.9) provides the mean or median value of the signal strength that can be expected if the distance between the transmitter and receiver is *d*. The actual RSS will vary around this mean value. This variation of the signal strength due to location is often referred to as *shadow fading* or *slow fading*. The reason for calling this shadow fading is that, very often, the fluctuations around the mean value are caused due to the signal being blocked from the receiver by buildings (in outdoor areas), walls (inside buildings), and other objects in the environment. It is called slow fading because the variations are much slower with distance than another fading phenomenon caused by multipath, which we discuss later. It is also found that shadow fading has less dependence on the frequency of operation than multipath fading or fast fading, as discussed later. The path loss, Eq. (2.9), will have to be modified to include this effect by adding a random component as follows:

$$L_{\mathrm{p}} = L_0 + 10\alpha\log_{10}(d) + X \tag{2.10}$$

Here, *X* is a random variable with a distribution that depends on the fading component. Several measurements and simulations indicate that this variation can be expressed as a log-normally distributed random variable. A log-normal absolute fading component ends up as a zero-mean Gaussian fading component in decibels. Therefore, the distribution function of X is given by

$$f_{\mathrm{SF}}(X) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{X^2}{2\sigma^2}\right) \tag{2.11}$$

where σ is the standard deviation of the RSS in decibels.

The problem caused by shadow fading is that all locations at a given distance may not receive sufficient signal strength for correctly detecting the information. In order to achieve sufficient signal coverage, the technique employed is to add a *fade margin* to the path loss or RSS. The fade margin is usually taken to be the additional signal power that can provide a certain fraction of the locations at the edge of a cell (or near the fringe areas) with the required signal strength. For computing the coverage we thus employ the following equation:

$$L_{\mathrm{p}} = L_0 + 10\alpha \log d + F_\sigma \tag{2.12}$$

where F_σ is the fade margin associated with the path loss to overcome the shadow fading component.

The distribution of X in Eq. (2.11) is employed to determine the appropriate fade margin. Since X is a zero-mean, normally distributed random variable that corresponds to log-normal shadow fading, this means that, at the fringe locations, the mean value of the shadow fading is 0 dB. Half of the locations have a positive fading component and half of the locations have a negative fading component. In Eq. (2.10), this will mean that the locations that have the positive fading component X will suffer a larger path loss, resulting in unacceptable signal strength. To overcome this, a fading margin is employed to move most of these locations to within an acceptable RSS value. This fading margin can be applied by increasing the transmit power and keeping the cell size the same, or reducing the cell size by setting a higher threshold for making a handoff. In mathematical terms, for γ% coverage the BS should have an additional fade margin of

$$1-\gamma = 0.5\,\mathrm{erfc}\left(\frac{F_\sigma}{\sigma\sqrt{2}}\right) = \int_{F_\sigma}^{\infty} f_{\mathrm{SF}}(x)\,\mathrm{d}x$$

Example 2.5: Computing the Fading Margin A mobile system is to provide 95% successful communication at the fringe of coverage with a location variability having a zero-mean Gaussian distribution with standard deviation of 8 dB. What fade margin is required?

Solution. Note that the location variability component X (dB) in this case is a zero-mean Gaussian random variable. In this example, the variance of X is 8 dB. We have to chose F_σ such that $Q(x)$ (see footnote) is 0.05, i.e. 95% of the locations will have a fading component smaller than the tolerable value. Using the complementary error function and a software package like Matlab, we can determine the value of F_σ as the solution to the equation $0.05 = 0.5\,\mathrm{erfc}(F_\sigma/\sigma\sqrt{2})$. For this example, the fade margin to be applied is 13.2 dB.[1]

So far, we have discussed achievable signal coverage in terms of the RSS and the path loss. In the following sections, we discuss parameters and path loss models for a variety of cellular environments. We will also discuss, where relevant, the important factors that lead to these path loss models.

[1] The function $Q(x) = \int_x^\infty f_X(x)dx = 0.05$, where $Q(x)$ is the probability that the normal random variable X has a value greater than x is tabulated or it can be determined using the complementary error function via the relation $Q(x) = 0.5\mathrm{erfc}(\mathrm{x}/\sqrt{2})$.

TABLE 2.1 Partition-dependent Losses

Attenuating material	Signal attenuation of 2.4 GHz (dB)
Window in brick wall	2
Metal frame, glass wall into building	6
Office wall	6
Metal door in office wall	6
Cinder wall	4
Metal door in brick wall	12.4
Brick wall next to metal door	3

2.3.3 Path-Loss Models for Indoor Areas

Path-loss models for indoor areas are employed for WLAN, wireless PCSs, and cordless telephones. Many researchers have performed measurements within buildings, primarily to determine the distance–power relationship and arrive at empirical path loss models for a variety of environments [Pah05]. We discuss some of these models below.

Wall-Partitioned Path-Loss Models. The simplest path-loss model expressing the behavior of the signal strength in indoor areas is the *partition-dependent* model [Rap03]. This model fixes the value of the path-loss gradient α at 2 for free space and introduces losses for each partition that is encountered by a straight line connecting the transmitter and the receiver. Therefore, the path loss is given by

$$L_p = L_0 + 20 \log d + \sum m_{\text{type}} w_{\text{type}} \tag{2.13}$$

where m_{type} refers to the number of partitions of that type and w_{type} the loss in decibels attributed to such a partition. Table 2.1 shows some decibel loss values measured at Harris semiconductors at 2.4 GHz for different types of partition. Longer tables are available in Rappaport [Rap03]. Appropriate fading margins have to be included to account for the variability in path loss for the same distance d.

Example 2.6: Using the same parameters as in Example 2.4, the coverage of the transmitter using the partitioned model when we have four office walls is given by

$$d = 10^{[104-40.05-(4\times 6)]/20} = 99\text{ m}$$

Comparing with the results of Example 2.4, here we have a coverage closer to 40 m that we obtained using simply a single gradient of $\alpha = 4$ and no wall losses.

This model is very simple and intuitive. It assumes a single path between the transmitter and the receiver with the free-space path loss; to add the effects of building layout, it adds additional fixed path losses per existing wall in between. There are some problems with this simple modeling. We need long tables to include all possibilities for different materials used in construction of the walls and the details of doors and windows. We also need to have specific information on the number of walls between the transmitter and the receiver, and it only considers the direct path, while in indoor areas we have numerous paths between the transmitter and the receiver. To address the first issue, the reader can find a long table of losses for different types of wall in Rappaport [Rap03]. The second and third issues are addressed using ray-tracing software, where we have the map of the building and we can use a computer simulation to find direct and all reflected paths using similar principles.

IEEE 802.11 Distance-Partitioned Models. In indoor areas, the area between the transmitter and the receiver is often not homogeneous with a single distance–power gradient. Results of wideband measurements in a partitioned indoor area show significant differences among the values of the distance–power gradient in different parts of a building. Figure 2.16 [Gan91] depicts the middle part of the third floor of the Atwater Kent Laboratories at WPI. The receiver is located at the center of Room 317, and the transmitter is moved to different locations in various rooms for measurements of the RSS. The area is divided into three segments: the interior of a small laboratory (Room 317), corridors around the laboratory, and offices on the opposite side of the corridor. Three different gradients of 1.76, 2.05, and 4.21 were calculated from the results of the measurements made in the three subareas. Inside the small laboratory all the locations provide a strong LOS connection and the gradient is 1.76, which is less than the free-space gradient. In the corridors there is at least one plaster wall with metal studs between the transmitter and receiver, and the gradient is close to that of free-space propagation. The third sub-area, with gradient 4.21, includes at least two walls, one of which contains a number of metal doors. Also, inside the rooms are several metal shelves, cabinets, and desks.

Most of the recent path-loss models developed for WLAN and WPAN applications consider this phenomenon by partitioning the path-loss behavior in different segments with multiple distance–power gradients each associated with a segment of the path between the transmitter and the receiver. Here, we describe the IEEE 802.11 recommended path-loss model. Figure 2.17 shows the basics of this model, in which the distance between the transmitter and the receiver is divided into two segments at a break-point distance d_{bp}. The distance–power gradients in the two segments are $\alpha_1 = 2$ and $\alpha_2 = 3.5$. The standard defines six different models with four different break points. Table 2.2 shows the path-loss-related parameters associated with these models. Model A is a flat fading model with one path between the transmitter and the receiver. The breakpoint for this model is at 5 m and the standard deviation of the shadow fading is 5 dB. Model B is recommended for a typical residential environment with LOS conditions and more than one effective path between the transmitter and the receiver. The path-loss parameters of this model are the same as Model A. Model C is recommended for a typical residential or small office environment with LOS and non-LOS (NLOS) conditions between the transmitter and the receiver. The breakpoint for this model is still at 5 m, but the standard deviation of the shadow fading is increased to 8 dB. Model D is recommended for a typical office environment with NLOS conditions with a 10 m breakpoint and 8 dB standard deviation of shadow fading. Model E is recommended for a typical large open space and office environments in areas with NLOS conditions and it has a breakpoint of 20 m and a standard deviation of 10 dB. Model F is recommended for a large open space with indoor and outdoor environment in the areas with NLOS conditions. The equations representing these models can be expressed as

$$L_{\mathrm{p}} = L_0 + \begin{cases} 10\alpha_1 \log_{10}(d) & d \le d_{\mathrm{bp}} \\ 10\alpha_1 \log_{10}(d_{\mathrm{bp}}) + 10\alpha_2 \log_{10}(d/d_{\mathrm{bp}}) & d > d_{\mathrm{bp}} \end{cases} \tag{2.14}$$

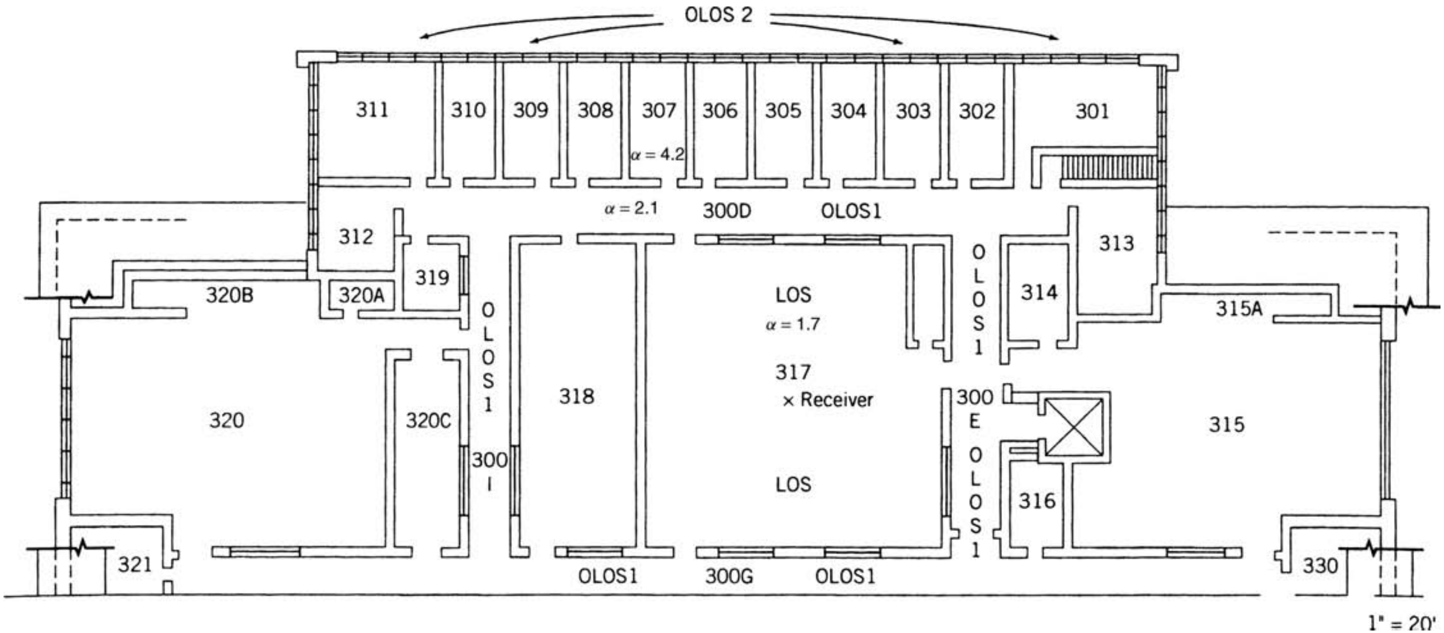


FIGURE 2.16 Layout of the third floor of the Atwater Kent Laboratories at the WPI, used for partitioned measurements.

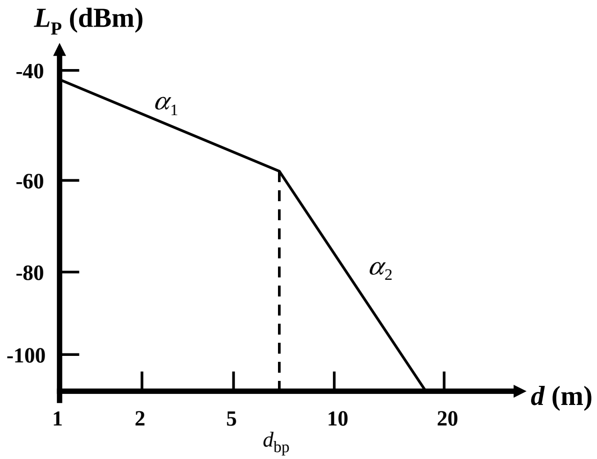


FIGURE 2.17 The IEEE 802.11 distance-partitioned path-loss model.

where using Eq. (2.8) one can calculate the path loss in the first meter L_0. For example, at 2.4 GHz and 5.2 GHz the values of the path loss at the first meter are 40.5 dB and 47 dB respectively.

Example 2.7: Coverage Using 802.11 Distance-Partitioned Model Using Eq. (2.14) and the parameters of Model D shown in Table 2.2, for calculation of the IEEE 802.11b/g with maximum path loss of 110 dB and path loss in the first meter of 40.5 dB, as we used in Example 2.6, in LOS/NLOS large office with indoor and outdoor conditions and the breakpoint of 10 m, the coverage is calculated from

$$110 = 40.05 + 20\log_{10}(10) + 35\log_{10}(d/10)$$

from which the coverage with 50% confidence is

$$d = 10 \times 10^{(110-40.5-20)/35} = 260\text{ m}$$

If we increase the confidence to 95%, with 8 dB, as shown in Example 2.5, we need an additional 13.2 dB fade margin; that reduces the coverage to

$$d = 10 \times 10^{(110-40.5-20-13.2)/35} = 109\text{ m}$$

TABLE 2.2 Parameters for Different IEEE 802.11 Recommended Path-loss Models for Six Environments

Environment	d_{bp} (m)	α_1	α_2	Shadow fading SD (dB)
A	5	2	3.5	5
B	5	2	3.5	5
C	5	2	3.5	8
D	10	2	3.5	8
E	20	2	3.5	10
F	30	2	3.5	10

The reader should note that statistical models for the coverage provide a statistical approximation for the actual coverage. So the results obtained from different models are not always the same.

2.3.4 Path-Loss Models for Outdoor Areas

Path-loss models for outdoor areas have been designed for the deployment of cellular telephone systems. In the cellular networking literature these are referred to as macrocells and microcells, while indoor models, such as the IEEE 802.11 model presented in the previous section, are called picocell models. The original frequency of operation for cellular networks was mostly around 900 MHz, and in mid 1990s it was extended to PCS bands at frequencies around 1800 and 1900 MHz. Macrocellular areas span a few kilometers to tens of kilometers, depending on the location. These are the traditional “cells” corresponding to the coverage area of a BS associated with traditional cellular telephony BSs. Microcells are cells that span hundreds of meters to a kilometer or so and are usually supported by below-rooftop-level BS antennas mounted on lampposts or utility poles. We start with introducing Okumura–Hata model for macrocells and then we introduce an extension to PCS bands.

Okumura–Hata Model. There have been extensive measurements in a number of cities and locations of the RSS in macrocellular areas and these have been reported in the literature. The most popular of these measurements corresponds to those of Okumura, who came up with a set of path-loss curves as a function of distance in 1968 [Oku68] for a range of frequencies between 100 and 1920 MHz. Okumura also identified the height of the BS antenna $h_{\rm b}$ and the height of the mobile antenna $h_{\rm m}$ as important parameters. Masaharu Hata [Hat80] came up with empirical models that provide a good fit to the measurements taken by Okumura for transmitter–receiver separations $d > 1$ km. The expressions for path loss developed by Hata are called the Okumura–Hata models, or simply the Hata models. Table 2.3 provides a summary of these models. As Example 2.8 shows, it is possible to express these equations in a form similar to Eq. (2.9) with the modification that distances between the transmitter and receiver are measured in kilometers.

Example 2.8: Path Loss Using Okumura–Hata Model Using the Okumura–Hata model, the path loss of a 900 MHz cellular system operating in a large city from a BS with a height of $h_{\rm b} = 100$ m and a mobile station installed in a vehicle with antenna height of $h_m = 2$ m is given by

$$\begin{aligned} a(h_{\rm m}) &= 3.2[\log_{10}(11.75h_{\rm m})]^2 - 4.97 = 1.05\ \text{dB} \\ L_{\rm p} &= 69.55 + 26.16\log f_{\rm c} - 13.82\log_{10}h_{\rm b} - a(h_{\rm m}) + [44.9 - 6.55\log_{10}h_{\rm b}]\log d \\ &= 118.2 + 31.8\log_{10}(d) \end{aligned}$$

Observe that this equation has the same general path-loss model format as Eq. (2.9), except that the distance d is in kilometers. The first term in the expression equaling 118.2 can be viewed as the path loss at the first kilometer, rather than the path loss in the first meter. From Eq. (2.8), the path loss in the first meter at 900 MHz is

$$L_0 = 20\log_{10}\left(\frac{4\pi}{\lambda}\right) = 20\log_{10}\left[\frac{4\pi}{(3\times10^8)/(900\times10^6)}\right] = 31.5\ \text{dB}$$

TABLE 2.3 Okumura–Hata Models for Macrocellular Path Loss

General formulation

$$L_p = 69.55 + 26.16\log f_c - 13.82\log h_b - a(h_m) + (44.9 - 6.55\log h_b)\log d$$

where f_c is in megahertz, h_b and h_m are in meters, and d is in kilometers

Suburban areas formulation

Use the general formulation above and subtract a correction factor given by

$$K_r\ (\text{dB}) = 2[\log(f_c/28)]2 + 5.4$$

where f_c is in megahertz

		Range of values
Center frequency f_c (MHz)		150–1500
h_b, h_m (m)		30–200, 1–10
$a(h_m)$ (dB) Large city	$f_c \leq 200$ MHz	$8.29[\log(1.54h_m)]^2 - 1.1$
	$f_c \geq 400$ MHz	$3.2[\log(11.75h_m)]^2 - 4.97$
Medium–small city	$150 \geq f_c \geq 1500$ MHz	$1.1(\log f_c - 0.7)h_m - (1.56\log f_c - 0.8)$

The difference of 86.7 dB is caused by the loss in the first three decades of distance. If we wanted to model this as a two-segment partitioned model, then the distance–power gradient in the first segment would be 86.7/30 = 2.89.

The above example shows that, regardless of detailed parameters obtained from extensive measurements in different urban areas resulting in a complex appearance for the Okumura–Hata model, the basic principle of the model is the same as that of the other models discussed in the previous sections.

Extension to Personal Communication Service Bands. To extend the Okumura–Hata model for PCS applications operating at 1800--2000 MHz, the European Co-operative for Scientific and Technical Research (COST) came up with the COST-231 model for urban radio propagation at 1900 MHz, which we provide in Table 2.4. In this table, $a(h_m)$ for different cities is chosen from Table 2.3. In a similar way, the Joint Technical Committee (JTC) of the Telecommunications Industry Association (TIA) has come up with the JTC models for PCS applications at 1800 MHz [Pah05].

TABLE 2.4 The COST-231 Model for PCS Applications in Urban Areas

General formulation

$$L_p = 46.3 + 33.9\log f_c - 13.82\log h_b - a(h_m) + (44.9 - 6.55\log h_b)\log d + C_M\, dx$$

where f_c is in megahertz, h_b and h_m are in meters, and d is in kilometers

		Range of values
Center frequency f_c (MHz)		1500–2000
h_b, h_m (m)		30–200, 1–10
d (km)		1–20
C_M (dB)	Large city	0
	Medium city/suburban areas	3

2.3.5 Effects of Multipath and Doppler

The received power in a multipath environment always varies with small local changes, on the order of the wavelength of the carrier frequency, in the location of the transmitter and receiver or the movement of the objects around them. However, the average received power over a small area is related to the distance from the transmitter to the center of the receiving area. Therefore, as the distance between a transmitter and a receiver increases, the received signal power will have *short-distance fluctuations* and *long-distance fluctuations* referred to as *multipath fading* and *shadow fading* respectively. Figure 2.18 illustrates variations of the received signal with respect to distance caused by multipath and shadow fading. Multipath fading is the rapid instantaneous changes in the received signal power caused by fast changes in the phase of the received signal from different paths due to small movements. Shadow fading is the long-term average changes in the RSS caused by changes in the relative position of large objects, such as buildings in urban areas, between the transmitter and the receiver.

Considering Fig. 2.18, the slope of the best-fit line to the observed data is the distance–power gradient, discussed in Section 2.2, representing the exponential rate of variation of power with the distance. The probability density function (PDF) of the variations of the average amplitude of the fade is the shadow fading characteristics of the channel. The statistics of the temporal fast multipath fading are characterized by the PDF of the sampled values of the fast variations of the channel. As we will see later in this chapter, the most popular distribution for this variation is the Rayleigh distribution, and for that reason sometimes this type of fading is referred to as Rayleigh fading. The Fourier transform of the samples of the variation of the signal is the Doppler spectrum of the channel.

So far we have considered achievable signal coverage in a variety of cell types based on the mean RSS or path loss suffered by a signal. Such a characterization of the RSS

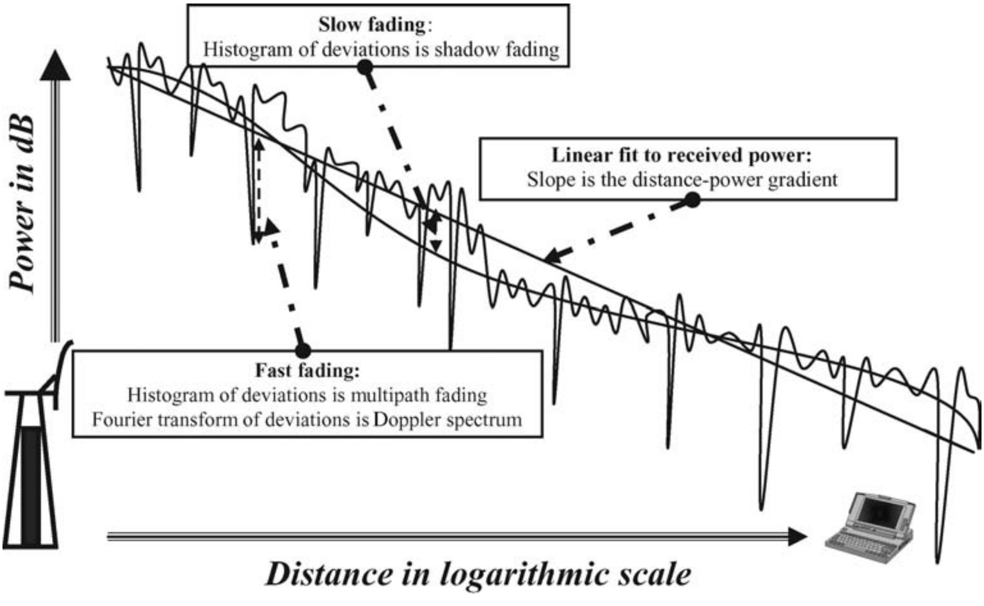

Slow fading:
Histogram of deviations is shadow fading
Linear fit to received power:
Slope is the distance-power gradient
Fast fading:
Histogram of deviations is multipath fading
Fourier transform of deviations is Doppler spectrum
Power in dB
Distance in logarithmic scale

FIGURE 2.18 Received power versus distance between a mobile terminal and a BS, linear fit with a fixed slope, multipath fading and shadow fading.

corresponds to *large-scale* average value. In reality, the received signal is rapidly fluctuating due to the mobility of the mobile terminal causing changes in multiple signal components arriving via different paths. This rapid fluctuation of the signal amplitude is referred to as *small-scale fading* and it is the result of movement of the transmitter, the receiver, or objects surrounding them. Over a small area, the *average* value of the signal is employed to compute the RSS or path loss. But the characteristics of the instantaneous signal strength are also important in order to design receivers that can mitigate these effects. We briefly describe the features of small-scale fading in this section.

Two effects contribute to the rapid fluctuations of the signal amplitude. The first, caused by the movement of the mobile terminal towards or away from the BS transmitter, is called *Doppler*. The second, caused by the addition of signals arriving via different paths, is referred to as multipath fading.

Modeling of Multipath Fading. Multipath fading results in fluctuations of the signal amplitude because of the addition of signals arriving with different *phases*. This phase difference is caused due to the fact that signals have traveled different distances by traveling along different paths. Since the phase of the arriving paths is changing rapidly, the received signal amplitude undergoes rapid fluctuation that is often modeled as a random variable with a particular distribution.

To model these fluctuations one can generate a histogram of the RSS in time. The density function formed by this histogram represents the distribution of the fluctuating values of the RSS. The most commonly used distribution for multipath fading is the Rayleigh distribution, whose PDF is given by

$$f_{\text{ray}}(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) \qquad r \geq 0$$

Here, it is assumed that all signals suffer nearly the same attenuation, but arrive with different phases. The random variable corresponding to the signal amplitude is r. Theoretical considerations indicate that the sum of such signals will result in the amplitude having the Rayleigh distribution. This is also supported by measurements at various frequencies [Pah05]. When a strong LOS signal component also exists, the distribution is found to be Ricean, and the PDF of such a distribution is given by

$$f_{\text{ric}}(r) = \frac{r}{\sigma^2}\exp\left[\frac{-(r^2+K^2)}{2\sigma^2}\right] I_0\left(\frac{Kr}{\sigma^2}\right) \qquad r \geq 0,\ K \geq 0$$

where K is a factor that determines how strong the LOS component is relative to the rest of the multipath signals.

These equations are used to determine what fraction of time a signal is received such that the information it contains can be decoded or what fraction of area receives signals with the requisite strength. The remainder of the fraction is often referred to as outage.

Small-scale fading results in very high bit-error rates (BERs). In order to overcome the effects of small-scale fading, it is not possible simply to increase the transmit power, because this will require an enormous increase in the transmit power. A variety of techniques are used to mitigate the effects of small-scale fading – in particular, error control coding with interleaving, diversity schemes, and using directional antennas. These techniques will be discussed in Chapter 3.

Doppler Spectrum. Distributions of the amplitude of a radio signal presented in the previous section demonstrate how signal is undergoing small-scale fading. In general, it is also important to know for what time a signal strength will be below a particular value (duration of fade) and how often it crosses a threshold value (frequency of transitions or fading rate). This is particularly important for designing the coding schemes and interleaving sizes for efficient performance. We see that this is a second-order statistic and it is obtained by what is known as the *Doppler spectrum* of the signal.

The Doppler spectrum is the spectrum of the fluctuations of the RSS. Figure 2.19 [How90] demonstrates the results of measurements of amplitude fluctuations in a signal and its spectrum under different conditions. In Fig. 2.19*a*, the transmitter and receiver are kept constant and nothing is moving in close to them. The received signal has a constant envelope and its spectrum is only an impulse. In Fig. 2.19*b*, the transmitter is randomly moved, resulting in fluctuation of the received signal. The spectrum of this signal is now expanded over a spectrum of around 6 Hz, reflecting the rate of variations of the RSS. This spectrum is referred to as the Doppler spectrum.

In mobile radio applications the Doppler spectrum for a Rayleigh fading channel is usually modeled by

$$D(\lambda) = \frac{1}{2\pi f_{\mathrm{m}}}\left[1-\left(\frac{\lambda}{f_{\mathrm{m}}}\right)^2\right]^{-1/2} \qquad \text{for } -f_{\mathrm{m}} \le \lambda \le f_{\mathrm{m}} \qquad (2.15)$$

Here, f_{m} is the maximum Doppler frequency possible and is related to the velocity of the mobile terminal via the expression $f_{\mathrm{m}} = v_{\mathrm{m}}/\lambda$, where v_{m} is the mobile velocity and λ is the

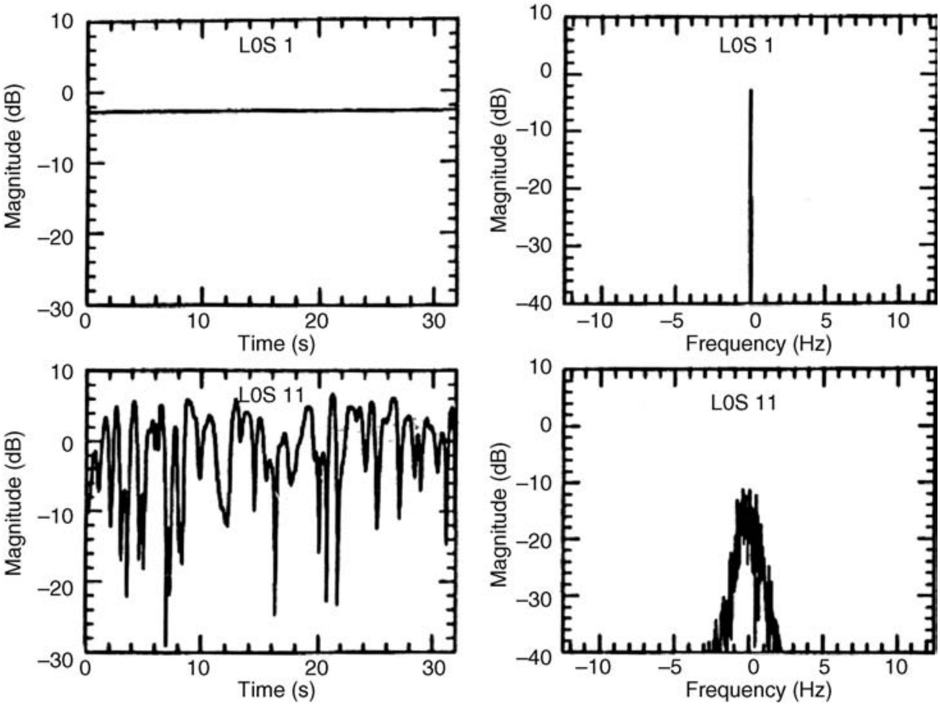


FIGURE 2.19 Measured values of the Doppler.

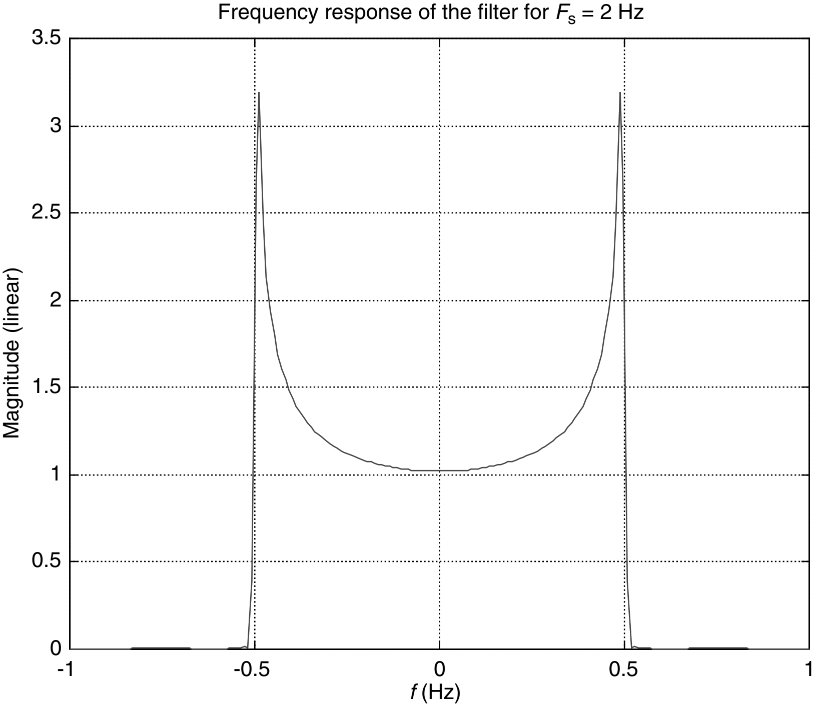
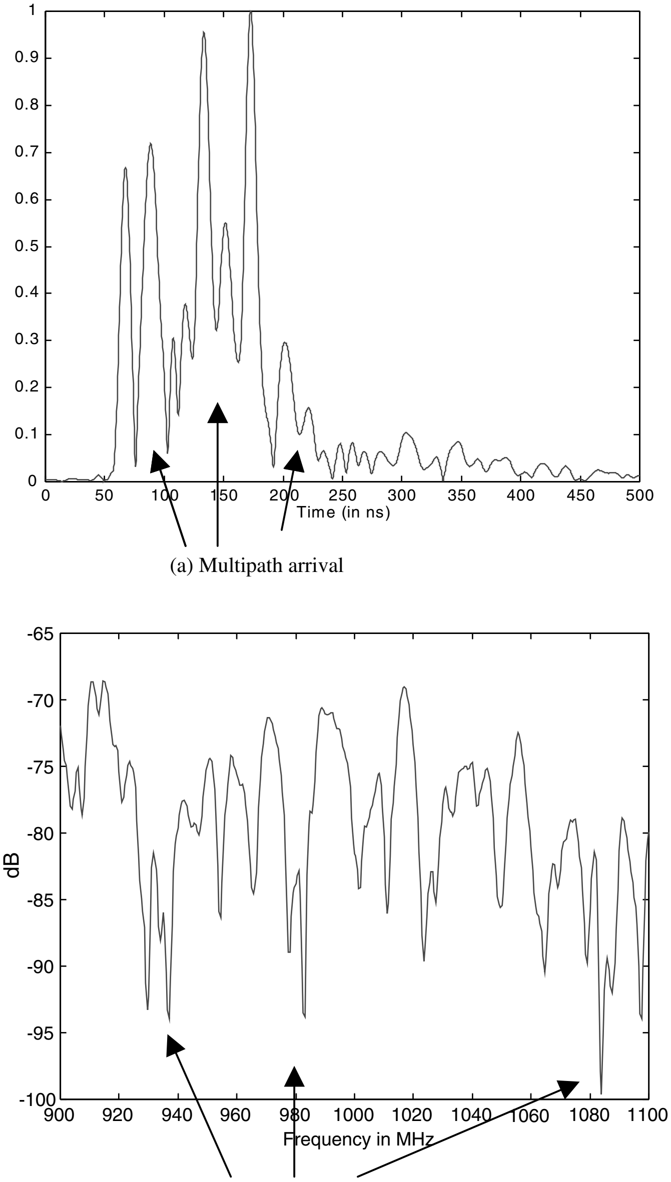


FIGURE 2.20 Classical Doppler spectrum.

wavelength of the radio signal. This spectrum, commonly used in mobile radio modeling, is also called the classical Doppler spectrum and is shown in Figure 2.20. Another popular model for the Doppler spectrum is the uniform distribution that is used for indoor applications [Pah05].

From the root-mean-square (RMS) Doppler spread, it is possible to obtain the fade rate and the fade duration for a given mobile velocity [Pah05]. These values can then be used in the design of appropriate coding and interleaving techniques for mitigating the effects of fading. Diversity techniques are useful to overcome the effects of fast fading by providing multiple copies of the signal at the receiver. Since the probability that all of these copies are in fade is small, the receiver is able to decode the received data correctly. Frequency hopping is another technique that can be used to combat fast fading. Because all frequencies are not simultaneously under fade, transmitting data by hopping to different frequencies is an approach to combat fading. This is discussed in Chapter 3.

Multipath Delay Spread. Figure 2.21 shows a sample measured time and frequency response of a typical radio channel. In the time domain, shown in Fig. 2.21*a*, a transmitted narrow pulse arrives as multiple paths with different strengths and arrival delay. In the frequency domain, shown in Fig. 2.21*b*, the response is not flat and it suffers from deep frequency-selective fades. From Fig. 2.21*a* we observe that radio signals arrive at a receiver via a multiplicity of paths. One of the significant problems caused by this phenomenon, along with fading, is intersymbol interference (ISI). If the multipath delay spread is comparable to or larger than the symbol duration, then the received waveform spreads



(a) Multipath arrival

(b) Frequency selective fading

FIGURE 2.21 Typical time and frequency response of a radio channel.

into neighboring symbols and produces ISI. The ISI results in irreducible errors that are caused in the detected signal. This effect is modeled using a *wideband* multipath channel model, sometimes called the delay power spectrum, shown in Figure 2.22, which is usually given by the impulse response:

$$h(t) = \sum_{i=1}^{L} \beta_i \delta(t-\tau_i) e^{\mathrm{j}\varphi_l} \tag{2.16}$$

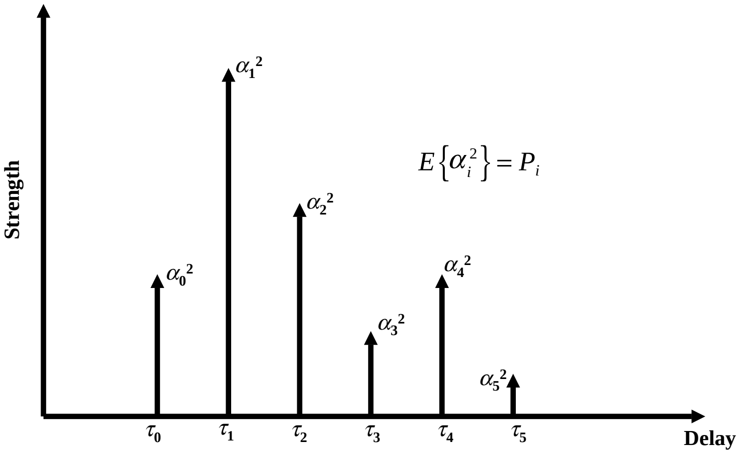


FIGURE 2.22 Wideband multipath model.

Here, most of the time the model assumes that β_i is a Rayleigh-distributed amplitude of the multipath with mean local strength $E\{\beta_i^2\} = P_i$ and the multipath arrives at a time delay τ_i with phase φ_i assumed to be uniformly distributed in (0, 2π).

The delays and multipath interarrival times have various models. In many of the models, the time delays are assumed to be fixed and only the mean-square values of the amplitudes are provided. More details on these models commonly used for the design of wireless networks are available in Pahlavan and Levesque [Pah05].

A measure of the data rate that can be supported over the channel without complex signal-processing techniques at the receiver is determined by the RMS multipath delay spread, which is the second central moment of the channel impulse response. The basic definition of the second central moment is given by

$$\tau_{\mathrm{rms}} = \sqrt{\bar{\tau^2} - (\bar{\tau})^2}$$

where moments are defined by

$$\bar{\tau^n} \equiv \frac{\sum_{k=1}^{N} \tau_k^n P_k}{\sum_{k=1}^{N} P_k} \qquad n = 1, 2$$

from which we have

$$\tau_{\mathrm{rms}} = \sqrt{\frac{\sum_{k=1}^{N} \tau_k^2 P_k}{\sum_{k=1}^{N} P_k} - \left(\frac{\sum_{k=1}^{N} \tau_k P_k}{\sum_{k=1}^{N} P_k}\right)^2} \tag{2.17}$$

Example 2.9: Calculation of the Multipath Delay Spread For a two-path channel impulse response with arrival delays $\tau_1 = 0\,\mathrm{ns}$ and $\tau_2 = 50\,\mathrm{ns}$ and path powers of $p_1 = 1$ (0 dBm) and $p_2 = 0.1$ (−10 dB m) the RMS delay spread is given by

$$\tau_{\mathrm{rms}} = \sqrt{\frac{0 \times 1 + 2500 \times 0.1}{1 + 0.1} - \left(\frac{0 \times 1 + 50 \times 0.1}{1 + 0.1}\right)^2} = 14.37\,\mathrm{ns}$$

A rule of thumb is that it is possible to support data rates that are less than the *coherence bandwidth* of the channel; that is, approximately $1/5\tau_{\text{rms}}$. The coherence bandwidth is the range of frequencies that are allowed to pass through the channel without significant distortion in the transmitted pulse shape. The RMS delay spread varies depending on the type of environment. In indoor areas, it could be as small as 30 ns in residential areas or as large as 300 ns in factories [Pah05]. In urban macrocells, the RMS delay spread is on the order of a few microseconds. This means that the maximum data rates that can be supported by a simple modem in indoor areas is around 6.7 Mb/s (at 30 ns) and 50 kb/s in outdoor areas (at 4 μs).

In order to support higher data rates, different receiver techniques are necessary. Equalization is a method that tries to cancel the effects of multipath delay spread in the receiver. Direct sequence spread spectrum enables resolving the multipath components and using them to improve performance. OFDM uses multiple carriers, spaced closely in frequency, each carrying low data rates to avoid ISI. MIMO antenna systems reduce the number of multipath components, thereby reducing the total delay spread itself. We discuss these topics in Chapter 3.

2.3.6 Emerging Channel Models

In this section we discuss some new radio channel models that are gaining importance for different applications. Position location is becoming important for emergency and location-aware applications (see Chapter 12), and models developed for communication systems are no longer sufficient to address the performance of geolocation schemes. The use of smart antennas and adaptive antenna arrays requires knowledge of the *angle of arrival* (AOA) of the multipath components in order to steer antenna beams in the right directions. We provide a brief discussion of these models below.

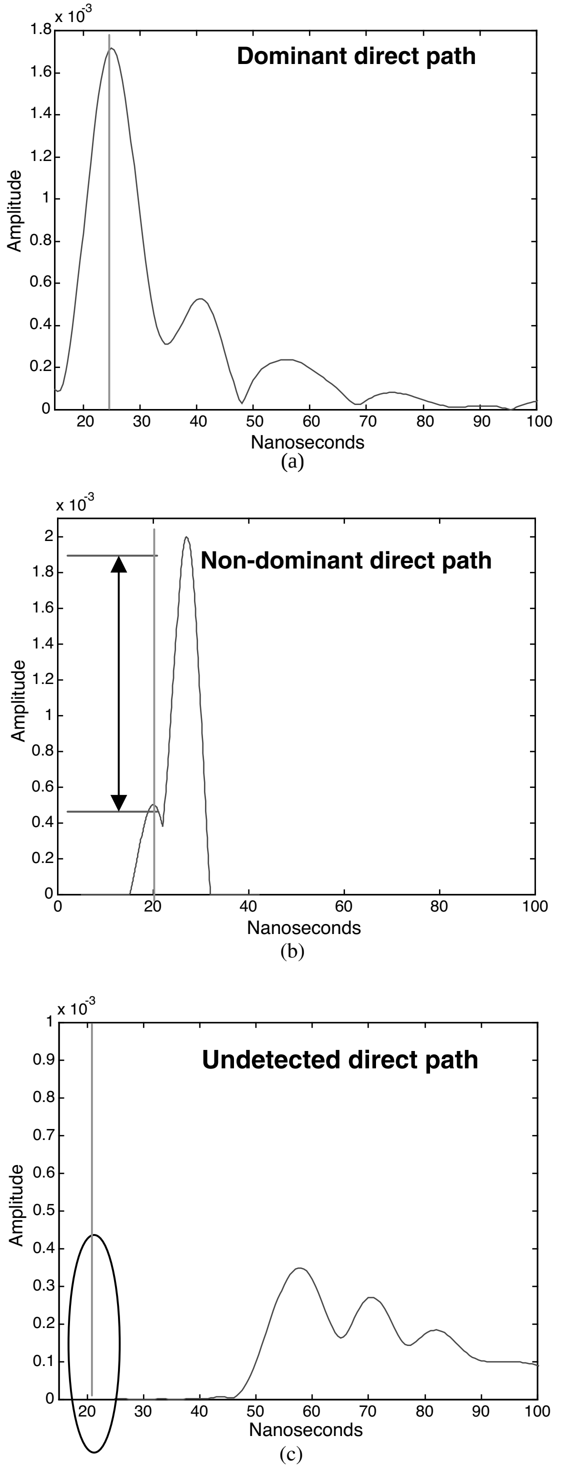
Wideband Channel Models for Geolocation. With the advent of widespread wireless communications, location of people, mobile terminals, pets, equipment and the like by employing radio signals is gaining importance as well. Several new position-location applications are rapidly emerging in the market. Civilian applications include intelligent transportation systems (ITSs), public safety (enhanced 911 or E-911 services), automated billing, fraud detection, cargo tracking, accident reporting, etc. It is possible to employ position location for additional benefits, such as cellular system design and futuristic *intelligent office* environments. Most tactical military units, on the other hand, are also heavily reliant on wireless communications. Ad hoc connectivity among individual warfighters in restrictive RF propagation environments, such as inside buildings, tunnels, and other urban structures, caves, mountainsides, and double canopy coverage in jungles and forests, requires *situation awareness* that enables the individual warfighters to determine their location and associated information. In either case, the position location service will have to operate within buildings where traditional geolocation techniques, such as the GPS, fail due to lack of sufficient signal power and the harsh multipath environment.

While RF propagation studies in the past have focused on telecommunications applications, position-location applications require a different characterization of the indoor radio channel [Pah98, Pah02]. For position-location applications, accurately detecting the *direct LOS* (DLOS) path between the transmitter and receiver is extremely important. The DLOS path corresponds to the straight line connecting the transmitter and receiver even if there are obstructions, like walls, in between. Detecting the DLOS path is important because the time of arrival (TOA) (or the AOA) of the DLOS path corresponds to the distance between the transmitter and receiver (or to the direction between them). This information is used in conjunction with multiple such measurements to locate either the transmitter or the receiver as the case may be. This is in contrast to telecommunications applications, where the emphasis is on how data bits can be sent over a link efficiently and without errors. Another issue in positioning systems is the relation between the bandwidth of the transmitted signal and the required accuracy in ranging. An error of 100 ns in estimating the delay of an arriving multipath component could result in an error of 30 m in calculation of the distance between a transmitter and a receiver. Therefore, positioning systems using TOA often require wide bandwidths to resolve multipath components and detect the arrival of the first path.

In wideband indoor radio propagation studies for telecommunication applications, often channel profiles measured in different locations of a building are divided into LOS and NLOS because the behavior of the channel in these two classes has a substantially different impact on the performance of a telecommunications system.

A logical way to classify channel profiles for geolocation applications is to divide them into three categories, as shown in Fig. 2.23. The first category is the dominant direct path (DDP) case, in which the DLOS path is detected by the measurement system and it is the strongest path in the channel profile. In this case, traditional GPS receivers designed for outdoor applications, where multipath components are significantly weaker than the DLOS path, lock on to the DLOS path and detect its TOA accurately. The second category is the non-DDP (NDDP) case, where the DLOS path is detected by the measurement system but it is not the dominant path in the channel profile. For these profiles, traditional GPS receivers may lock to the strongest path making an erroneous decision on the TOA that leads to an error in position estimation. The amount of error made by a traditional receiver is the distance associated with the difference between the TOA of the strongest path and the TOA of the DLOS path. For the second category locations with NDDP profiles, a more complex RAKE-type receiver [Pah05] can resolve the multipath and make an intelligent decision on the TOA of the DLOS path. The third category of channel profiles is the undetected direct path (UDP) profiles. In these profiles, the measurement system cannot detect the DLOS path; therefore, both traditional GPS and RAKE-type receivers cannot detect the DLOS path. If we define the ratio of the power of the strongest path to the power of the weakest detectable path of a profile as the dynamic range of a receiver, then in NDDP profiles the strength of the DLOS path is within the dynamic range of the receiver and in UDP profiles it is not. If practical considerations regarding the dynamic range are neglected, then one can argue that we have only two classes (DDP and NDDP) of profiles, because the DLOS path always exists but sometimes we cannot detect it with a practical system. Figure 2.24 shows the results of ray-tracing simulations of regions on the first floor of Atwater Kent Laboratories at WPI with different types of multipath profile for a centrally located channel sounder.

In the same way that the BER is the ultimate measure for comparing performance of different digital communication receivers, the error in the measurement of the TOA or AOA of the DLOS path is a measure of the performance of the geolocation receivers. Traditional RF studies consider the path loss and τ_{rms} as mentioned above and these are not sufficient for the geolocation problem. The relative power and delay of the signal arriving via other paths, the channel noise, the signal bandwidth, and interference all influence the detection of the DLOS path and, thus, the error in estimating the range (distance) between the transmitter and receiver. A variety of models of the indoor radio propagation for geolocation are



(a)

(b)

(c)

FIGURE 2.23 Different multipath profile conditions for indoor geolocation: (*a*) DDP; (*b*) NDDP; (*c*) UDP.

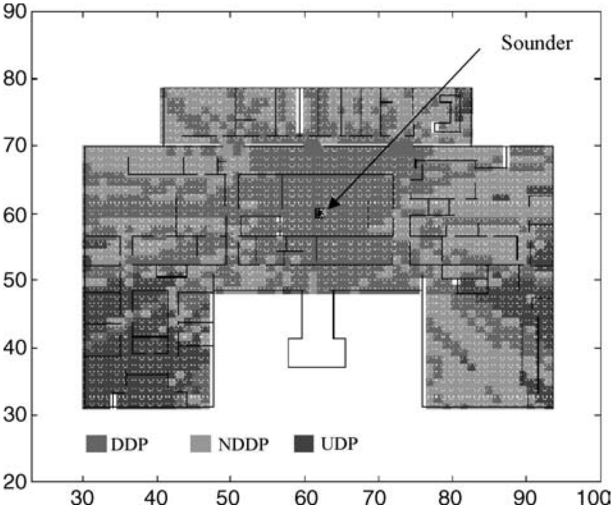


FIGURE 2.24 Simulated regions in the first floor of Atwater Kent Laboratories showing regions with different types of multipath profile for a centrally located channel sounder.

reported in the literature; in these efforts, the effects of multipath on distance measurements using TOA of the received signal are modeled and they are used for the design of indoor geolocation algorithms [Pah05].

Single-Input–Multiple-Output and Multiple-Input–Multiple-Output Channel Models. Recently, there has been a lot of attention placed on *spatial wideband channel models*, which not only provide the delay-power spectrum discussed in Eq. (2.16), but also the AOA of the multipath components. The advent of antenna array systems that are used for interference cancellation and position-location applications has made it necessary to understand the spatial properties of the wireless communications channel.

A lot of research has been carried out in the area of *single-input–multiple-output* (SIMO) radio channel models [Ert98]. In these models, a typical cellular environment is considered where it is assumed that the mobile transmitters are relatively simple and the BS can have a complex receiver with MIMO antenna systems. As shown in Fig. 2.25, the multipath

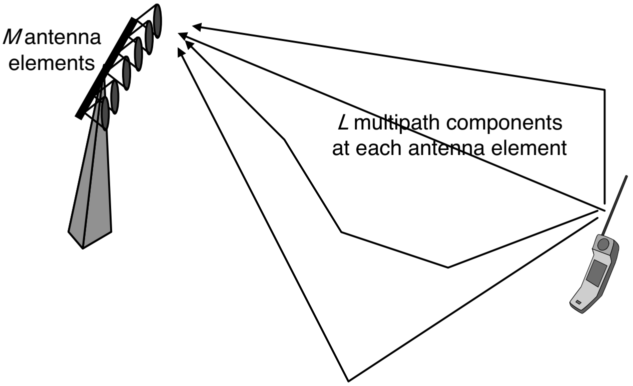


FIGURE 2.25 SIMO model.

environment is such that up to L signals arrive at the BS from different mobile terminals l with different amplitudes and phases φ at different delays τ from different directions θ. These are, in general, time invariant; as a result, the channel impulse response is usually represented by

$$\vec{h}(t) = \sum_{l=1}^{L(t)} \beta_l(t) e^{j\varphi_l(t)} \delta(t-\tau_l(t)) \vec{a}(\theta_l(t))$$

Note that the channel impulse response is now a *vector* rather than a scalar function of time. The quantity $\vec{a}(\theta(t))$ is called the array response vector and will have M components if there are M antenna array elements. Thus, there are M channel impulse responses each with L multipath components. A variety of models are available in Ertel *et al.* [Ert98]. The amplitudes are usually assumed to be Rayleigh distributed, although they are now dependent on the array response vector $\vec{a}(\theta(t))$ as well.

An extension of this model to the scenario where there are N mobile antenna elements and M BS antenna elements for MIMO systems is available [Ped00]. In this case, the channel impulse response is an $M \times N$ matrix that associates a *transmission coefficient* between each pair of antennas for each multipath component. Experimental results and models are considered by Pedersen *et al.* [Ped00] and Kermoal *et al.* [Ker00]. The IEEE 802.11 standardization committee recommends models for the MIMO systems for performance evaluation of IEEE 802.11n WLAN devices [Erc03].

QUESTIONS

1. What are the typical bandwidths of Cat-3, Cat-5, and Cat-7 TP wirings? Which one is used for telephony and which one for Ethernet?
2. What are NEXT and FEXT and how do they affect the bandwidth of TP wiring?
3. Compare cable, twisted pair, and fiber optic lines in terms of length of coverage, availability in existing buildings, cost of installation, and radiated interference.
4. What are the different grades of optical fiber lines and how do they differ in\bandwidth and path-loss?
5. How do the coverage in guided and wireless media differ?
6. What are the three important radio propagation phenomena at high frequencies? Which of them is predominant in indoor environments?
7. Explain what path-loss gradient means. Give some typical values of the path-loss gradient in different environments.
8. Explain the meaning of the expression "a loss of 37 dB per decade of distance" in terms of the path loss gradient α.
9. Why does multipath in wireless channels limit the maximum symbol transmission rate? How can we overcome this limitation?
10. What is the Doppler spectrum and how can one measure it?
11. Differentiate between shadow fading and fast fading.
12. What distributions are used to model fast fading in LOS situation? In OLOS situations?
13. What are the differences between multipath, shadow and frequency-selective fading?
14. For position location applications, how are wideband radio channels classified? How is this classification useful?
15. What is the difference between a SIMO and a MIMO radio channel?

PROBLEMS

Problem 1:

What is the minimum SNR required to carry a Gigabit Ethernet signal over one pair of 100 meters cat-3 TP wiring with a bandwidth of 25 MHz?

Problem 2:

Determine the maximum length of a Gigabit Ethernet Fiber Optic cable using multimode 62.5/125 fiber at 1300 nm wavelength. Assume the link budget of the LED and the photo sensitive diodes used for transmission is 4 dB and the loss per Km of the fiber optics cable is 1.5 dB. Leave a 3 dB safety margin in your calculations with allowance of 2 splices each with 0.1 dB loss.

Problem 3:

Assuming wireless devices use an antenna length of one fourth of the wavelength for the transmitted frequency, what are the typical antenna lengths for the cellular phones operating at 900 MHz and 1800 MHz PCS bands and WLANs operating at 2.4 GHz and 5.2 GHz.

Problem 4:

Figure P2.1 shows the measured relation between data rate and distance of a DSL modem over Cat-3 TP wiring.

(a) Find a best fit exponential function to relate the data rate to the distance.

(b) Use equations for path-loss of the Cat-3 TP wiring at the frequency of 1.1 MHz to plot the signal loss in dB versus distances identified in Fig. P2.1.

(c) Use results of (a) and (b) to plot the relation between data rate and path-loss requirement.

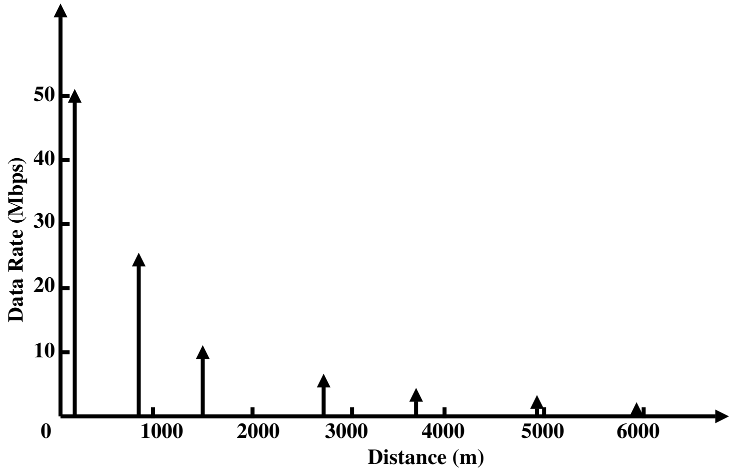


FIGURE P2.1 Data rate versus distance over a Cat-3 TP used for DSL modems.

Problem 5:

(a) Using MATLAB or any other computational software tool plot the path gain of Cat-3 and Cat-5 cables as a function of distance for a frequency of 10 MHz.

(b) Use the plot to find the path-loss for a 100 meter Cat-3 cable.

(c) What would be the length of a Cat-5 for the same of path-loss?

(d) What would be the maximum length of a Cat-5 cable if it carries legacy Ethernet 10 Mbps signal? Explain.

Problem 6:

What is the received power (in dBm) in free space of a signal whose transmit power is 1 W and carrier frequency is 2.4 GHz if the receiver is at a distance of 1 mile (1.6 km) from the transmitter? What is the path loss in dB?

Problem 7:

Use the Okumura-Hata and COST-231 models to determine the maximum radii of cells at 900 MHz and 1900 MHz respectively having a maximum acceptable path loss of 130 dB. Use $a(h_m) = 3.2\ [\log\ (11.75\ h_m)]^2 - 4.97$ for both cases.

Problem 8:

The path loss in a building was discovered to have two factors adding to the free space loss: a factor directly proportional to the distance and a floor-attenuation-factor (FAF). In other words, the path loss = free space loss + βd + FAF. If the FAF is 20 dB, and the distance between transmitter and receiver is 40 m, determine what the value of β should be so that the path loss suffered is less than 115 dB. Use $f_c = 900$ MHz.

Problem 9:

Table P2.1 provides the minimum required RSS for an IEEE 802.11b device to operate at different rates.

(a) Calculate the coverage associated with each data rate in the table.

(b) Plot the staircase function of the Data Rate vs RSS

(c) Plot the staircase function of the Data Rate vs Distance.

(d) If a mobile terminal moves away from an 802.11 AP and it goes out of the coverage area, calculate the average data rate that terminal observes during the movement in the coverage area of the AP?

TABLE P2.1 Data rate and minimum power requirement for IEEE802.11b

Data Rate (Mbps)	RSS (dBm)	Coverage (m) Using IEEE 802.11 Path-loss Model D
11	−82	
5.5	−87	
2	−91	
1	−94	

Problem 10:

The transmitted power in IEEE 802.11g is 100 mW. When the terminal is close to the access point (AP) the maximum data rate is 54 Mbps that requires a −72 dBm received signal strength (RSS). The minimum supported data rate is 6 Mbps that requires a minimum of −90 dBm of RSS.

(a) Determine the coverage of the AP for 54 Mbps and 6 Mbps in a small office using the IEEE 802.11 channel model.

(b) Repeat (a) assuming a single distance-power gradient of $\alpha = 2.5$ and compare your results with the results of part (a).

Problem 11:

In a mobile communications network, the minimum required signal to noise ratio is 12 dB. The background noise at the frequency of operation is -115 dBm. If the transmit power is 10 W, transmitter antenna gain is 3 dBi, the receiver antenna gain is 2 dBi, the frequency of operation is 800 MHz and the base station and mobile antenna heights are 100 m and 1.4 m respectively, determine the maximum in building penetration loss that is acceptable for a base station with a coverage of 5 km if the following path loss models are used.

(a) Free space path loss model with $\alpha = 2$.

(b) Two-ray path loss model with $\alpha = 4$.

(c) Okumura-Hata model for a small city

Problem 12:

Signal strength measurements for urban microcells in the San Francisco Bay area in a mixture of low-rise and high-rise buildings indicate that the path loss *Lp* in dB as a function of distance *d* is given by the following linear fits:

$$L_P = \begin{cases} 81.14 + 39.40 \log f_c - 0.09 \log h_b + [15.80 - 5.73 \log h_b] \log d, for\ d < d_{bk} \\ [48.38 - 32.1 \log d_{bk}] + 45.7 \log f_c + (25.34 - 13.9 \log d_{bk}) \\ \quad \log h_b + [32.10 + 13.90 \log h_b] \log d \\ + 20 \log(1.6/h_m), for\ d >_{bk} \end{cases}$$

Here, d is in kilometers, the carrier frequency f_c is in GHz (that can range between 0.9 and 2 GHz), h_b is the height of the base station antenna in meters, and h_m is the height of the mobile terminal antenna from the ground in meters. The *breakpoint* distance d_{bk} is the distance at which two piecewise linear fits to the path loss model have been developed and it is given by $d_{bk} = 4h_b h_m/(1000\lambda)$, where λ is the wavelength in meters (when you calculate d_{bk} using units of meters for h_b, h_m, and λ, the scaling factor of 1000 makes the units of d_{bk} as km). What would be the radius of a cell covered by a base station (height 15 m) operating at 1.9 GHz and transmitting a power of 10 mW that employs a directional antenna with gain 5 dBi? The sensitivity of the mobile receiver is −110 dBm. Assume that $h_m = 1.2$ m. How would you increase the size of the cell?

Problem 13:

A mobile system is to provide 95% successful communication at the fringe of coverage with a location variability having a zero mean Gaussian distribution with standard deviation of 8 dB. What fade margin is required?

Problem 14:

Sketch the power-delay profile of the following wideband channel. Calculate the excess delay spread, the mean delay, and the RMS delay spread of the multipath channel described in Table P2.2. A channel is considered "wideband" if the inverse of the rms multipath spread is smaller than the data rate of the system. Would the channel be considered a wideband channel for a binary data system at 25 kbps? Why?

TABLE P2.2 Delay-power profile for Problem 14

Relative delay in microseconds	Average relative power in dB
0.0	−1.0
0.5	0.0
0.7	−3.0
1.5	−6.0
2.1	−7.0
4.7	−11.0

Problem 15:

The modulation technique used in the existing Advanced Mobile Phone System (AMPS) is analog FM. The transmission bandwidth is 30 kHz per channel and the maximum transmitted power from a mobile user is 3 W. The acceptable quality of the received SNR is 18 dB and the power of the background noise in the system is −120 dBm. Assuming that the height of the base and mobile station antennas are $h_b = 100$ m and $h_m = 3$ m respectively and the frequency operation of is $f = 900$ MHz, what is the maximum distance between the mobile station and the base station for an acceptable quality of communication?

(a) Assume free space propagation with transmitter and receiver antenna gains of 2.

(b) Use Hata's equations for Okumura's model in a large city.

Problem 16:

The IEEE 802.11 WLANs operate at a maximum transmission power of 100 mW (20 dBm) using multiple channels with different carrier frequencies. The IEEE 802.11g use 2.402–2.480 GHz bands and the IEEE 802.11a uses 5.150–5.825 GHz bands. Both standards use OFDM modulation with a bandwidth of 20 MHz.

(a) Calculate received signal strength in dBm at one meter distance of an IEEE 802.11g access point for the smallest and the largest possible carrier frequencies in the band.

Assume that transmitter and receiver antenna gains are one and in one meter distance signal propagation follows the free-space propagation rules.

(b) Repeat (a) for the IEEE 802.11a WLANs.

(c) Compare the received signal strengths at one meter distance of the IEEE 802.11g and IEEE 802.11a devices. Use the middle of the allocated band for each standard as the carrier frequency in your calculations.

(d) Compare the rate of the received signal fluctuations (Doppler shift), due to the change in frequency of operation, for the IEEE 802.11g and the IEEE 802.11a. Use the middle of the allocated for each standard as the carrier frequency in your calculations.

Problem 17:

A multipath channel has three paths at 0, 50, and 100 nsec with the relative strengths of 0, -10, and -15 dBm, respectively.

(a) What is the multipath spread of the channel?

(b) Calculate the rms multipath spread of the channel.

(c) What would be the difference between multipath spreads and rms multipath spreads of this three-path channel and a two-path channel formed by the first and the third path of this profile?

Problem 18:

In the 1900 MHz bands, measurements [Bla92] show that the RMS delay spread increases with distance. An upper bound on the RMS delay spread is given by the equation $\tau = e^{0.065L(d)}$ in ns where $L(d)$ is the mean path loss in dB as a function of distance d between the transmitter and receiver. The path loss itself is given by the equation $L(d) = L_0 + 10\alpha \log_{10}(d/d_0)$ where $L_0 = 38$ dB and $\alpha = 2.2$ for $d < 884$ m and $\alpha = 9.36$ for $d > 884$ m. The standard deviation of shadow fading is 8.6 dB. Assume that you are using a transmission scheme that has a symbol rate of 135 ksps without equalization. If the maximum allowable path-loss is 135 dB, what limits the size of the cell – the RMS delay spread, or the outage at 90% coverage at the cell-edge? Explain clearly all of your steps.

PROJECT 1: SIMULATION OF MULTIPATH FADING

Figure P2.2 shows two mobile Automatic Guided Vehicles (AGVs) communicating in a large open indoor area with a ceiling with a height of 5 m and two antennas that are 1.5 meter above the ground. Communication between the terminals is taking place through three paths: the direct path, the path reflected from the ground and the path reflected from the ceiling. The reflection coefficients from the ground and the ceiling are 0.7 and each reflection causes an additional 180 degrees phase shift.

(a) If the transmitted power is *1 mW*, derive an expression for calculation of P_o, the free space received signal strength in *1 m* distance from the transmitter, as a function of frequency of operation, *f*.

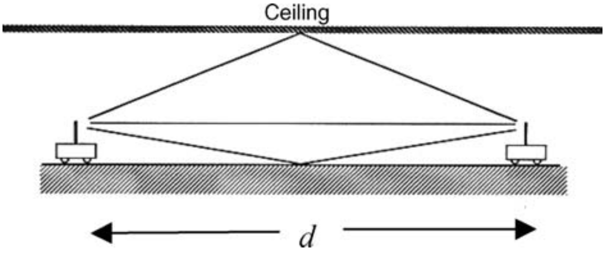


FIGURE P2.2 An indoor scenario for path-loss modeling using Ray tracing.

(b) Derive an expression for calculation of the amplitude, delay, and phase of each of the arriving paths as a function of distance, *d*, and the frequency of operation, *f*.

(c) Derive an expression for calculation of the received signal strength (RSS) as a function of *d* and *f*.

(d) Use MATLAB to plot the RSS versus distance for $1\,m < d < 100$ m for center frequencies of 900 MHz, 2.4 GHz, and 5.2 GHz (similar to the plot provided in Fig. P2.3).

(e) Discuss the relation between the received signal strength, rate of fluctuations, and frequency of operation.

(f) Use the results for 2.4 GHz to design a two-piece path-loss model for the RSS by determining a suitable break-point and two distance power gradients in the two regions. Compare your model with the IEEE 802.11 models and discuss your observations.

PROJECT 2: THE RSS IN IEEE 802.11

There are a number of software tools (e.g. WirelessMon by PassMark) which can be used to gather information about access points in close proximity. These tools provides multiple

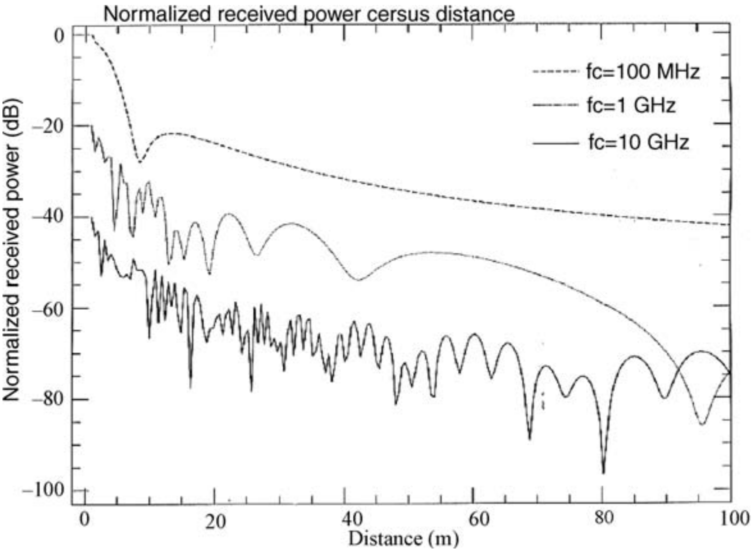


FIGURE P2.3 Results of simulation in different frequencies.

features but we are going to use them to log the received signal strength (RSS) from chosen access points (APs) at different locations to check with 802.11 models. The following steps can be used to make an RSS measurement using these tools.

- Install a software tool for measurement of RSS (e.g. you can download wirelessmon.exe from http://www.passmark.com/products/wirelessmonitor.htm) in your laptop.
- Set the software to monitor the access point of your choice, the access points can be distinguished from each other by their MAC addresses or SSIDs.
- Modify the logging options of the software for recording the characteristics of an AP.
- Record the RSS readings from a specific AP.
  (a) Do war driving in a specific floor of your building, for which you have an schematic available, to find the exact location of the AP in that floor. Show the locations in the schematics of the building.
  (b) Select five different locations on the floor of your choice which are approximately 1, 5, 10, 20, and 30 meters away from your AP of choice. Spread the points over the entire floor and mark them on your schematic floor plan. Determine the distance from the selected points to each of the AP locations in that floor.
  (c) Measure the RSS at each location for at least 1 minute. Calculate the average power received from each AP in each location and record them in a table which relates the distance to RSS from your target AP.
  (d) Use the table to generate a scatter plot of the average RSS (in dBm) vs distance in logarithmic form for all APs in your target floor.
  (e) Find the best fit 802.11 model for your data in the scattered plot.
  (f) Use www.speakeasy.net/speedtest to record the measured data rate in each of the five locations.
  (g) Explain the correlation among the throughput from speakeasy and the power and distance in each location.

PROJECT 3: COVERAGE AND DATA RATE PERFORMANCE OF THE IEEE 802.11B/G WLANS

I. Modeling of the RSS

To develop a model for the coverage of the IEEE 802.11b/g WLANs, a group of undergraduate students at WPI measured the received signal strength (RSS) in six locations in the third floor of the Atwater Kent Laboratory (AKL) at WPI, shown in Figure P2.4. After subtracting the RSS from the transmitted power recommended by the manufactures they calculated the path-loss for all the points that are shown in Table P2.3.

To develop a model for the coverage of the WLANs they used the simple distance-power gradient model:

$$L_p = L_0 + 10\alpha \log_{10}(d)$$

Where d is the distance between the transmitter and the receiver, L_p is the path-loss between the transmitter and the receiver, L_o is the path-loss at one meter distance from the transmitter, and α is the distance-power gradient. One way to determine L_o and a from the results of measurements is plot the measured L_p and $\log_{10}$ (d) and find the best fit line to the results of measurements.

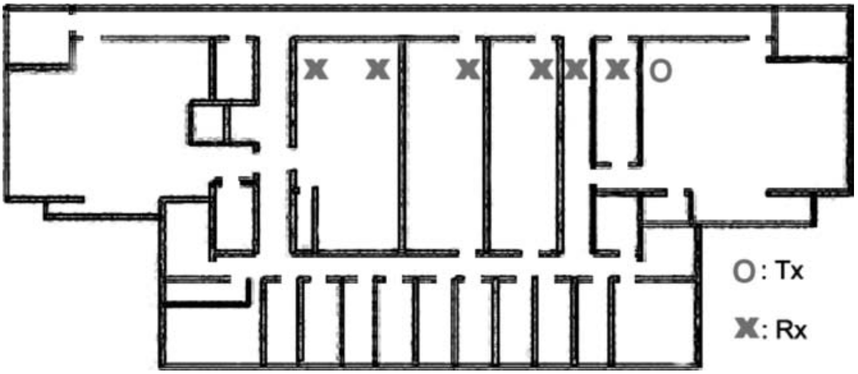


FIGURE P2.4 Location of the transmitter and first five locations of the receiver used for calculation of the RSS and path-loss.

TABLE P2.3 The distance between the transmitter and the receiver and the associated path-loss for the experiment

Distance (m)	Number of Walls	Lp (dB)
3	1	62.7
6.6	2	70
9.5	3	72.75
15	4	82.75
22.5	5	90
28.8	6	93

(a) Use the results of measurements by students to determine the distance-power gradient, α, and path loss in the one meter distance from the transmitter, L_o. In your report provide the Matlab code and the plot of the results and the best fit curve.

(b) Manufacturers often provide similar measurement tables for typical indoor environments. Table P2.4 shows the RSS at different distances for open areas (an area without wall), semi-open areas (typical office areas), and closed areas (harsher indoor environments) provided by PROXIM, one of the manufacturers of WLAN products. Use the results of measurements from the manufacturer and repeat part (a) for the three areas used by the manufacturer. Which of the measurements areas used by the manufacturer resembles the 3rd floor of the Atwater Kent laboratory used by the MQP students? Assume that the transmitted power used for these measurements was 20 dBm. In your report give the curves used for calculations of the distance-power gradient in different locations.

II. Coverage Study

IEEE 802.11b/g WLANs support multiple data rates. As the distance between the transmitter and the receiver increases the WLAN reduces its data rate to expand its coverage. The IEEE 802.11b/g standards recommend a set of data rates for the WLAN. The first column of the Table P2.4 shows the four data rates supported by the IEEE

TABLE P2.4 Data rate, distance in different areas, and the RSS for IEEE 802.11b (Source Proxim)

Data Rate (Mbps)	Closed area (m)	Semi-Open area (m)	Open area (m)	Signal Level (dBm)
11	25	50	160	−82
5.5	35	70	270	−87
2	40	90	400	−91
1	50	115	550	−94

TABLE P2.5 Data rates and the RSS for the IEEE 802.11g (Source Cisco)

Data Rate (Mbps)	RSS (dBm)
54	−72
48	−72
36	−73
24	−77
18	−80
12	−82
9	−84
6	−90

802.11b standard and the last column represents the require RSS to support these data rates. Table P2.5 shows the data rates and the RSS for the IEEE 802.11g provided by Cisco.

(a) Plot the data rate versus coverage (staircase functions) for the IEEE 802.11b WLANs for closed, open, and semi-open areas using Table P2.4 provided by Proxim. Discuss the coverage vs data rate performance in different areas and relate them to the value of α of different areas, calculated in part I of the project.

(b) Using α and L_0 found for the third floor of the AKL, plot the data rate versus coverage (staircase functions) for IEEE 802.11b and g WLANs operating in that area. Discuss the differences in data rate vs coverage performance of the 802.11b and g in the third floor of AKL.

3

FUNDAMENTALS OF PHYSICAL LAYER TRANSMISSION

3.1 INFORMATION TRANSMISSION

In Chapter 2 we described the behavior of different wired and wireless media. To transfer the information over a medium we need information transmission techniques. A number of analog and digital information transmission techniques have evolved in the past century. Fundamentally, these techniques are either used for transmitting a waveform over the transmission medium or to process the information so that it uses the transmission facilities efficiently. In this chapter we provide an overview of transmission techniques which gradually moved into implementations of popular physical layers in modern information networks. In Chapter 4 we will present a summary of the coding techniques and protocols used for reliable packet transmission. The material presented in these chapters prepares the reader for understanding the reasons behind the development of WAN, LAN, and PAN information networking alternatives in modern times. The material presented is carefully

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

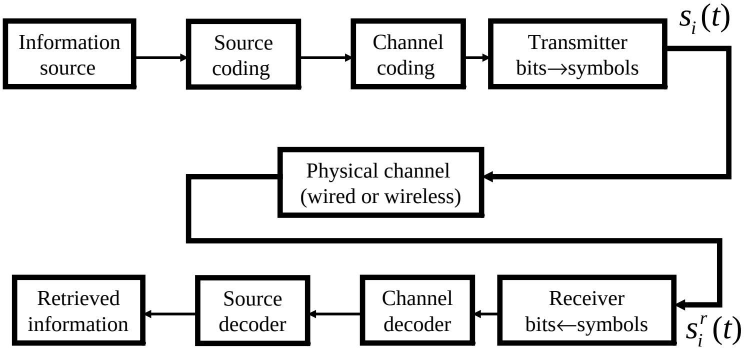


FIGURE 3.1 Overview of transmission and reception of digital data in the PHY layer.

selected to avoid complicated signal processing details that would need training in electrical engineering. The audience of the book is assumed to be computer engineering and science students. However, electrical engineering students may have an easier time grasping this material because they may have studied similar material in other courses.

Figure 3.1 shows a general block diagram for different components involved in information transmission. The information source is first encoded into digital information or a bit stream which then passes through a channel coding process that increases its integrity and protects the data against disturbances caused by the transmission channel. The encoded data stream is then mapped into a set of analog waveforms s_i(t) each representing a set of transmitted information bits. Symbols are transmitted every T_s seconds, referred to as the symbol duration or symbol interval. The waveform representing the symbols passes through the channel, which may make changes to the shape of the signal and disturb it with additive noise. The received symbol is then processed at the receiver, which extracts the best processed estimate of the transmitted bits associated with the received symbol which is then decoded to retrieve the transmitted information. We have structured this chapter to address bit transmission techniques. In networking we use packets or frames for transmission. Source coding, channel coding, and techniques used for reliable packet transmission are presented in Chapter 4.

In the rest of this section we discuss issues related to wired and wireless transmission followed by introducing the simplest implementation of the physical layer using baseband transmissions. Section 3.2 describes multisymbol transmission and signal constellations used to express the details of more complex transmission techniques. Section 3.3 is devoted to performance analysis of the PHY layer. In this section we describe the effects of noise and fading on the performance of modems and we introduce the concept of diversity techniques and its impact on the performance of wireless networks. Section 3.4 provides an overview of the modern techniques used for wideband or high data rate wireless networks: spread spectrum, OFDM, space–time coding (STC), and MIMO transmission techniques are described in this section.

3.1.1 Wired and Wireless Transmission

In principle, transmission techniques used in information networks are applicable to all wired and wireless modems because the basic design issues are common to both systems. In general, we would like to transmit data with the highest achievable data rate with a minimum expenditure of signal power, channel bandwidth, and transmitter and receiver complexity. In other words, we usually want to maximize both bandwidth efficiency and power efficiency and minimize the transmission system complexity. However, the emphasis on these three objectives varies according to the application requirement and medium for transmission, and there are certain details that are specific to particular applications and transmission media. Also, these design objectives are often conflicting and the tradeoffs decide what factors are considered more important than others.

In most legacy wired LANs or optical wireless channels, transmission schemes over twisted pair, coaxial cable, or wireless/fiber optical media were very simple. The received data from the higher layers were coded to facilitate clock synchronization at the receiver and the coded signals were applied directly to the medium. These transmission techniques are often referred to as *baseband transmission schemes*. In voice-band modems, DSL, and coaxial cable modem applications, the transmitted signal is modulated over a *carrier*. In voice-band modems this carrier is around 1800 Hz; that is, the center of the *passband* of 300–3300 Hz of telephone channels. The purpose of modulation here is to eliminate the DC component from the transmission spectrum and allow the usage of more bandwidth-efficient modulation to support higher data rates. For DSL services, the spectrum in which DSL is utilized is shifted away from the lower frequencies used for voice applications. Discrete multitone transmission, a form of OFDM, is employed there. In cable modems, modulation is employed to shift the spectrum of the signal to a particular frequency channel and improve the bandwidth efficiency of the channel to support higher data rates. In industry, cable modems are referred to as broadband modems because they provide a much higher data rate (broader band) than the voice-band modems. High bandwidth efficiency in the voice band has a direct economic advantage to the user, as it can reduce connect time and avoids the necessity for leasing additional circuits to support the application at hand. The typical telephone channel is less hostile than a typical radio channel, providing a fertile environment for examining and employing complex transmission and coding techniques such as QAM and TCM and complex signal-processing algorithms, such as equalizers and echo cancellers. Specific impairments seen on telephone channels are amplitude and delay distortion, phase jitter, frequency offset, and effects of nonlinearity. Many of the practical design techniques of wired modems have been developed to deal efficiently with these categories of impairments.

Wireless channels are characterized by multipath fading and Doppler spread, and a key impediment in the radio environment is the relatively high levels of average signal power needed to overcome fading. However, there are other considerations that impact the selection of a modem technique for a wireless application. For example, in radio systems, bandwidth efficiency is an important consideration, since the radio spectrum is limited and many operational bands are becoming increasingly crowded. These requirements can vary somewhat from one system to another, depending upon the type of system, the requirements for delivered service, and the users' equipment constraints.

Most wireless networks that support mobile users have a need for bandwidth-efficient transmission. One of the major incentives of the cellular telephone industry for moving from analog to digital and then from TDMA to CDMA was to increase the bandwidth efficiency and, consequently, the number of users. A cellular carrier company is assigned a specified amount of licensed bandwidth in which to operate their system; therefore, an increase in system capacity leads directly to increased revenues. This defines another clear need for transmission techniques that provide efficient utilization of available bandwidth.

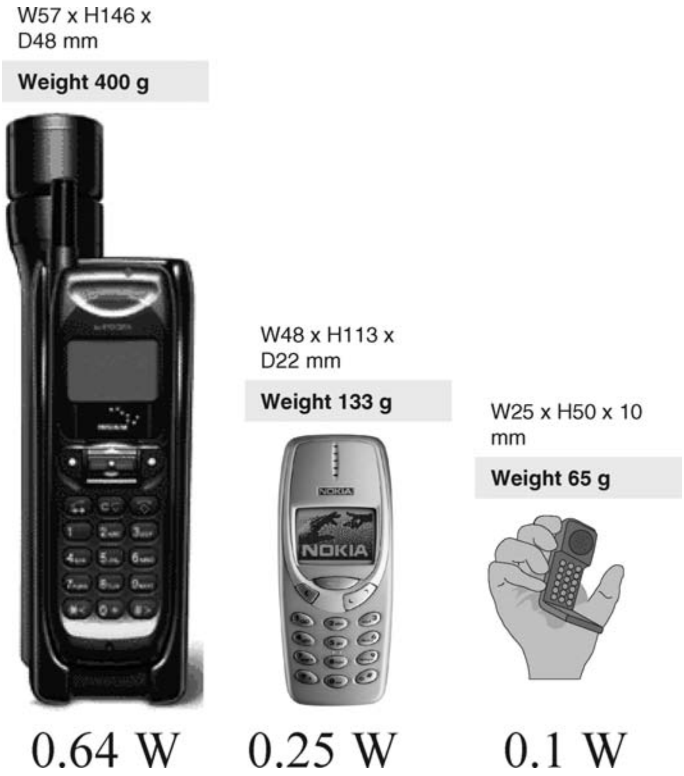


FIGURE 3.2 Power consumption and the size/weight of a mobile terminal.

Power efficiency is another parameter which is not of major importance for wireless equipment using AC power sources, such as WLAN APs or cellular BSs, but is of crucial importance in battery-oriented applications such as handheld cellular or WLAN cards used in laptops. In these applications, power consumption translates into battery size and recharging intervals and, even more important to the mobile user, into the size and weight of portable terminals. Figure 3.2 demonstrates how the power consumption is directly related to device size and weight. Power efficiency will become increasingly important as consumers become accustomed to the convenience of small hand-held communication devices.

There are two facets of the power requirement: one is the power needed to operate the electronics in the terminal; the other is the amount of power needed at the input to the power amplifier in order to radiate a given amount of signal power from the antenna. The radiated signal power, of course, translates directly into signal coverage and it is a function of the data rate and the complexity of the receiver. Higher data rates require higher operating levels of transmitted power. More complex systems using computationally demanding coding techniques or employing adaptive signal processing algorithms require less transmission power. However, a more complex receiver design increases the power consumed by the electronics and, consequently, reduces battery life. In some applications a compromise has to be made between the complexity of the receiver and the electronic power consumption. For example, in the case of handheld cordless telephone devices, some manufacturers avoid the use of complex speech coding techniques in order to reduce battery consumption.

Also, in the design of high-speed data communication networks, for laptop or pen-pad computers, some designers find it difficult to justify the additional electronic power consumption necessitated by the inclusion of adaptive algorithms.

In spread-spectrum CDMA systems, power efficiency and overall system bandwidth efficiency are closely related. The use of a more power-efficient modulation method allows a system to operate with a lower transmission power. The performance of a CDMA system is limited by the interference from other users on the system, and an improvement in power efficiency in turn increases the bandwidth efficiency of the system. To support higher data rates, OFDM and MIMO technologies provide a more attractive solution for transmission. Most of the modern WLAN and WPAN techniques employ CDMA, OFDM, and MIMO transmission techniques to support a better quality streaming for voice and video and higher data rates for information transmission.

3.1.2 Baseband Transmission

We start our technical discussions on transmission techniques with *baseband transmission* using line coding techniques. In baseband transmission, the digital signal is transmitted without modulating it over a carrier at a higher frequency. In line-coded transmission, the digital data stream is line coded to facilitate synchronization at the receiver and avoid the DC offset during transmission. Baseband line-coded signaling is commonly used in short-distance wired and wireless applications. In wired applications, baseband signaling using differential Manchester line-coding is used in the IEEE 802.3 Ethernet, the dominant standard for LANs, as well as IEEE 802.5 token ring, the competitor of Ethernet in the early days of the LAN industry. In wireless applications, baseband transmission with line coding is popular in high-speed diffuse and directed beam IR wireless LANs. We discuss this topic first to provide a clear understanding of the issues through simple examples.

If the data stream produced by a computer is applied directly to the wires, then the receivers will have difficulty in synchronizing with the transmitted symbols. To provide better synchronization at the receiver, the format of the incoming data stream is modified before transmission. This modification process is often referred to as line coding. We now provide more details of several popular line-coding techniques with a discussion on why they have evolved and examples of how they work.

Figure 3.3 shows examples of popular line-coding techniques. The non-return to zero (NRZ) line coding, shown in Fig. 3.3*a*, just uses two different amplitudes for the two different binary digits. On wired links, these two amplitudes are selected at the same level with opposite polarities. With this setup, the average transmitted amplitude of the waveform is zero, providing a zero DC component for the transmission. It can be shown that this form of transmitted symbols provides the best performance with a fixed transmission power. In optical communications, however, we have LEDs, which cannot implement polarity. Usually, the light is either on or off. Hence, the transmission does not have a zero average or DC value in this case. In the case of NRZ-I, shown in Fig. 3.3*b*, in the computer communication community the letter “I” stands for inverted. Here, the term “inverted” does not mean that the amplitude is inverted; rather, it means that the information bit is encoded at the edge of the bit. In Fig 3.3*b*, whenever we have a “0” we have no transition at the edge and when we have a “1” we have a transition at the edge. This arrangement in the telecommunication literature is referred to as differential coding. The advantage of this type of coding is that we can afford 180° phase ambiguity, which means that, if the polarization of the received data is the reverse of the transmitted data, coding still works. In baseband

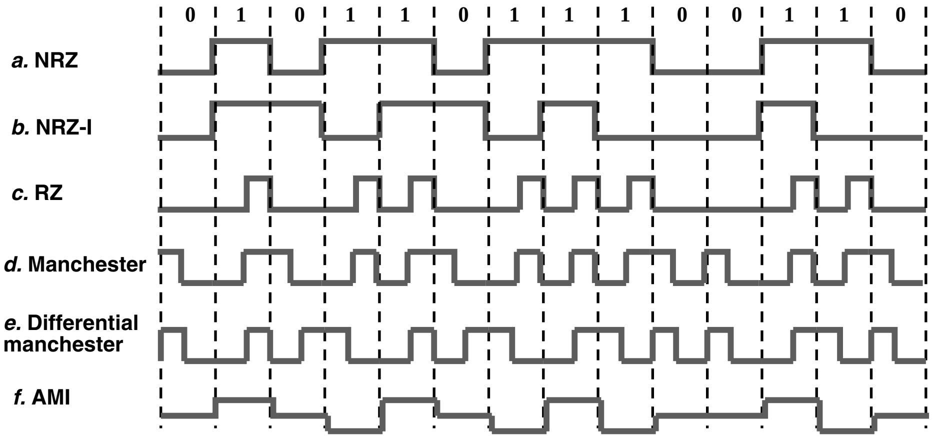


FIGURE 3.3 Examples of popular line coding techniques: (*a*) NRZ; (*b*) NRZ inverse; (*c*) RZ; (*d*) Manchester coding; (*e*) differential Manchester coding used in 802.3 Ethernet; (*f*) three-level AMI.

communications, receiver digital circuitry is often edge triggered, which better suits this type of coding.

NRZ in Fig. 3.3*a* means that the signal amplitude does stay at the same level throughout the transmission of a bit. In contrast, with return to zero (RZ) in Fig. 3.3*c*, during transmission of the symbol "1," in the middle of the transmission of a bit, the amplitude is changed. The transition in the middle of the bit in RZ coding assures that if we have a string of "1s," even then, for every symbol we have a transition enabling clock recovery. In general, the receiver synchronizes based on the transitions in the received signal; more transitions provide better synchronization at the receiver. There is an advantage of RZ over NRZ is these additional transitions. The disadvantage of the RZ approach is that the transmitted pulses are two times narrower, consuming twice the bandwidth of NRZ coding.

Figure 3.3*d* shows Manchester-coded information bits. In this coding technique, the digit "0" is represented by a pulse starting at a high level that switches to a low level in the middle of the transmission of the bit. The symbol for "1" is the inverse of "0," starting low and ending at the high amplitude. This code has the basic concepts of the RZ transmissions with the additional advantage that it has a transition for both symbols, whereas the transitions with RZ were only used on the digit "1." In optical communications, where we turn an LED on and off to transfer two digits for binary communications, RZ can be easily implemented by reducing the on time of the LED to a half. This will also reduce the power consumption of the signal to a half, which is very useful for applications such as remote controls, where we want to have a very long battery life. In remote controls for devices like TVs, to save even more in terms of the transmission power, rather than keeping the LED on for half of the time, we may turn it on and off several times [Pah95]. The edge-triggered variation of the Manchester code, called differential Manchester coding, is shown in Fig. 3.3*e*. In this code we always have a transition in the middle of the bit. However, if we have a "0" then we have a transition in the start of the bit and if we have "1" then there is no transition at the beginning of the bit. Differential Manchester coding was the first physical layer transmission technique recommended by the IEEE 802.3 Ethernet standardization committee.

Figure 3.3*f* shows a three-level line-coding technique called alternate mark inversion (AMI), in which we transmit the "1" symbols as a pulse with alternating polarities. In a manner similar to RZ, this approach provides for better synchronization, but unlike RZ it does not double the bandwidth of the system. In this section we have only referred to some basic examples and the issues which are involved in the design of line-coding techniques. More advanced line-coding techniques, such as ML-3, 4B3T, 8B10B, etc., are commonly used in design of fast and gigabit Ethernet, which will be described in Chapter 8.

3.2 TRANSMISSION TECHNIQUES AND SIGNAL CONSTELLATION

The POTS transmits the analog voice to the network where it is digitized before further transmission. The idea of ISDN, which was expected to bring digital transmission to the home, never became popular, and analog transmission with POTS remained as the dominant transmission technique to connect homes to the PSTN. To connect to the Internet, first voice-band modems and then high-speed or broadband xDSL and cable modems dominated the market. Voice-band modems and xDSL use the same telephone wiring which is used by POTS, and cable modems use the cable TV wiring. All of them use *bandpass* modem transmission techniques.

In wireless communications, the situation was different. The first-generation wireless cellular and cordless telephone systems used analog frequency modulation (FM). With the emergence of second-generation wireless networks, digital transmission techniques replaced analog transmission. To increase the capacity, analog voice was source coded into digital format at the mobile terminal for digital transmission over the wireless medium. Speech coding at the terminal also facilitates the integration of voice and data services in a single terminal. After the emergence of second-generation systems, digital transmission has become the dominant choice for wireless communication networks. As far as computer communication networks are concerned, Ethernet connected the LANs using baseband digital transmission techniques and IEEE 802.11 WLANs (the wireless Ethernet) used different bandpass transmission techniques. As a result of the popularity of these diversified PHY layers, network operations with an understanding of the PHY layer of the networks requires careful attention. In this book we are not concerned with analog techniques, and in the rest of this chapter we describe digital transmission techniques used for the implementation of the PHY layer of all of these modern digital communication networks.

Popular digital transmission techniques can be divided into four categories according to their applications. The first category is pulse transmission techniques used for baseband transmission in Ethernet, optical communication links and the so-called *impulse radio UWB* networks [Sch00]. The second category is the bandpass or carrier-modulated techniques widely used in TDMA cellular networks and a number of mobile data networks. The third category consists of spread-spectrum systems used in optical, some cable, and satellite communications, as well as the CDMA digital cellular networks and the original IEEE 802.11 WLANs operating in ISM bands. More recently, a fourth category of transmission technology has emerged to increase the data rates of all of these systems to support broadband access for popular Internet access and other data-oriented applications. Variations of the spread-spectrum technology and OFDM have been adopted by WLAN standard organizations and are being considered for incorporation into the future voice-oriented cellular networks. In the following sections we provide an overall description of all of these popular modulation techniques to provide the reader with an understanding of the applied physical layer alternatives.

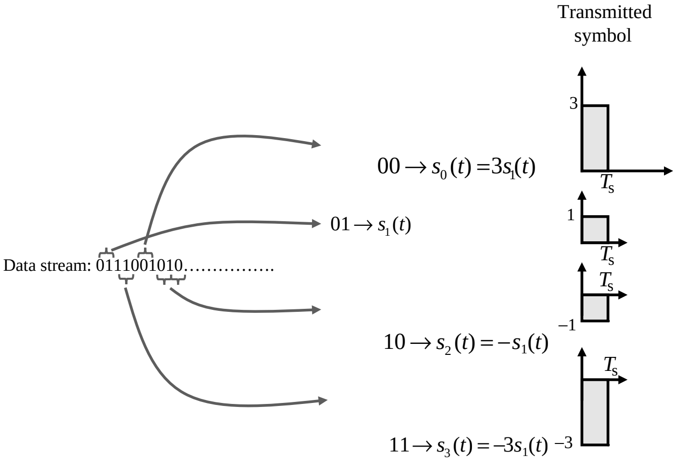


FIGURE 3.4 Basic concept of digital communications.

3.2.1 Multisymbol Digital Communications

In Fig. 3.4, symbols are transmitted every T_s seconds, which we may refer to as the symbol time. In digital communications these symbols are represented by a set of electrical waveforms, each representing one of the transmitting symbols.

Example 3.1: PAM Figure 3.4 shows a digital communication system using four symbols. Each symbol carries 2 bits of encoded data. The stream of data is packed into blocks of length 2 bits to associate with the transmitted symbols. Since the shape of the waveforms is the same and only their amplitude is different, we call this digital communication technique the *PAM* technique. The four possible pulses used in this system take (±1, ±3) amplitudes of the basic pulse s_1(t).

Example 3.2: RF Communications In RF communications, pulses are formed by sinusoids at different frequencies. In the computer communications literature these pulses are sometimes referred to as bandpass pulses, in contrast to baseband pulses. Figure 3.5 shows a 4-PAM system using RF pulses with a basic pulse shape of $s_1(t) = \sin \omega_c(t)$, $0 \leq t < T_s$. The center frequency of the RF pulse is $f_c = \omega_c/2\pi$. The advantage of RF pulses is that we can change the center frequency and have several of these streams operate on a single medium at the same time. This method of transmitting multiple streams over a single medium is referred to as FDM, which is used in all wireless networks. The wireless medium is not like wired media, where we always have the choice to add another cable at an additional cost; therefore, FDM is essential for wireless networks, while in wired media FDM is a luxury for efficient use of the medium.

At the receiver, the recovered bits are detected based on the shape of the received symbol. If the channel distorts the transmitted symbol significantly then it may be mistaken as another symbol at the receiver, causing an error. The probability of occurrence of this error is a function of harshness of the channel, strength or power of the transmitted signal, and the modulation techniques used for transmission. The digital communications literature [Pro00] provides details of a number of transmission techniques. These details include the relation between their error rate and the ratio of the transmitted power against the

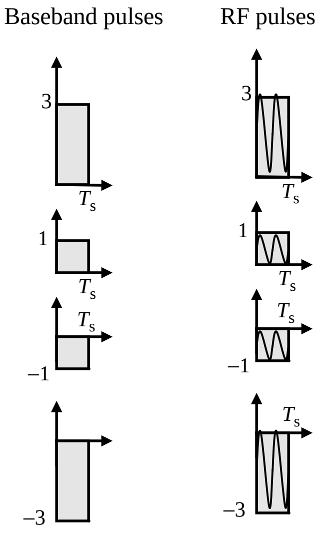


FIGURE 3.5 Baseband versus bandpass or RF transmission.

background noise, as well as sophisticated signal processing techniques to improve the performance.

3.2.2 Signal Constellation in Digital Communications

A popular graphical method to illustrate the relative energy of the symbols used in a digital communication system is to employ a "signal constellation." The signal constellation is a way of showing the signals as points in space rather than referring to the actual time-dependent waveform representing symbols, shown in Fig. 3.4. We use the square root of the energy of the signal as a representation of its amplitude, as shown in Figure 3.6. This way the transmitted symbols in the case of a PAM-based digital communication system are illustrated in a one-dimensional graph. The average energy per symbol is the average energy of the constellation, which is simply the average of the square of the distance of the points from the center of the coordinate system.

Table 3.1 provides a summary of important parameters that are used in digital communications to relate the transmission to the SNR of a transmission technique. The first set of parameters is related to the transmission rates and their relation to the number of bits and the number of symbols. The data transmission rate R_b is the coded information bit rate of the link in bits per second (bits/s). The symbol transmission rate $R_s = 1/T_s$ is the symbol transmission rate in symbols per second (S/s). If we use M symbols for digital

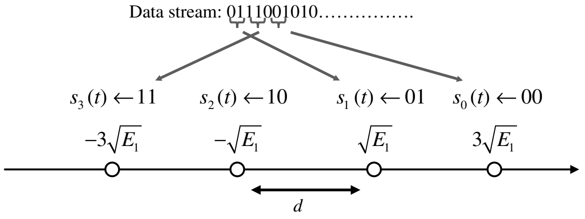


FIGURE 3.6 The signal constellation for four symbols (PAM).

TABLE 3.1 Summary of Important Parameters Used in Digital Communications

Data transmission rate	R_b
Symbol transmission rate	R_s
Number of symbols	M
Average number of bits per symbol	$m = \sum_{i=0}^{M} p_i m_i$
Energy per symbol	$E_{s_i} = \int \lvert s_i(t)\rvert^2 \, \mathrm{d}t$
Average energy per symbol	$E_\mathrm{s} = \sum_{i=0}^{M} p_i E_{s_i}$
Energy per bit	$E_\mathrm{b} = E_\mathrm{s}/m$
Variance of the noise	N_0
SNR	E_s/N_0
SNR per bit	E_b/N_0

communications and each symbol is transmitted with the probability of p_i and it carries m_i bits, then the average number of bits per symbol is given by

$$m = \sum_{i=0}^{M-1} p_i m_i \tag{3.1}$$

With an average of m bits per transmitted symbol, the bit rate $R_\mathrm{b} = mR_\mathrm{s}$.

The second set of parameters is related to the energy. The transmitted energy per symbol is given by

$$E_{s_i} = \int \left| s_i(t) \right|^2 \mathrm{d}t \tag{3.2}$$

and the average transmitted energy is

$$E_\mathrm{s} = \sum_{i=0}^{M-1} p_i E_{s_i} \tag{3.3}$$

where p_i is the probability of transmission of a symbol. For equally likely symbols (symbols transmitted with equal probability), $p_i = 1/M$. The average energy per bit is given by

$$E_\mathrm{b} = \frac{E_\mathrm{s}}{m} \tag{3.4}$$

Example 3.3: Signal Constellation and Energy per Symbol Figure 3.6 shows a signal constellation for the digital communication system described in Fig. 3.4 for which we have

$$M = 4, \quad m = 2, \quad p_0 = p_1 = p_2 = p_3 = \frac{1}{4}, \quad R_\mathrm{b} = 2R_\mathrm{s}, \quad E_\mathrm{b} = \frac{E_\mathrm{s}}{2}$$

$$E_\mathrm{s} = \sum_{i=0}^{3} \frac{1}{4} E_i = \frac{9+1+1+9}{4} E_1 = 5E_1$$

$$d = 2\sqrt{E_1} = \sqrt{\frac{4E_\mathrm{s}}{5}}$$

$$\therefore E_\mathrm{s} = 1.25d^2$$

This signal constellation can be implemented using baseband pulses for wired communications or RF pulses for wired or wireless data communications.

The constellation shown in Fig. 3.6 for PAM systems can be expanded for eight points with 3 bits per symbol and more points in a similar manner. Each time that we double the number of points in the constellation we send one more bit per symbol. It can be shown (see problems at the end of the chapter) that, for PAM systems, the relation between the average energy in the constellation and the minimum distance between the points is given by

$$E_{\mathrm{s},m} = \frac{4^m - 1}{12} d^2 \tag{3.5}$$

Since the exponential term on the numerator dominates over -1, we can say that, for a normalized minimum distance between points in the constellation, the required average energy is proportional to 4^m. This observation implies that addition of a bit to the bit per symbol of a PAM constellation requires four times (or $10\log 4 = 6\,\mathrm{dB}$) more transmission power to maintain the same minimum distance between the points in the constellation. We will see later that this minimum distance impacts the error probability.

Assuming that the variance of the background noise is denoted by N_0, the SNR per symbol and per bit will be given by E_s/N_0 and E_b/N_0 respectively. The SNR per symbol is representative of the physical transmitted signal power compared with the background noise power.

The received symbols are corrupted by noise, and if we map them to the constellation they will be typically in a location close to the transmitted symbol. On rare occasions, the received point in the constellation gets closer to another symbol, causing the receiver to make an error in detection. The probability of the symbol error can be approximated by [Pah05]

$$P_\mathrm{s} \approx \frac{1}{2}\mathrm{erfc}\left(\frac{d}{\sqrt{N_0}}\right) \geq \frac{1}{2}\mathrm{e}^{-d/\sqrt{N_0}} \tag{3.6}$$

where d is the distance between the symbols in the constellation, shown in Fig. 3.6.

Example 3.4: Energy in a 5-PAM System Figure 3.7 shows a signal constellation for the 5-PAM digital communication system. In this system, the first 2 bits of the stream of data are checked and if we have 00, 01, or 10 then we map them to their associated symbols in the middle of the constellation. If we have a 11, however, we wait for the next bit to form a block of 3 bits. The two possible 3-bit patterns of 110 and 111 are then mapped to the corner points

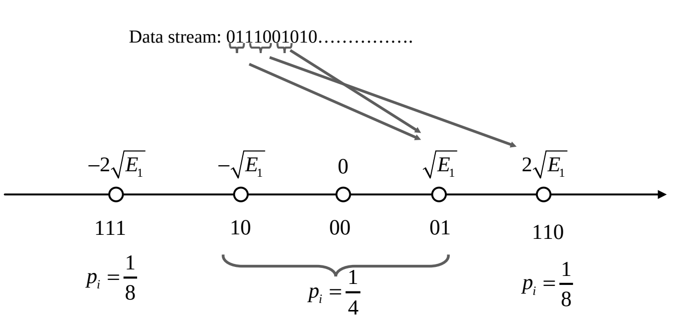


FIGURE 3.7 The signal constellation for 5-PAM.

of the constellation. For this constellation, $M=5$ and

$$p_0 = p_4 = \frac{1}{8}; \quad m_0 = m_4 = 3\ (\text{bits})$$
$$p_1 = p_2 = p_3 = \frac{1}{4}; \quad m_1 = m_2 = m_3 = 2\ (\text{bits})$$

Therefore:

$$m = \sum_{i=0}^{4} p_i m_i = 2 \times \frac{1}{8} \times 3\ (\text{bits}) + 3 \times \frac{1}{4} \times 2\ (\text{bits}) = 2.25\ (\text{bits})$$

and R_b=2.25R_s, which indicates that with the same duration of pulses (the same bandwidth) we have a system which provides a higher data rate than the 4-PAM system.

$$E_\mathrm{s} = \sum_{i=0}^{5} p_i E_i = \frac{1}{8} \times 4E_1 + \frac{1}{4} \times E_1 + \frac{1}{4} \times 0 + \frac{1}{4} \times E_1 + \frac{1}{8} \times 4E_1 = \frac{3}{2} E_1$$
$$d = \sqrt{E_1} = \sqrt{\frac{2E_\mathrm{s}}{3}} \Rightarrow E_\mathrm{s} = 1.5d^2$$

Therefore, for the same distance (the same symbol error rate), 5-PAM needs $1.5/1.25 = 1.2$ times the transmitted energy per symbol (the same as the power). Power is usually measured in decibels, in which we need $10\log(1.2) \simeq 0.8(\mathrm{dB})$ more power to support a 2.25 times higher data rate.

The 5-PAM constellation can be implemented for wired baseband communications, but it is not suitable for wireless communications because the symbol represented by zero energy cannot be easily created. In Chapter 8 we show that the variations of this basic concept are used in the design of a high-speed PHY layer for Ethernet.

3.2.3 Two-Dimensional Signal Constellations

Two-dimensional constellations are widely used in transmission systems used in telecommunication networks. Historically, they were first employed in the early 1970s in voiceband data communications to achieve 9600 bits/s transmission over four-wire telephone lines; today, they are employed in high-speed Ethernet, DSL modems, cable modems, wireless LANs and high data-rate cellular networks.

We start our discussions with an example.

Example 3.5: Two-Dimensional 16-QAM Constellation Assume we have two independent 4-PAM transmission systems each having their own four symbols and associated bits. We can create blocks of 4 bits and map each of 2 bits to one of the two 4-PAM systems. If we define a two-dimensional signal constellation for this system, then we can represent the constellation as shown in Fig. 3.8.

$$M = 16, \quad m = 4, \quad p_i \frac{1}{16}, \quad R_\mathrm{b} = 4R_\mathrm{s}, \quad E_\mathrm{b} = \frac{E_\mathrm{s}}{4}$$
$$E_\mathrm{s} = \sum_{i=0}^{15} \frac{1}{16} E_i = \frac{4 \times 18 + 8 \times 10 + 4 \times 2}{16} E_1 = 10E_1$$
$$d = 2\sqrt{E_1} = \sqrt{\frac{2E_\mathrm{s}}{5}}$$
$$\therefore E_\mathrm{s} = 2.5d^2$$

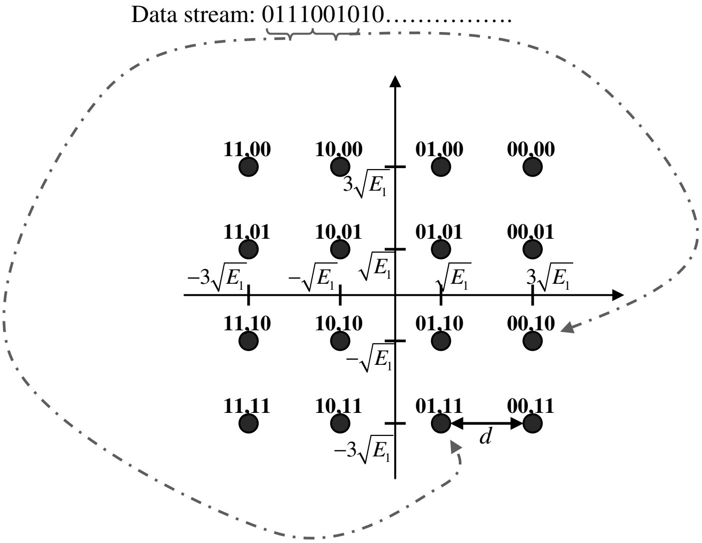


FIGURE 3.8 The two-dimensional signal constellation associated with the one-dimensional constellation shown in Fig. 3.5.

This constellation is called a QAM constellation because it uses two independent or orthogonal PAM systems. In this case it is 16-QAM with 16 signal points.

The constellation shown in Fig. 3.8 for QAM systems can be expanded to 64 points with 6 bits per symbol and more. Each time we double the number of points in the constellation we send one more bit per symbol. It can be shown (see problems at the end of the chapter) that for QAM systems the relation between the average energy in the constellation and the minimum distance between the points is given by

$$E_{s,m} = \frac{2^m - 1}{6} d^2 \tag{3.7}$$

where m is an even number. For odd values of m, the constellation cannot have a rectangular form. Since the exponential term on the numerator dominates the – 1, we can say that, for a normalized minimum distance between signal points, the required average energy is proportional to 2^m. This observation implies that the addition of a bit to the bit per symbol of a QAM constellation requires two times (or 10log2 = 3 dB) more transmission power to maintain the same minimum distance between signal points. Comparing the 3 dB per bit of QAM with the 6 dB per bit of PAM, we can conclude that the two-dimensional QAM is more power efficient than one-dimensional PAM. This additional gain is obtained because, on average, for the same minimum distance between the points of the constellation, in two-dimensional QAM constellation we can bring points closer to the center than the one-dimensional PAM constellation. Therefore, if we can implement a two- dimensional constellation, then we have an edge. But how can we implement a two-dimensional constellation?

In general, we can implement a two-dimensional modulation scheme for baseband or bandpass transmission techniques by using two different frequency bands or two different time slots. In wired networks we can also implement a two-dimensional constellation by using two different lines rather than one line for one-dimensional transmission. In the case of bandpass signals, we can use two *orthogonal* carrier frequencies in the same band to implement QAM signals. To describe this concept, first consider the implementation of

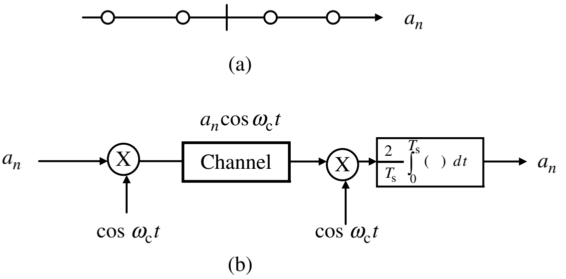


FIGURE 3.9 Bandpass implementation of one-dimensional constellations: (*a*) the signal constellation; (*b*) implementation of the transmitter and the receiver.

a one-dimensional constellation scheme using bandpass signals, as shown in Fig. 3.9. Consider the implementation with rectangular pulses shown in Fig. 3.9. The amplitude of the pulse obtained from the constellation is multiplied by the cosine of the carrier frequency ω_c to form the transmitted signal $a_n \cos\omega_c t$ for a duration T_s of the symbol. Assuming an ideal channel, the received signal is the same as the transmitted signal. To recover the transmitted symbols we multiply the received signal by the cosine *at the same frequency* and then integrate over the duration of the symbol. If the carrier frequency and the symbol duration are adjusted so that we have full cycles of the cosine during the symbol interval, then we have

$$\frac{2}{T_s}\int_0^{T_s} a_n \cos^2\omega_c t\,\mathrm{d}t = \frac{a_n}{T_s}\int_0^{T_s}(1+\cos 2\omega_c t)\,\mathrm{d}t = a_n$$

and the transmitted symbol is recovered successfully.[1]

The basic concept described in Fig. 3.9 can be extended to the implementation of two-dimensional constellations shown in Fig. 3.10. Each symbol in the constellation is identified

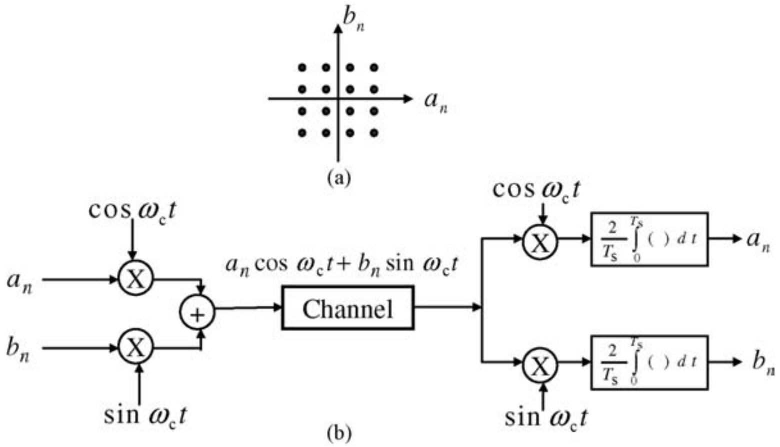


FIGURE 3.10 Bandpass implementation of two-dimensional constellation: (*a*) the signal constellation; (*b*) implementation of the transmitter and the receiver.

[1] Note that $2\cos^2\alpha = 1 + \cos 2\alpha$, $2\sin^2\alpha = 1 - \cos 2\alpha$, and $2\sin\alpha\cos\alpha = \sin 2\alpha$.

by a two-dimensional coordinate (a_n, b_n), referred to in the digital communication literature as in-phase and quadrature-phase symbols. The in-phase symbol is multiplied by a cosine and the quadrature-phase symbol by a sine; the resulting signals are added and then passed through the channel. The received signal $a_n \cos \omega_c t$ is passed through two braches, similar to the one-dimensional implementation for bandpass signals. The upper branch multiplies the signal by the cosine to recover the in-phase symbols:

$$\frac{2}{T_s}\int_0^{T_s}(a_n\cos^2\omega_c t + b_n\sin\omega_c t\,\cos\omega_c t)\,\mathrm{d}t = \frac{a_n}{T_s}\int_0^{T_s}(1+\cos 2\omega_c t)\,\mathrm{d}t + \frac{b_n}{T_s}\int_0^{T_s}\sin 2\omega_c t\,\mathrm{d}t = a_n + 0 = a_n$$

and the lower branch recovers the quadrature-phase symbol as follows:

$$\frac{2}{T_s}\int_0^{T_s}(a_n\cos\omega_c t\,\sin\omega_c t + b_n\sin^2\omega_c t)\,\mathrm{d}t = \frac{a_n}{T_s}\int_0^{T_s}\sin 2\omega_c t\,\mathrm{d}t + \frac{b_n}{T_s}\int_0^{T_s}(1-\cos 2\omega_c t)\,\mathrm{d}t = 0 + b_n = b_n$$

The advantage of the two-dimensional bandpass implementation shown in Fig. 3.10 over the one-dimensional implementation is that the two-dimensional bandpass implementation can be used to send twice as many bits per symbol in the same bandwidth and the same minimum distance between the symbols. As we discussed earlier, a rectangular two-dimensional constellation consumes 3 dB per additional bit, whereas the one-dimensional version needs 6 dB per bit. If we represent the bandwidth efficiency (sometimes called the spectrum efficiency) by η and the bandwidth by W, the bandwidth efficiency is defined as

$$\eta = \frac{R_b}{W}$$

where R_b is the bit rate of the data stream. The bandpass implementation has a separate advantage allowing the multiplexing of two streams over the same band, resulting in doubling the bandwidth efficiency of the transmission technique. The unit of bandwidth efficiency is bits per second per hertz; indeed, it is the data rate of the system normalized by the bandwidth. If we assume that the symbol transmission rate is the same as the bandwidth (usually it is of the same order), $R_s = W$, then $\eta = (R_b/R_s) = m$, which is the average number of bits per points of the constellation, which can have the same unit as bandwidth efficiency.

We can now provide some examples of applying the signal constellation concept for the design of the physical layer of networks.

Example 3.6: 64-QAM in IEEE 802.11g WLANs Figure 3.11 shows the 64-QAM signal constellation recommended by IEEE 802.11g for 54 Mb/s data transmission. This constellation has 6 bits per symbol and the symbol transmission rate is 12 kS/s. Therefore, the base transmission rate is 6(bits/S) × 12(kS/s) = 72(Mb/s). However, the transmitted data is encoded by a rate $R = 3/4$ convolution codes, which turn the effective data rate to 72 (Mb/s) × 3/4 = 54(Mb/s). Convolution codes add the equivalent of 1 bit to every 3 bits of data stream to increase its integrity and provide for error corrections. These codes are very popular in all wireless networks.

Example 3.7: TCM in Voice-band Modems Figure 3.12 shows the signal constellation used for CCITT standard V.32 voice-band modems operating at 9600 bits/s. The symbol transmission rate of the modem is 2400 S/s and the modem uses TCM. This coding

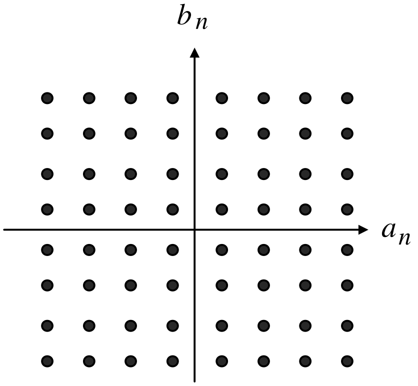


FIGURE 3.11 QAM constellations for 4 and 6 bits/S.

technique doubles the number of bits per constellation (equivalent to adding 1 bit per symbol) but correlates the transmitted symbols using a pattern that actually improves the overall performance by 3–6 dB. The constellation has $2^5 = 32$ symbols with 5 bits per symbol, but one the bits is used for the coding overhead, resulting in an effective data rate of 4(bits/S) × 2400(S/s) = 9600(bits/s). Two other interesting issues related to this constellation are that the constellation has a cross shape because five of the bits are odd and with that number of bits we cannot create a square constellation similar to Fig. 3.11. The constellation is rotated 45°, which allows one to use a subset of points in the constellation to implement a simple binary constellation for the start-up procedure of the modem. During the start up these modems need simple modulation techniques to initiate the adaptive signal processing algorithms. A similar cross pattern with 128 points is also used as a standard for the 14.4 kb/s modems (see the problems associated with this chapter).

Example 3.8: The 5 × 5 PAM Constellation in Ethernet Figure 3.13 shows the signal constellation recommended by the IEEE 802 community for several versions of the Ethernet. In that community this constellation is referred to as 5 × 5 PAM. This is the

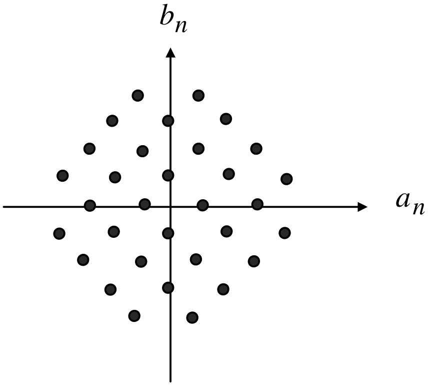


FIGURE 3.12 Signal constellation for TCM-coded 9600 bits/s modems recommended by CCITT for four-wired voice-band operation.

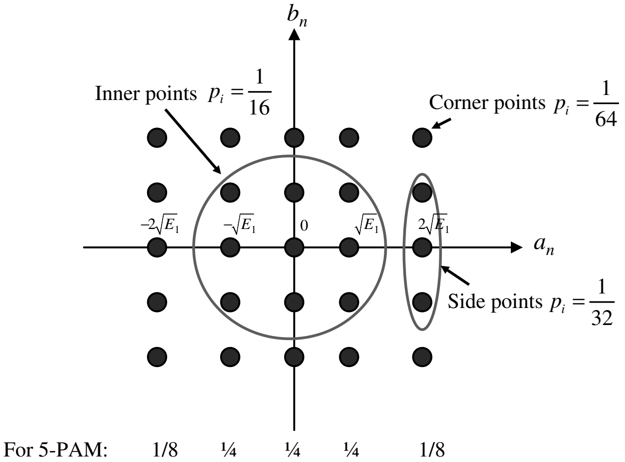

b_n
Inner points $p_i = \frac{1}{16}$
Corner points $p_i = \frac{1}{64}$
$-2\sqrt{E_1}$
$-\sqrt{E_1}$
0
$\sqrt{E_1}$
$2\sqrt{E_1}$
a_n
Side points $p_i = \frac{1}{32}$
For 5-PAM: 1/8 ¼ ¼ ¼ 1/8

FIGURE 3.13 5 × 5 PAM signal constellation used in different versions of Ethernet.

two-dimensional version of the 5-PAM shown in Fig. 3.7. For this constellation we have

$$M = 25$$

$$\text{for inner points :} \qquad p_i = \frac{1}{16}, \quad m_i = 4\ \text{(bits)}$$

$$\text{for side points :} \qquad p_i = \frac{1}{32}, \quad m_i = 5\ \text{(bits)}$$

$$\text{for corner points :} \qquad p_i = \frac{1}{64}, \quad m_i = 6\ \text{(bits)}$$

$$m = \sum_{i=0}^{24} p_i m_i = 9 \times \frac{1}{16} \times 4 + 12 \times \frac{1}{32} \times 5 + 4 \times \frac{1}{64} \times 6 = 4.5\ \text{(bits)}$$

$$R_\mathrm{b} = 4.5 R_\mathrm{s}, \quad E_\mathrm{b} = \frac{E_\mathrm{s}}{4.5}$$

$$E_\mathrm{s} = \sum_{i=0}^{24} p_i E_i = 0 + 4 \times \frac{1}{16} \times E_1 + 4 \times \frac{1}{16} \times 2E_1 + 4 \times \frac{1}{32} \times 4E_1 + 8 \times \frac{1}{32} \times 5E_1$$
$$+ 4 \times \frac{1}{64} \times 8E_1 = 3E_1$$

$$d = \sqrt{E_1} = \sqrt{\frac{E_\mathrm{s}}{3}}$$

$$\therefore E_\mathrm{s} = 3 \times d^2$$

To map the binary data stream to this constellation one may treat the stream as two sequential streams of 5-PAM-encoded data using the encoding technique described in Example 3.4. Similar to that example, if the first 2 bits in the stream are 00, 01, or 10 then they are assigned to a_n values representing the three middle horizontal coordinates; if it is 11, then the next bit is read to make the a_n assignment. After completion of horizontal bit assignment, the next 2 or 3 bits in the stream are read and mapped with the same rules to determine the appropriate b_n value to identify the point in the constellation. Figure 3.14

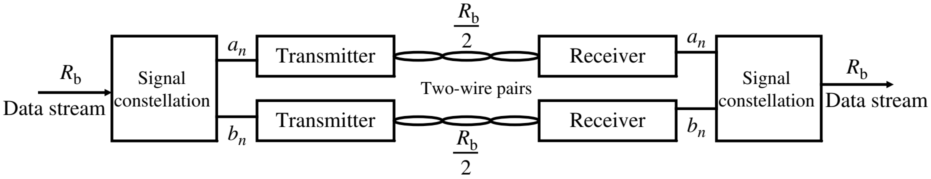


FIGURE 3.14 Baseband implementation of a two-dimensional constellation on two pairs of wires.

shows a block diagram for implementation of this constellation over two pairs of wires. Comparing this with just the 5-PAM transmission scheme, this constellation has two times higher data rates at the expense of 10 log(3/1.5) = 3 dB more power. Like the 5-PAM scheme, this constellation is only good for baseband data transmission that is restricted to wired transmission and we need two sets of wires to implement it. When we use two independent 5-PAMs over the two sets of wires, then we would have to double the transmission power (3 dB more power).

3.2.4 Channel Capacity

In our discussions thus far we have shown how the design of the signal constellation relates to the achievable data rate and the bandwidth and energy efficiency. At this time, it is useful to consider the ultimate limits on data rate and efficiency that are theoretically achievable. This is best done by examining Shannon–Hartley's well-known formula:

$$C = W \log_2\left(1 + \frac{E_s}{N_0}\right) \text{ bits/s}$$

where $C = R_{b\text{-max}}$ is the maximum achievable information transfer rate in bits per second in a bandwidth of W Hz, and the signal-to-noise power ratio of E_s/N_0 [Sha48]. This simple and powerful theorem relates the three most important transmission parameters, namely power, bandwidth, and maximum achievable data rate, and it is the considered the ultimate guideline for understanding the limits to the performance of a transmission technique. For example, if we apply the above equation to the voice-band telephone channel which has a bandwidth of 4 kHz and a typical SNR of 30 dB, then we obtain a theoretical channel capacity of about 40 kb/s. If we want to increase the data rate, since bandwidth and the maximum transmitted power are fixed, then we need to improve the line conditions to reduce the noise at the receiver. In V.90 modems this is done in the downstream channel by eliminating the analog circuitry in the connection between the user and the PSTN.

If we express the equation in terms of bandwidth efficiency we have

$$\eta_{\max} = \frac{C}{W} = \log_2\left(1 + \frac{E_s}{N_0}\right) \text{ bits/(s Hz)} \tag{3.8}$$

in which $\eta_{\max}$ is the maximum bandwidth efficiency. As we explained in Section 3.2.3, in symbol transmission techniques, we can reasonably replace the bandwidth by the symbol transmission rate R_s, in which case $\eta_{\max} = m_{\max}$, the maximum number of average bits per symbol of a constellation. This simple bound provides us with an idea to relate the bandwidth, data rate, and the SNR to find achievable transmission rates over a specific channel.

In Eq. (3.8), if we use

$$\frac{E_\mathrm{s}}{N_0} = m_\mathrm{max}\frac{E_\mathrm{b}}{N_0} = \frac{C}{W}\frac{E_\mathrm{b}}{N_0} = \frac{C}{W}\gamma_\mathrm{b}$$

then we will have

$$\frac{C}{W} = \log_2\left(1+\frac{C}{W}\frac{E_\mathrm{b}}{N_0}\right) = \log_2\left(1+\frac{C}{W}\gamma_\mathrm{b}\right)$$

Or

$$\gamma_\mathrm{b} = \frac{E_\mathrm{b}}{N_0} = \frac{1}{C/W}(2^{C/W}-1)$$

It will be useful, therefore, to determine how closely this limit can be approached with various signal constellations.

This equation describes channel capacity in terms of two convenient normalized parameters, $\gamma_\mathrm{b} = E_\mathrm{b}/N_0$ and $\eta = C/W = R_\text{b-max}/W$. The first parameter is the minimum value of *SNR per bit* required for reliable transmission of data at capacity over a channel with bandwidth W. The second parameter, $\eta = R_\mathrm{b}/W$, simply normalizes the bit rate normalized by the bandwidth. Therefore, this equation relates the achievable data rate to the bandwidth of the system and the SNR, which provides us with a convenient framework for assessing the communications efficiency of any chosen modulation scheme.

In Fig. 3.15 we show the capacity formula as a plot of $\eta = R_\mathrm{b}/W$ versus $\gamma_\mathrm{b} = E_\mathrm{b}/N_0$. Note that the lower portion of the scale is expanded for convenience in drawing the figure. This figure essentially represents a plane of bandwidth versus efficiency, and the capacity curve divides the plane into two regions. The shaded area to the left of the curve defines the region in which reliable communication cannot be achieved; that is, no modulation or coding scheme can be devised to operate in that region with low error rates in delivered data. In the right-hand area of the figure, which defines the region of achievable signal designs, design points are shown for several modulation methods, which we have discussed earlier. For all the cases shown, the delivered BER is 10^{-5}. The displacement of each design point from the capacity boundary indicates how close the communication efficiency of the corresponding modulation scheme comes to the capacity limit. The horizontal displacement measures the shortfall in terms of SNR per bit, while the vertical displacement measures the shortfall in terms of bandwidth utilization. Note that the points would all move to the right (i.e. further away from the capacity boundary) if we were to plot the modem design points for a lower level of delivered BER, whereas they would move closer to the capacity boundary if we used a higher BER.

It is conventional to call the region of $R/W > 1$ the *bandwidth-limited region* of operation and to call the region of $R/W < 1$ the *power-limited region* of operation. The bandwidth-limited region includes all the modulation schemes we have described for use on voice-band telephone circuits, where rigid channel bandwidth limitations are imposed by the existing design of the public network. There, we see that the M-ary modem signal constellations provide steadily increasing bandwidth utilization as M is increased. It can be seen from the figure that the QAM schemes are closest to the capacity boundary.

By inspecting the figure we can conclude that if the bandwidth is much more than the data rate (e.g. original direct-sequence spread-spectrum 802.11 with a 2 Mb/s data rate and 26 MHz bandwidth) then we can operate at a smaller SNR per bit (which results in larger

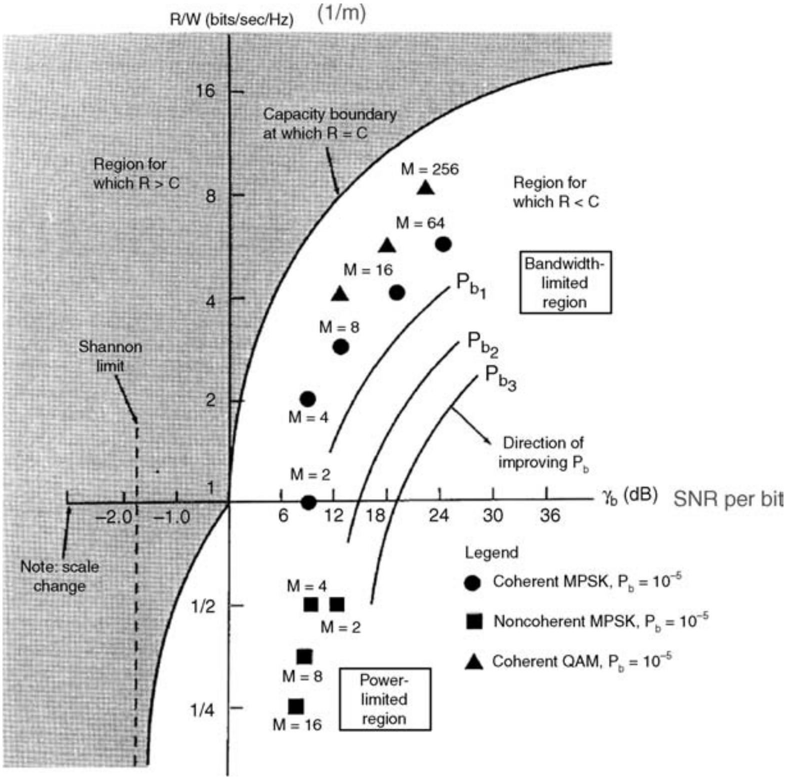


FIGURE 3.15 Channel capacity and comparison of several modulation methods at a BER of 10^{-5}. (*Source*: [Skl01] © Prentice Hall PTR.)

coverage for the radio). If we have small bandwidth and we need high data rates (e.g. voice-band modems with 4 kHz bandwidth and 36 kb/s modems) then we need to use multiple symbols and a high SNR per bit.

3.3 PERFORMANCE OF THE PHYSICAL LAYER

In a wired or wireless transmission scheme, the transmitted symbols are corrupted by additive background noise. In digital communications, the additive noise causes erroneous decisions in detecting the transmitted symbols at the receiver. To measure the performance of these transmission techniques, the symbol transmission error P_s or probability of bit transmission error P_b in logarithmic form is often plotted against the SNR $\gamma_s = E_s/N_0$ or SNR per bit $\gamma_b = E_b/N_0$ in decibels for the particular transmission technique. These plots allow us to determine the required level of the transmitted power to achieve a certain BER, which is a subjective criterion imposed by the application. For example, in digital voice transmission we may accept error rates on the order of 10^{-2} (1 in 100 bits) while for voice-band data communications we expect error rates on the order of 10^{-5} and for applications in wired LANs one may think of error rates lower than 10^{-8}. SNR is a measure of how much received power is required to detect information in a signal correctly with a given probability. As we described in Section 3.2.3, the SNR is related to the energy in the constellation and the variance of the background noise as

$$\gamma_s = \frac{E_s}{N_0} = \frac{R_b}{W}\frac{E_b}{N_0} = m\frac{E_b}{N_0} \tag{3.9}$$

To relate the SNR to the error rate, we start with Eq. (3.6), which relates the probability of occurrence of a symbol detection error to the minimum distance in the constellation and the variance of the background noise. On the other hand, in all our examples on signal constellation we have derived a linear relation between the average energy in the signal constellation and the square of the minimum distance. We can represent this relation by the general equation $E_s = \alpha d^2$, in which α is a parameter which differs among different constellations. For example, $\alpha = 2$ for the 5-PAM constellation and $\alpha = 2.5$ for 16-QAM. Substituting this overall relation into Eq. (3.6) we have

$$P_s \approx \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{E_s}{\alpha N_0}}\right) \geq \frac{1}{2}\mathrm{e}^{-E_s/\alpha N_0} \tag{3.10}$$

which relates the probability of symbol error to the SNR per symbol. To find out the relation between the BER and the symbol error rate, consider Fig. 3.8. If a symbol is erroneously detected as its neighboring symbol, then depending on the string of the bits which represent that symbol most of the time we make only one error (e.g. we detect the received symbol as corresponding to 0100 or 0001 instead of 0000). But rarely, we may yet make more than one bit error (e.g. to detect it as 1001 instead of 0101). In general, when we assign bit blocks to the points in the constellation, we pay attention to minimize the bit error occurrences in neighboring points in the constellation. This procedure is referred to as Gray coding, which assigns bit blocks to neighboring points in the constellation so that they are only different in one bit of the block. This way, if a symbol transmission error occurs, only one bit will be in error and all other bits remain the same. As an example, consider the top row of the points in the constellation shown in Fig. 3.8. Any symbol error between the points and neighboring points causes only one bit to be in error. Based on this discussion, we can assume that the approximated symbol error rate given by Eq (3.6) is actually a good approximation of the BER as well. Therefore, we may refer to the symbol error rate or BER as the probability of error and assume that $P_e = P_s \simeq P_b$.

Equation (3.8) is good for comparison of different constellations and their relative performance based on the same transmitted power. Constellations, however, may have different bits per symbol and represent different data rates. In the literature, the SNR is sometimes normalized by the number of bits per symbol by using the fact that $E_s = mE_b$, in which case Eq. (3.9) changes to

$$P_e \approx \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{m}{\alpha}\gamma_b}\right) \geq \frac{1}{2}e^{-(m/\alpha)\gamma_b} \tag{3.11}$$

In this format, $\gamma_b = E_b/N_0$ is referred to as the SNR per bit normalizing the transmitted power by the number of bits, making the comparisons based on the same data rate. In the

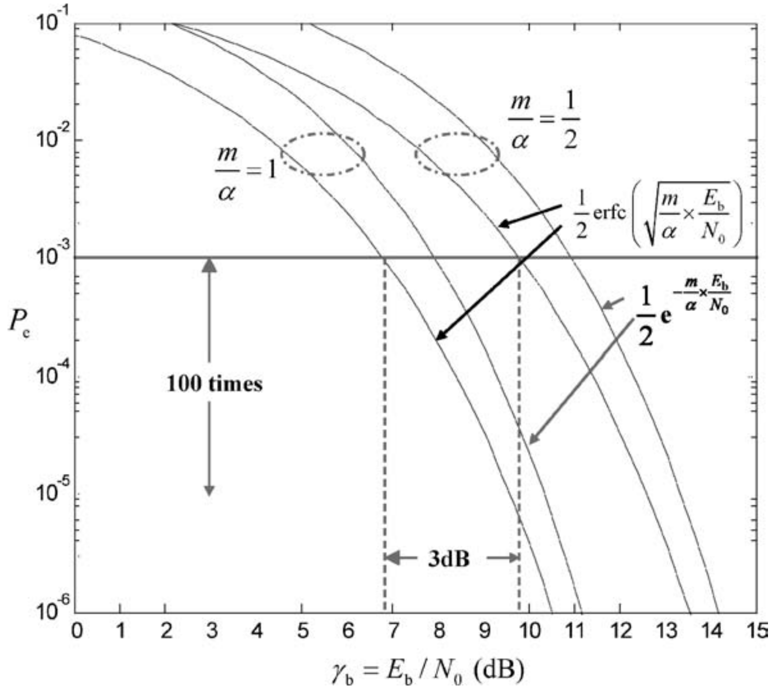


FIGURE 3.16 BER in logarithmic form versus SNR/bit in decibels for two values of m/α.

literature it is customary to plot P_e in logarithmic form versus γ_b in decibels. Figure 3.16 gives an example of such plots produced with Matlab code.[2]

In Fig 3.16 we have plots of erf and its exponential bounds in Eq. (3.10) for two values of $m/\alpha = 1$ and $m/\alpha = 1/2$. These plots are very popular in the digital communication literature to compare the performance of different physical layer alternatives. There are a few

[2] Matlab code similar to the following is used for generation of these plots:

```
gb_db = linspace(0, 30, 100);
gb  = 10.^(gb_db/10);
pe_0 = 1e-5;

% m/∝ = 1
a = 0.5;
b = 1;
pe_bpsk = a * erfc (sqrt(b * gb));
gb_0 = (1/b) * (erfinv(1-(pe_0/a)))^2;
gb_bpsk_db = 10*log10(gb_0);
figure(1)
semilogy(gb_db, pe_bpsk)
hold on
semilogy([0 15], [1e-5 1e-5],‘r’)
axis([0 15 1e-6 1e-1])
set(gca, ‘XTick’, [0:1:15])
xlabel(‘SNR/bit (dB)’)
ylabel(‘BER’)
```

interesting observations from this figure. The first observation is that the difference between the set of erfc plots or exponential plots is 3 dB, which was what we expect because they are different by a factor of two. The second observation is that the exponential provides an asymptotic bound to the erfc function. This means that, with an increase in SNR and reduction in the error rate, the two curves get closer to one another and they ultimately merge when the error rate is close to zero. The third observation is that the error rate drops 100 times for each 3 dB of additional SNR. For example, if we want to improve the error rate from 10^{-3} to 10^{-5} we need 3 dB or two times more transmission power with the same background noise. These are very general and useful observations for systems engineering applications to relate the data rate, power, and error rate to one another. The purpose of the plots is to demonstrate how much performance degradation results when we design simpler receivers.

3.3.1 Effects of Fading on Performance over Wireless Channels

One of the main characteristics of the wireless medium affecting the performance of a transmission technique is large fluctuations of the received power level, referred to as *fading*. As opposed to wired channels, the received signal from wireless channels suffers from strong amplitude fluctuations (on the order of 30–40 dB) that cause fading in the received signal. The error rate of the transmission system increases substantially during periods of signal fading, and the error rate becomes negligible when the system is out of fade.

Figure 3.17 shows the basic concept behind the strange and complex behavior of transmission techniques over fading wireless channels. Owing to the fading effect the SNR per bit γ_b fluctuates randomly in time. The average SNR per bit γ_b represents the transmitted power which can be regulated by the designer. Any application has an acceptable error rate, when γ_b crosses a specified "threshold" and the error rate drops below the acceptable error rate. The fraction of time in which the error rate is unacceptable is called the *outage rate* or outage probability for those applications. The error rate when the signal is above the threshold is always very small (close to zero), and when it is below the threshold it is very high (close to 0.5). Therefore, most errors occur during deep fades when the signal level crosses the threshold. This is a very important observation; to remedy the effects of fading we need to find methods that can recover bits corrupted during occurrence of deep fading.

To evaluate the performance over a fading radio channel, either the average BER $\bar{P}_e$ or the probability of outage versus average received SNR per bit $\bar{\gamma}_b$ is used. The average BER and

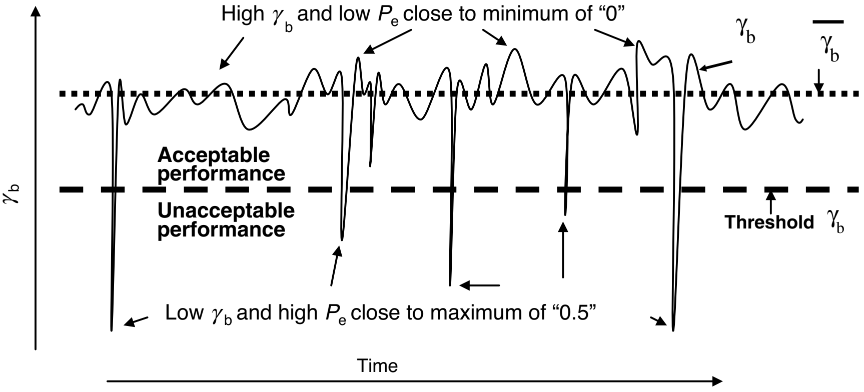

High γ_b and low P_e close to minimum of "0"
γ_b
$\bar{\gamma}_b$
γ_b
Acceptable performance
Unacceptable performance
Threshold γ_b
Low γ_b and high P_e close to maximum of "0.5"
Time

FIGURE 3.17 Relation between error rate, outage rate and fading characteristics.

the outage rate may look different. However, they represent the same phenomenon, and in many cases they look similar. To calculate any of these performance measures we need a model for the variations of $\bar{\gamma}_b$. The model most commonly used for variations of signal amplitude in fading channels is the Rayleigh distribution, for which $\bar{\gamma}_b$ follows an exponential distribution:

$$f_\Gamma(\gamma_b) = \frac{1}{\gamma_b} e^{-\gamma_b/\bar{\gamma}_b}$$

in which γ_b is the instantaneous SNR per bit and $\bar{\gamma}_b$ is the average SNR per bit over a long time. If we consider the exponential asymptotic bound for errors described in Eq. (3.10), then the average BER $\bar{P}_e$ over all possible values of the SNR per bit γ_b is given by

$$\bar{P}_e = \frac{1}{2\bar{\gamma}} \int_0^\infty e^{-\gamma_b/\bar{\gamma}_b} e^{-(m/\alpha)\gamma_b}\, d\gamma_b = \frac{1/2}{1+\dfrac{m}{\alpha}\bar{\gamma}_b} \simeq \frac{1/2}{\dfrac{m}{\alpha}\bar{\gamma}_b} \tag{3.12}$$

The average BER $\bar{P}_e$ is an inverse function of the average SNR per bit $\bar{\gamma}_b$. To compare this relation with the relation between the error rate P_e and SNR per bit γ_b for nonfading wired channels, remember that a 3 dB change in transmission power in nonfading channels decreased the BER by two orders of magnitude (hundred times). Equation (3.12) reveals that we need 20 dB (100 times) more average power to increase the average BER by two orders of magnitude for fading channels. Figure 3.18a shows the average probability of error in logarithmic form versus average SNR per bit in decibels for $m/\alpha = 2$. This relation is a line with slope of unity. Comparing this curve with the waterfall-like plots of BER versus signal to noise per bit, shown in Fig 3.18b, we need much more power to overcome the effects of

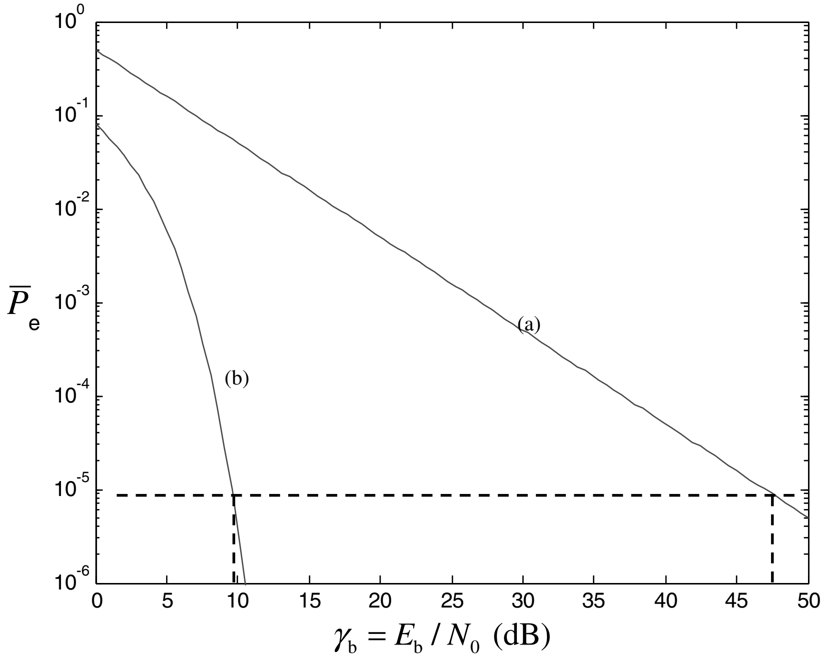


FIGURE 3.18 Average BER in logarithmic form versus average SNR/bit in decibels for $m/\alpha = 2$: (*a*) over a Rayleigh fading wireless channel; (*b*) over a nonfading wired channel.

fading to achieve the same error rates. For example, to achieve an error rate of 10^{-5} we need a γ_b of less than 10 dB on a nonfading wired channel compared with an average signal to noise per bit $\bar{\gamma}_b$ or more than 45 dB needed to achieve the same average error rate on a fading wireless channel. This difference means that achievement of the same error rate for the radio channel requires > 3000 times (35 dB) more power. Designers of radio modems have worked hard in the past half a century to close this gap between the performance over nonfading wire and fading wireless channels and they have come up with a number of innovative solutions such as STC and MIMO antenna systems, which are used as the latest advancement in design of the physical layer of wireless networks. All of these solutions take advantage of the so-called diversity techniques that we will discuss next.

3.3.2 Diversity Techniques

As we observed in the previous section, fading is manifested as signal amplitude fluctuations over a wide dynamic range. In particular, during short periods of time, the channel goes into deep fades causing significant numbers of errors that virtually dominate the overall average error rate of the system. In order to compensate for the effects of fading when operating with a fixed-power transmitter, the power must typically be increased by several orders of magnitude relative to nonfading operation. This increase of power protects the system during the short intervals of time when the channel is deeply faded. A more effective method of counteracting the effects of fading is to use diversity techniques in transmission and reception of the signal. The concept here is to provide multiple copies of the received signals whose fading patterns are different (and hopefully independent, as we will see later). With the use of diversity, the probability that all the received signals are in a fade at the same time reduces significantly, which in turn can yield a large reduction in the average error rate of the system.

Figure 3.19 shows fluctuations in two branches of a diversity channel and how they help in the reduction of overall error rates. When one of the branches is in deep fade, causing a large number of errors, the correct data can be retrieved from the other branch. In a diversity

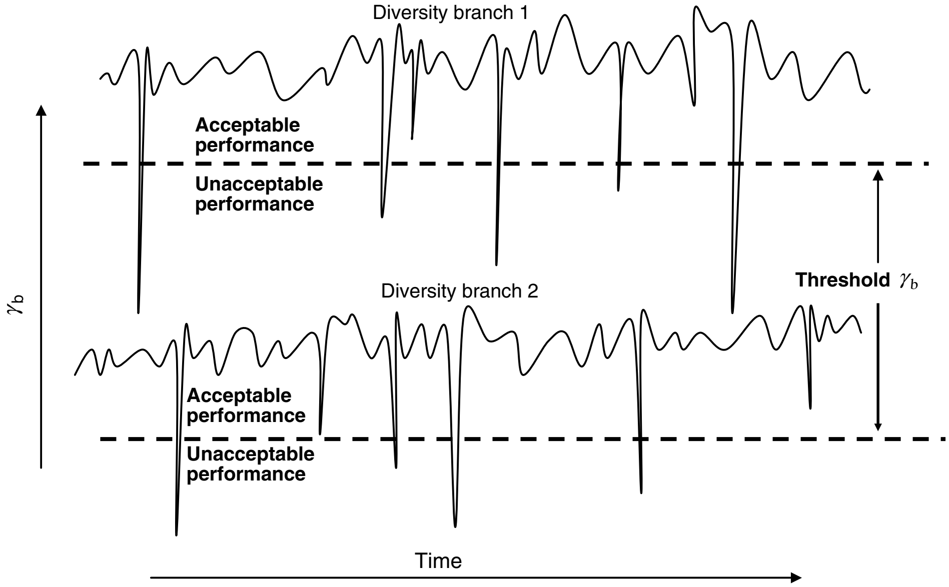


FIGURE 3.19 Fading in two branches of a diversity channel.

channel, large numbers of errors can occur when all branches are in deep fade at the same time. Since the probability of a deep fade occurring in all branches is much lower than in only one branch, the error rate on a diversity channel is much less than on a single-branch fading channel. The occurrence of deep fading on all branches is a function of correlation among different branches and of the number of diversity channels. As the correlation among the diversity branches decreases and they become independent and the number of branches increases, the error rate decreases.

Diversity can be provided spatially by using multiple antennas, in frequency by providing signal replicas at different carrier frequencies, or in time by providing signal replicas with different arrival times. It is conventional to refer to the diversity components as *diversity branches*. We assume that the same symbol is received along different branches, with each branch exposed to a *separate* random fluctuation. This has the effect of reducing the probability that the received signal will be faded simultaneously on all the branches; this in turn reduces the overall outage probability, as well as the average BER.

A variety of techniques are available for reception of diversity signals. In the most popular and the optimum method of combining, called *maximal-ratio combining* (MRC), the diversity branches are weighted prior to summing them, each weight being proportional to the received branch signal amplitude.

Let us assume that the amplitudes of the signals received on different branches are all uncorrelated Rayleigh-distributed random variables and all branches of diversity have the same average received signal power and that the average SNR on each branch is denoted by $\bar{\gamma}_{\mathrm{b}}$. The probability distribution function of the post-combining of a maximum ratio combiner SNR is then given by the gamma function

$$f_{\Gamma}(\gamma_{\mathrm{b}}) = \frac{1}{(D-1)!\bar{\gamma}_{\mathrm{b}}{}^{D}} \gamma_{\mathrm{b}}^{D-1} \mathrm{e}^{-\gamma_{\mathrm{b}}/\bar{\gamma}_{\mathrm{b}}} \tag{3.13}$$

where D is the order of diversity. It can be shown [Pro00, Pah05] using Eqs (3.10) and (3.13) that the average probability of error for the maximal-ratio combiner output is given by

$$\bar{P}_{\mathrm{e}} = \int_{0}^{\infty} f_{\Gamma}(\gamma_{\mathrm{b}}) P_{\mathrm{b}}(\gamma_{\mathrm{b}})\,\mathrm{d}\gamma_{\mathrm{b}} \approx \left(\frac{1}{8\frac{m}{\alpha}\bar{\gamma}_{\mathrm{b}}} \right)^{D} \begin{pmatrix} 2D-1 \\ D \end{pmatrix} \tag{3.14}$$

where we use the standard notation for a binomial coefficient:

$$\begin{pmatrix} N \\ k \end{pmatrix} = \frac{N!}{(N-k)!k!}$$

The expression in Eq. (3.14) shows that average BER performance at the maximal-ratio combiner output improves exponentially with increasing D, the order or number of branches of diversity. Figure 3.20 shows the average probability of error $\bar{P}_{\mathrm{e}}$ versus average SNR per bit $\bar{\gamma}_{\mathrm{b}}$ for different orders of diversity. Included in the figure is the error rate curve for steady-signal reception. As we saw earlier, with a single antenna we lose 30–35 dB in performance relative to steady-signal reception at reasonable levels of error rate. With two independent diversity branches, the performance loss is reduced to about 25 dB, and with four orders of diversity the signal-to-noise penalty is reduced to around 10 dB. With additional orders of diversity the penalty relative to nonfading can be further reduced. There will, of course, be a practical limit to the order of diversity implemented because, for

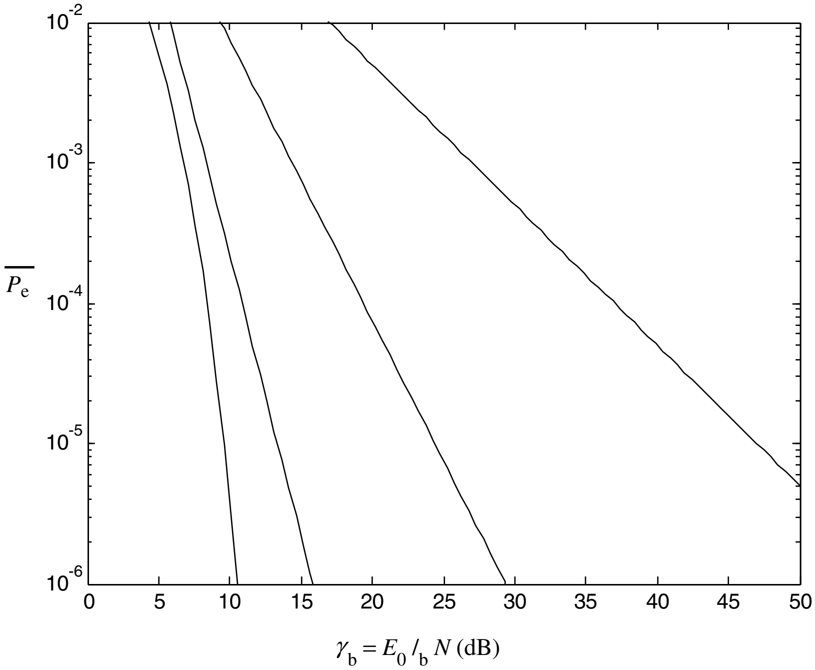


FIGURE 3.20 Average BER versus average SNR per bit for different orders of diversity.

example, one cannot put an arbitrarily large number of antennas into a communications terminal.

3.4 WIDEBAND MODEMS

Another major difference between wireless and wired channels is that the wireless channel is a multipath channel. In a multipath channel, the shape of the received pulse and the time duration of the signaling pulse are both changed due to multipath arrivals. The difference between the first and the last arriving pulses at the receiver due to the same transmitted pulse is the *delay spread* of the channel. If the symbol duration is much larger than the multipath spread of the channel, then all pulses received via different paths arrive roughly on top of one another, causing only amplitude fluctuations and fading that was discussed in the previous section. If the ratio of the delay spread to the pulse duration becomes considerable, then the received pulse shape is severely distorted and it also interferes with neighboring symbols, causing ISI. In addition to SNR fluctuations due to fading effects, the interference power also degrades the performance. However, the ISI effect of multipath degrades the performance in a different manner than fading. The effects of fading can be compensated via an increase in the transmit power by a fading margin. However, increasing the transmit power cannot compensate for the effects of ISI. This is because an increase in the transmit power increases the signal as well the ISI interference power, keeping the signal-to-interference ratio at the same level.

The effect of ISI caused by multipath is the main obstacle for high-speed communications over wireless channels. As we increase the data rate, the duration of the transmitted symbols decreases and the ISI caused by multipath increases. As a result, handling the ISI effects of multipath in the hope of using the diversity of the received signal from different paths has been another major area of research for the past several decades. As a result of this research, a number of signal processing techniques, such as OFDM, adaptive equalization, spread spectrum, and UWB communications, have emerged to take advantage of multipath arrival and to achieve higher data rates, referred to as wideband modems, for wireless RF transmissions. The most popular of these techniques is OFDM, adopted by the IEEE 802.11 community. OFDM combined with MIMO technologies has enabled WLANs to increase their first-generation data rates of 2 Mb/s using spread-spectrum technology to over 100 Mb/s in IEEE 802.11n. In the rest of this section we provide an overview of the most important of these technologies: spread-spectrum transmission, OFDM, and MIMO.

3.4.1 Spread-Spectrum Transmissions

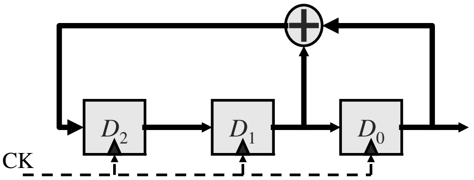
Spread-spectrum technology was first invented during the Second World War and it has dominated military communication applications, where it is attractive because of its resistance to interference and interception, as well as its amenability to high-resolution ranging. In the later part of the 1980s, commercial applications of spread-spectrum technology were investigated, and today it is the transmission technique used in 3G cellular, a number of proprietary cordless telephones, the original IEEE 802.11 WLAN, IEEE 802.15 Bluetooth, ZigBee and UWB WPANs. The voice-oriented digital cellular and PCS industries have selected spread-spectrum technology to support CDMA networks as an alternative to TDMA/FDMA networks in order to increase system capacity, provide a more reliable service, and to provide soft handoff of cellular connections, which will be discussed in more detail in subsequent chapters. In the WLAN industry, spread-spectrum technology was adopted primarily because the first unlicensed frequency bands suitable for high-speed radio communication were ISM bands, which were initially released by the FCC under the condition that the technologies use spread spectrum. In Bluetooth, ZigBee, and UWB WPANs, spread spectrum is adopted because of the simplicity of implementation and low power consumption, which is necessary for ad hoc and sensor networks implementations.

The main difference between spread-spectrum transmission and traditional radio modem technologies is that the transmitted signal in spread-spectrum systems occupies a much larger bandwidth than the traditional radio modems do. Compared with baseband impulse transmission techniques, the occupied bandwidth by spread spectrum is still restricted enough so that the spread-spectrum radio can share the medium with other spread-spectrum and traditional radios in a frequency-division multiplexed format. There are two basic methods for spread-spectrum transmission: DSSS and frequency-hopping spread spectrum (FHSS).

Frequency-Hopping Spread Spectrum. The FHSS technique was first invented to protect guided torpedoes from jamming by the German movie star Hedy Lamarr, who had no technical training – therefore, it must be a relatively simple technology. In order to avoid a jammer, the FHSS transmitter shifts the center frequency of the transmitted signal. The shifts in frequency, or *frequency hops*, occur according to a random pattern that is only known to the transmitter and the receiver. If we move the center frequency randomly among 100 different frequencies, then the required transmission bandwidth is 100 times more than the original transmission bandwidth. We call this new technique a *spread-spectrum technique* because the spectrum is spread over a band that is 100 times larger than with original traditional radio. FHSS can be applied to both analog and digital communications, but it has been applied primarily for digital transmissions.

The FHSS modulation technique can be thought of as a two-stage modulation technique. In the first stage, the input data stream is modulated with a traditional modem and in the second stage the center frequency is changed according to a random hopping pattern generated by a random number generator. Ideally, the random pattern or spreading code is designed so that the occurrence of frequencies is statistically independent of one another. At the receiver, first a de-hopper, synchronized to the transmitter, repeats the hopping pattern of the transmitted signal and then a traditional demodulator detects the received data. In a digital implementation of this system, the sampling rate is the same as the sampling rate of the traditional system, leaving the complexity of the implementation in the same range as traditional modems. As we will see later, DSSS needs much higher sampling rates and, consequently, a more complex hardware implementation.

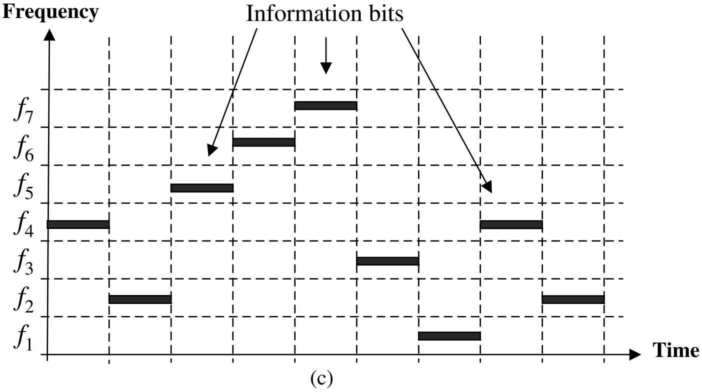
Figure 3.21 shows the hopping pattern and associated frequencies for a frequency-hopping system transferring data packets over the air. A three-state recursive machine is used for



(a)

CK	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
D_2	1	0	1	1	1	0	0	1
D_1	0	1	0	1	1	1	0	0
D_0	0	0	1	0	1	1	1	0
f_c	f_4	f_2	f_5	f_6	f_7	f_3	f_1	f_4

(b)



(c)

FIGURE 3.21 FHSS using a length-7 PN sequence: (*a*) three-stage state machine for implementation of the code; (*b*) states of the memory in time and frequency; (*c*) transmission of information bits in time and frequency in FHSS.

generation of the code that determines the frequency to be hopped to. This machine, shown in Fig. 3.21a, generates a sequence of seven states representing all seven nonzero values which can be formed by three binary digits periodically. Fig. 3.21b shows all seven states of the machine and their associated number, which is used as the frequency index for FHSS. Each packet is transmitted using one of these frequencies. The sequence of frequencies is $f_3, f_5, f_6, f_1, f_4, f_8, f_2, f_7$ before returning to the first frequency f_3. Fig. 3.21c shows transmission of the information bits in time and frequency in this FHSS system. The random sequences used in these situations are referred to as a pseudo noise (PN) sequences; more details on how they are generated are given in Example 4.24 of Section 4.3.4.

In FHSS, the hopping of the carrier frequency does not affect the performance in the presence of additive noise because the noise level in each hop remains the same as the noise level of traditional modems. Therefore, the performance of the FHSS systems in noninterfering environments remains exactly the same as the performance of traditional systems without frequency hopping. In the presence of a narrowband interference the signal-to-interference ratio of a traditional modem operating at the frequency of the interferer becomes very low, corrupting the integrity of the received digital information. The same situation happens in frequency-selective fading channels when the center frequency of a traditional system coincides with a deep frequency-selective fade. In an FHSS system, since the carrier frequency is constantly changing, the interference or frequency-selective fading only corrupts a fraction of the transmitted information and transmission in the rest of the center frequencies remains unaffected. This feature of FHSS is exploited in the design of wireless networks to provide a reliable transmission in the presence of interfering signals or when a system works over a frequency-selective fading channel.

Multipath conditions in wireless channels cause frequency-selective fading, which results in very poor performance in certain frequency regions while performance in other frequencies is acceptable.

When frequency-selective fading occurs, traditional systems operating over center frequencies coinciding with the faded frequencies cannot operate properly. As shown in Fig. 3.22, an FHSS system can be designed so that the deep fades in the environment only

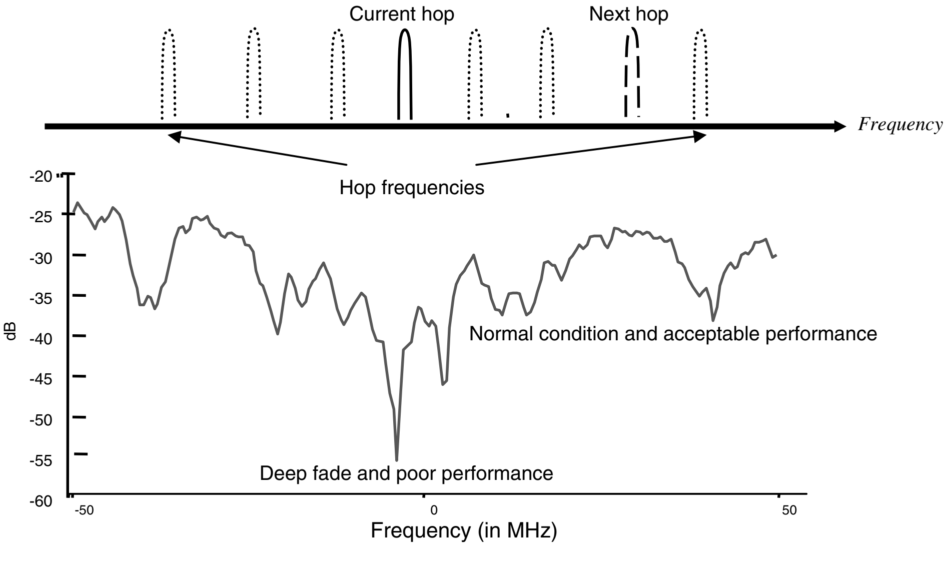


FIGURE 3.22 Frequency-selective fading and FHSS.

corrupt a small fraction of hops, leaving the rest of the hops for successful retransmissions. In an indoor environment, the width of the fade is around several megahertz, and the FHSS system used in IEEE 802.11 or Bluetooth uses hops that are 1 MHz apart from one another. Therefore, if a hop occurs in a deep fade and the data transmitted in that hop is not reliable, then the retransmitted data packet in the next hop will be successful.

FHSS allows the coexistence of several transmissions in the same frequency band using different hoping codes. Different users could be members of the same network following a coordinated hop pattern or two different networks each using their own pattern. For example, IEEE 802.11 FHSS and Bluetooth both use the same bandwidth of 1 MHz and the same 78 channels in the 2.4 GHz ISM bands. Members of each network coordinate their transmissions while both coexist in the same band. Then multiuser interference occurs when two different users transmit on the same hop frequency. If the codes are random and independent from one another, then the "hits" will occur with some calculable probability. If the codes are synchronized and the hopping patterns are selected so that two users never hop to the same frequency at the same time, then multiple-user interference is eliminated.

Direct Sequence Spread Spectrum. In a manner similar to FHSS, DSSS can be thought of as a two-stage transmission technique. In the first stage, each transmitted information bit is coded into N smaller pulses referred to as chips using a PN sequence (described in Section 4.3.4). In the second stage, the chips are transmitted over a traditional digital modulator. At the receiver, the transmitted chips are first demodulated and then passed through a correlator to calculate their autocorrelation function (ACF). We define the ACF of a sequence $\{b_i\}$ as

$$R(k) = \sum_{i=0}^{N-1} b_i b_{i-k} = \begin{cases} N; \ k = mN \\ -1; \ \text{otherwise} \end{cases}$$

The PN sequence is a periodic sequence of length N; therefore, the correlation is calculated over one period of the sequence. The ACF of a good random code has a very high peak of height N at $k = 0$, which is usually referred to as the *processing gain* of the receiver. The value of this ACF for $k \neq 0$ is far below the peak value. Therefore, DSSS systems use the peak of the ACF to detect the transmitted bit.

Example 3.9: DSSS Using Barker Code in IEEE 802.11 A Barker code of length 11, used in the IEEE 802.11 as the spreading signal for the DSSS physical layer, is given by $\{1\ 1\ 1\ -1\ -1\ -1\ 1\ -1\ -1\ 1\ -1\}$.[3] Figure 3.23 shows a data bit "1" in a binary communication DSSS system, the transmitted Barker code for the data bit, and the ACF at the receiver with its high peak and low side lobes.

Barker code has the same autocorrelation properties as other maximum-length PN sequences introduced in Section 4.3.4; however, its length does not need to follow $2^m - 1$ values. Barker code was adopted by IEEE 802.11 for the original DSSS system at 1 and 2Mb/s because the FCC allowed a minimum processing gain of 10 for systems operating in unlicensed ISM bands. The closest maximum-length codes would be of length 15, which has 15/11 times lower bandwidth efficiency than Barker code. The maximum-length PN sequences described in Section 4.3.4 are widely used in CDMA cellular networks and time-of-arrival-based geolocation techniques.

[3] Note that we use the digit "-1" instead of "0". This way, the exclusive-OR operation used in digital addition changes to standard multiplication, which is easier for calculations.

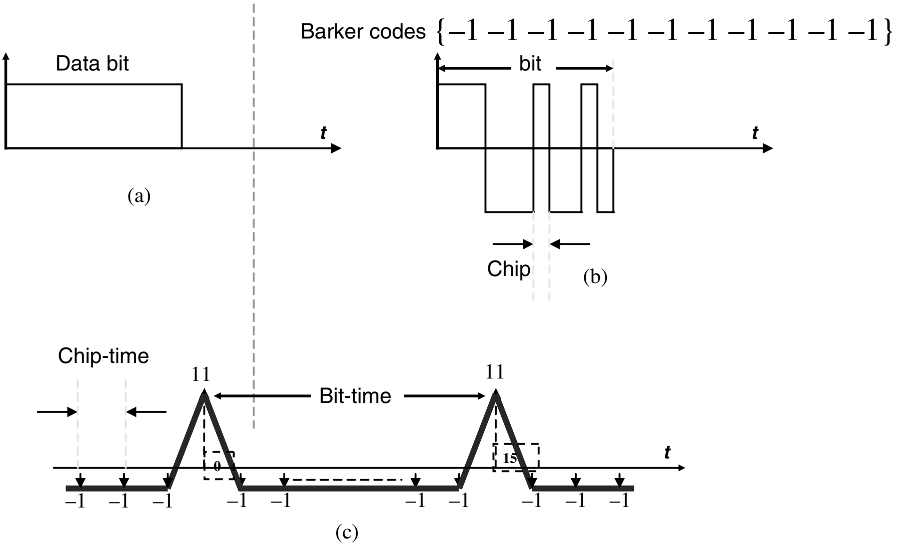


FIGURE 3.23 Barker code used in DSSS signal in IEEE 802.11: (*a*) a transmitted bit; (*b*) the 11-chip Barker code; (*c*) circular ACF of the code.

In cellular networks, DSSS is used for CDMA. In a multiuser direct sequence CDMA environment, different codes are assigned to different users. In other words, each user has their own unique "key" code that is used to spread and despread only their messages. The codes assigned to other users are selected so that, during the despreading process at the receiver, they produce very small signal levels (like noise) that are on the order of the side lobes of the ACF. Consequently, they do not interfere with the detection of the peak of the ACF of the target receiver. In this manner, each user is a source of noise for the detection of other users' signals. As the number of users increases, the multiuser interference increases for all of the users. This phenomenon continues up to a point where the mutual interference among all terminals stops the proper operation for all of them.

In time-of-arrival-based geolocation (see Chapter 12), the peak of the ACF is used to determine the relative TOA of the signal to calculate the time of flight of the signal between the transmitter and the receiver. Since radio waves propagate at the speed of light, the time of flight is used to determine the distance between a transmitter and a receiver. GPS uses DSSS transmission to determine the distance between different satellites and a mobile user, which is further processed for localization of a terminal. The accuracy of the estimate of the flight time is a function of the SNR. In DSSS we can increase the SNR by increasing the processing gain of the codes, which is related to the averaging time for calculation of the ACF. Therefore, more accurate estimates are obtained over longer periods of time. In GPS systems, the initial navigation takes a longer time than the updates because in the updates we use tracking capabilities which improve the accuracy based on the previous location estimates.

The bandwidth of any digital system is inversely proportional to the duration of the transmitted pulse or symbol. Since the transmitted chips are N times narrower than data bits, the bandwidth of the transmitted DSSS signal is N times larger than a traditional system without spreading. As a result, N is also referred to as the *bandwidth expansion* factor. In a manner similar to FHSS, DSSS is also anti-interference and resistant to frequency-selective fading. The transmission bandwidth of the DSSS is always wide, whereas the FHSS is a narrowband system hopping over a number of frequencies in a wide spectrum. As a result, there is some distinction between the two methods. The DSSS systems provide a robust signal with better coverage area than FHSS systems do. An FHSS system can be implemented with much slower sampling rates, saving in implementation costs and power consumption of mobile units.

As shown in Fig. 3.23, for any transmitted bit, the output of the correlator function at the receiver produces a narrow pulse with a height N and a base that is twice the chip duration. Therefore, a DSSS system can be thought of as a pseudo pulse transmission technique receiving a narrow pulse designating each transmitted bit. In a multipath radio channel environment, different paths will bring different pulses to the receiver at different times; an intelligent receiver can use these pulses as a source of *time diversity* to improve its performance. These receivers are referred to as RAKE receivers, which are commonly implemented in DSSS wireless transmission systems to achieve reliable communications.

3.4.2 Orthogonal Frequency-Division Multiplexing

What is known as OFDM today is a method of implementation of multicarrier modulation (MCM) using orthogonality of the adjacent carriers. An MCM system is indeed an FDM system for which a single user uses all the FDM channels together. OFDM is an implementation of MCM that takes advantage of the orthogonality of the channels and develops a computationally efficient implementation based on the fast Fourier transform (FFT) algorithm.

MCM was first evaluated for high-speed voice-band modems in the early 1960s and it was augmented with an FFT implementation for the same application in the early 1980s. It found its way into WLANs (IEEE 802.11), DSL modems, cable modems (IEEE 802.14), WMANs (IEEE 802.16), WPANs (IEEE 802.15) and many other wired and wireless applications in the 1990s. The concept here is very simple. Instead of transmitting a single stream at a rate of R_s symbols per second, we use N streams over N carriers spaced by about R_s/N hertz, each carrying a stream at the rate R_s/N symbols per second. The primary advantage of MCM is its ability to cope with severe channel conditions, such as high attenuations at higher frequencies in long copper lines or the effects of frequency-selective fading in wireless channels. Sub-channels provide a form of frequency diversity, which can be exploited by applying error-control coding *across symbols* in different sub-channels. In the OFDM implementation, this latter technique is referred to as coded OFDM (COFDM). To further improve the performance of an OFDM transmission, one may measure the received signal power in different sub-channels and using a feedback channel may adjust the specification of the transmitted sub-carriers to optimize performance. With these features, OFDM has become an ideal solution for broadband transmissions. In OFDM, increasing the data rate is simply a matter of increasing the number of carriers. The limitations are complexity of implementation and the limitation on the transmitted power.

Fig 3.24 presents an MCM system with N carriers and a channel frequency response with frequency-selective fading. As shown in this figure, carriers operating at different frequencies are exposed to different channel gains. Therefore, the received signals in individual carriers have different SNRs and BER qualities. If some redundancy in the carriers is imposed on the system, then the errors caused by low SNR in poor channels can be recovered. This redundancy can be easily achieved by scrambling the data before transmission. Although the entire bandwidth used by the system is exposed to frequency-selective fading, individual channels are only exposed to flat fading that does not cause ISI,

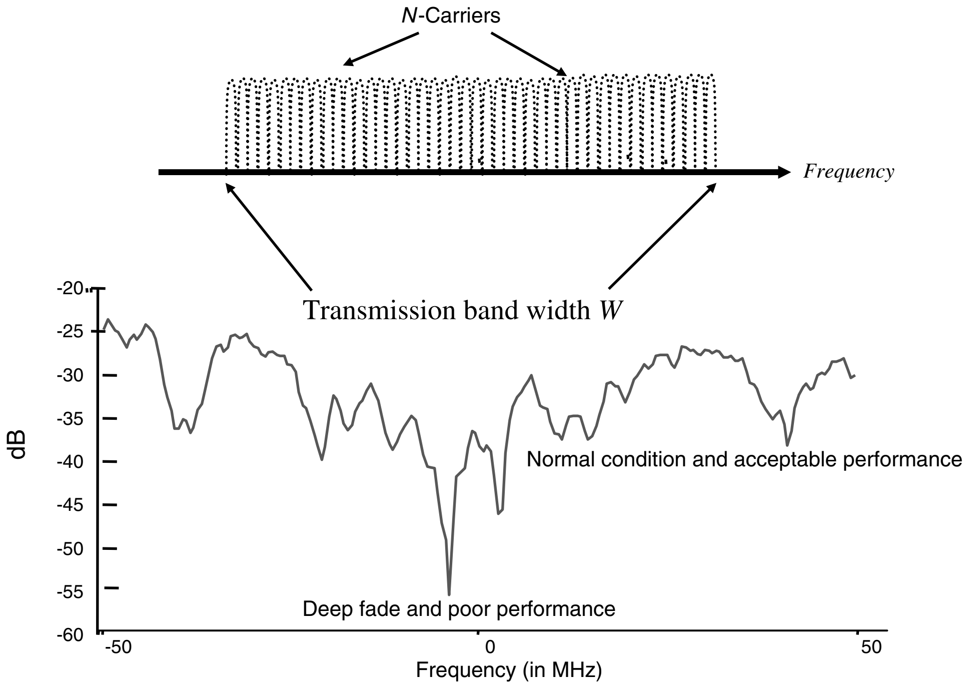


FIGURE 3.24 Frequency-selective fading and MCT.

which needs computationally intensive receivers making use of adaptive equalization techniques. In practice, to avoid overlap between consecutive transmitted symbols, a time guard is enforced between transmissions of two OFDM pulses that will reduce the effective data rate. Also, some of the carriers are dedicated to the synchronization signal and some are reserved for redundancy. An example will bring together all details of implementation.

Example 3.10: Implementation of OFDM in IEEE 802.11a Figure 3.25 shows the 64 sub-channel implementation of OFDM for the IEEE 802.11a,g physical layer specifications. Each channel carries a symbol rate of 250 kS/s. We have 48 sub-carriers devoted to information transmission, four sub-carriers for pilot tones used for synchronization, and 12 reserved for other purposes. The user symbol transmission rate is 48 × 250 kS/s = 12 MS/s. The bit transmission rate depends on the number of bits per symbol (constellation used for transmission) and the rate of the convolutional code used for transmission (see Chapter 4). For BPSK, an $R = 1/2$ coding rate, and 1 bit/S, the data rate is 12 (MS/s) × 1/2 × 1(bit/S) = 6 (Mb/s) and for 64-QAM with 6 bits/S and convolutional coding of rate $R = 3/4$ we have 12 (MS/s) × 3/4 (bits/S) × 6 (bits/S) = 54 (Mb/s). As the distance between the transmitter and the receiver is increased, the data rate is reduced by adjusting the coding rate and the symbol transmission rate (size of the constellation). The fallback data rates are 54 Mb/s to 36, 27, 18, 12, 9, and finally 6 Mb/s to cover distances of up to around 100 m. The guard time between two transmitted symbols is 800 ns compared with the symbol duration of 1/250 kS/s = 4000 ns with a time utilization efficiency of 4000/4800 = 83%. The occupied bandwidth is 20 MHz, providing a channel occupancy of 20 MHz/64 = 312.5 kHz per sub-channel. Therefore, the bandwidth efficiency is 250 kS/s/312.5 kHz = 0.8 S/(s Hz).

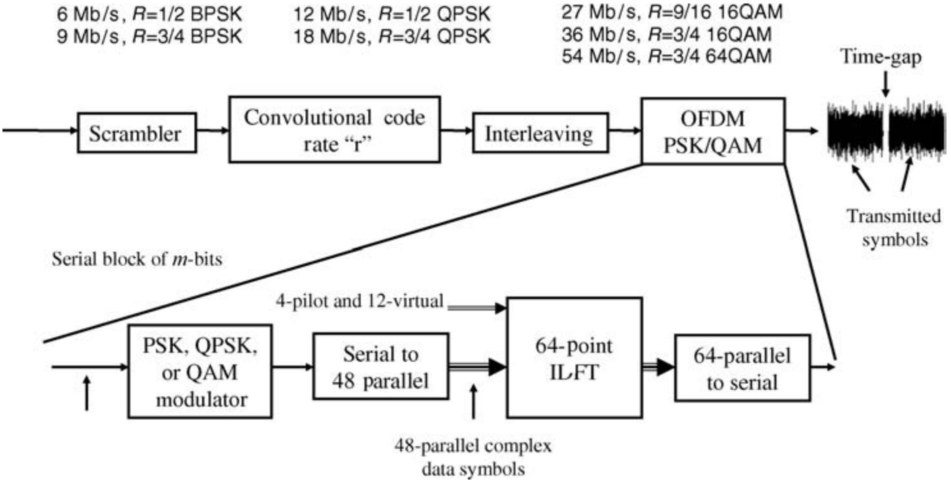


FIGURE 3.25 Frequency-selective fading and OFDM.

3.4.3 Space–Time Coding

STC techniques are used for wireless communication systems with multiple transmit antennas and single or multiple receive antennas. STC techniques are realized by introducing temporal and spatial correlation into the signals transmitted from different antennas. Using STC does not require increasing the total transmitted power or transmission bandwidth. The overall diversity gain of the STC technique results from combining the time diversity obtained from coding with the space diversity obtained from using multiple antennas. In wireless networks, a number of antennas can be deployed in the AP or BS, while the mobile terminal's receiver usually has one main antenna with some possible support from other antennas, depending on the size of the terminal. In traditional multiple AP or BS antenna systems, all transmit antennas carry the same signal and the signal received at each receiver antenna is the summation of all received signals from different transmit antennas. The mobile station combines the diversified signal from different antennas to optimize the performance. The basic principle of STC is to encode the transmitted symbols from different antennas at the BS and modify the receiver to take advantage of the space and time diversity of the arriving signal from multiple transmitter antennas. Using STC at the BS, we can improve the performance of the downlink (base to mobile) channel significantly to support asymmetric applications, such as Internet access, where the downlink data stream operates at a much higher rate than does the uplink data stream.

Using STC, significant increases in throughput over a single-antenna system are possible with only two antennas at the BS and one or two antennas at the mobile terminal. It can be implemented for block [Ala98] or convolutional codes [Tar98, Nag98] with simple receiver structures. To show the basic concept of STC we describe the simple two transmit and one received antenna block coding system known as the Alamouti coded STC system [Ala98] in Appendix B. Also, Alamouti [Ala98] introduced a simple MIMO scheme for two transmit and two received antennas, and the simulation results show that the performance is identical to a system that uses MRC with one transmit and four receiver antennas. Therefore, Alamouti has shown that, with his simple block-coded STC for two transmit and one and two receiver antennas, one can obtain the same diversity performance as is achieved with optimum MRC with two and four received antennas. More elegant approaches that combine transmit diversity with channel coding, similar to TCM, are also available [Tar98, Nag98]. A good overall overview of STC and its applications is provided by Al-Dhahir *et al.* [Dha02].

3.4.4 Capacity Multiple-Input–Multiple-Output Antenna Systems

MIMO antenna systems have recently emerged as one of the promising technologies for next generation wireless networks. In general, MIMO systems combine the transmitting scheme and the detection process in a way that the overall performance of the system is improved. In Section 3.3.2, we demonstrated the general diversity concept (e.g. by using a single transmit antenna and multiple receiver antennas we can take advantage of space diversity) to substantially improve the performance of transmission techniques over wireless fading channels. In the modern literature, this traditional approach to provide space diversity is referred to as single-input multiple-output (SIMO) antenna systems. In Section 3.4.3 we introduced STC as a method for implementation of MIMO systems where we have a number of transmitted antennas at the AP or the BS with one or a few receiver antennas at the mobile terminal. In this section we discuss the bounds on the performance of a general MIMO system to show why in the recent years this area has gained so much of attention for the design of next generation wireless networks.

As we showed in Section 3.2.4, Eq. (3.8), the normalized channel capacity in bits per second per hertz is given by

$$\frac{C}{W} = \log_2\left(1+\frac{E_s}{N_0}\right)$$

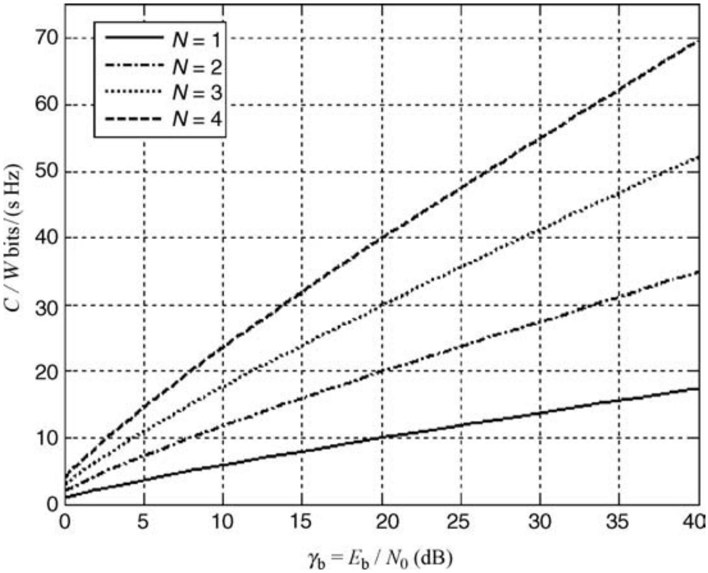


FIGURE 3.26 Comparison of the capacity of the SISO $N \times N$ MIMOs under ideal conditions.

Following the original derivations for the capacity of MIMO provided by Telatar [Tel95] and Foschini and Gans [Fos98b], the capacity of a MIMO channel is given by

$$\frac{C}{W} = \sum_{i=1}^{M} \log_2\left(1 + \frac{E_s}{N_0}\frac{1}{N}\lambda_i\right) = \sum_{i=1}^{M} \log_2\left(1 + \frac{C}{W}\frac{E_b}{N_0}\frac{1}{N}\lambda_i\right) \tag{3.15}$$

where λ_i are the eigenvalues of the $N \times M$ cross-correlation matrix of the *channel gains* between the elements of the transmitter and the receiver antennas. For $M = N = 1$ and $\lambda_i = 1$, this equation is reduced to Eq. (3.8), providing the bounds on the capacity for single-transmitter and single-receiver antenna systems.

To illustrate the bounds on performance improvement using MIMO systems we assume the same number of transmitter and receiver antennas, $N = M$, and no interference among the signals from different receiving antennas so that the received signals are equal and uncorrelated, $\lambda_i = 1$. With these ideal assumptions, Eq. (3.15) reduces to

$$\frac{C}{W} = N\log_2\left(1 + \frac{C}{W}\frac{E_b}{N_0}\frac{1}{N}\right) \tag{3.16}$$

Following the same algebraic manipulations used in Section 3.2.4, we can write this equation as

$$\gamma_b = \frac{E_b}{N_0} = \frac{N}{C/W}\left[2^{(1/N)(C/W)} - 1\right] \tag{3.17}$$

Figure 3.26 shows the capacity in bits per second per hertz versus SNR per bit in decibels for different values of N. As we increase the number of antennas from 1 × 1 single input–single output (SISO) to 2 × 2, 3 × 3, and 4 × 4 MIMO antenna systems we observe a substantial increase in the channel capacity. This substantial increase of the capacity obtained by using MIMO antenna systems has motivated extensive research, resulting in the emergence of new commercial technologies such as IEEE 802.11n.

QUESTIONS

1. In computer networking literature sometimes they use the word bandwidth instead of data rate. Is that correct all the time? Explain why.
2. Discuss the advantages and disadvantages of differential Manchester coding, used in legacy Ethernet, in terms of usefulness to establish synchronization between the transmitter and the receiver and the bandwidth efficiency.
3. The word non return-to-zero used in line coding techniques refers to what feature of the transmitted symbols?
4. Why are line transmission techniques not used in radio communications but they are used in wired and IR communications?
5. How are number of symbols, symbol transmission rate, bandwidth, and data rate related to one another?
6. How do we implement a system to carry the symbols of a 2D signal constellation over a guided or a wireless media?
7. Why is a 5 × 5 PAM constellation not implemented for communication over the wireless medium?
8. Why are bandwidth and power efficiency so important for wireless networking?

9. In order to maintain approximately the same error rates with addition of one bit per symbol to a PAM system, how much increase in transmission power is needed?
10. What is the additional power for transmission of one additional bit per symbol in a QAM system?
11. Use Shannon-Hartley bounds to determine the maximum data rate achievable over a voiceband telephone channel with a bandwidth of 4 KHz and minimum signal to noise ratio of 30 dB.
12. Use Fig. 3.15 to give an estimate for the signal to noise ratio requirement for 16-PSK and 16-QAM modulation and compare them with the Shannon bound for 16-symbol transmission.
13. Figure 3.15 classifies transmission techniques into band limited and power limited regions. What is the meaning of these terminologies?
14. Use Fig. 3.16 to explain the difference between error patterns in a fading wireless and a non-fading guided medium.
15. Explain why in Fig. 3.18 for the same bit error rate we need a much higher signal to noise ratio for the fading channel.
16. Use Fig. 3.19 to explain why diversity techniques are so effective in improving the performance over fading channels.
17. Differentiate between frequency hopping and direct-sequence spread spectrum.
18. What is the difference between a MIMO system and a traditional system using multiple antennas to obtain space diversity?
19. What is the difference between OFDM and COFDM and why does COFDM have a better performance?

PROBLEMS

Problem 1:

We want to transmit a 620×620 pixel image with 3 bytes per pixel using a 56 Kbps voiceband modem connecting Boston to San Francisco.

(a) How long would it take to transmit the picture over the channel?

(b) What is the total delay for communication between two terminals? Assume the speed of propagation in the wired media is 200,000 Km/s and there are four routers between the terminals, each causing an average of 25usec processing delay.

Note: you can use Google map or other direction finding software to find the approximate distance between two cities.

Problem 2:

The standard pulse shape used in most short-distance cable communication applications such as RS232 is a rectangular pulse. If the voltage used for the amplitude of the pulses is $\pm A$ volts, the data rate is R, and the variance of the received noise is N_0, determine the signal-to-noise ratio per bit.

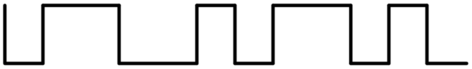
Problem 3:

In the following differential encoded Manchester coded signal,

(a) Show the beginning and the end of each bit.

(b) Identify all the bits in the data sequence.

(c) Identify the bits if it was non-differential Manchester coded.



Problem 4:

For the given sequence of data

Data sequence: | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |

(a) Draw the waveform if the data is NRZ-I encoded.

(b) Draw the waveform if the data is RZ encoded.

(c) Draw the waveform if the data is Manchester encoded.

(d) Draw the waveform if the data is Differential Manchester encoded.

(e) Draw the waveform if the data is AMI coded.

Problem 5:

Show how two binary PSK transmissions can operate simultaneously in the same radio channel by using two carriers at the same frequency in phase quadrature. Draw a block diagram for the transmitters and receivers. Give the overall data rate for this quadrature carrier system as a function of available channel bandwidth.

Problem 6:

Assume that we have a BPSK modem operating on a voiceband channel with a symbol transmission rate of 2400 symbols per second. We want to increase the data rate to 19.2 kbits/sec.

(a) If we increase the number of points in the constellation until the data rate becomes 19.2 kbits/sec while the baud rate remains at 2400 symbols per second, what is the number of points in the constellation?

(b) What is the bandwidth efficiency of the modulation technique?

(c) What is the additional power requirement for the transmitter to keep the error rate the same as before?

Problem 7:

(a) Draw the 5 × 5 PAM signal constellation and show the probability of transmission of each symbol.

(b) How can we implement this 2D constellation on wires?

(c) Calculate the average number of bits per symbol for this constellation.

(d) If we use this constellation over a TP line with bandwidth of 20 MHz what would be the effective bit rate of the modem?

Problem 8:

(a) Derive the relation between the minimum distance and average energy in the constellation for 16-QAM.

(b) Repeat (a) for 5 × 5 PAM constellations.

(c) Compare the two constellations in terms of power requirement and supporting data rates when they both use the same symbol transmission rates.

Problem 9:

The multi-carrier modem in IEEE 802.11g uses 48 carriers for data transmission. If each carrier uses 64-QAM and the transmission rate of each carrier is 250 KSps, what is the overall transmission rate of the modem in symbols/second and bits/second?

Problem 10:

Consider the signal constellations for binary 16-PSK and 16-QAM modulations.

(a) Determine the average energy in each constellation E_s as a function of the minimum distance between the points in the constellation, d.

(b) Starting with Eq. (3.6)

$$P_s \approx \frac{1}{2} erfc\left(\frac{d}{\sqrt{N_0}}\right) \geq \frac{1}{2} e^{-\frac{d}{\sqrt{N_0}}},$$

give an approximate equation for calculation of error rate of each of these modems which relate the probability of the symbol error P_s to the average energy in the constellation E_s and the variance of the background noise, N_0.

(c) For the same expected error rate determine the difference in power requirement for the two modems in dB.

(d) What are the advantages and disadvantages of 16-PSK versus 16-QAM modems?

Problem 11:

Suppose the maximum fade duration over a radio channel is 0.001 ms. Assume that all the bits are in error when a signal encounters a fade. What is the maximum number of consecutive bits in error for a transmission through this channel if the data rate is 10 kbps? If the data rate is 11 Mbps?

Problem 12:

For a 64-QAM modem

(a) Give the SNR at which the error rate over a telephone line is 10^{-5}.

(b) Give the average SNR at which the average error rate over a flat Rayleigh fading radio channel is 10^{-5}.

Note: If you decided to use erfc function to represent the error rate, you can use Matlab erfinv or erfcinv function to calculate the inverse of this function.

Problem 13:

The IEEE 802.11a/g transceivers use multiple modulation techniques in the same transmission bandwidth to provide different data rates. When the mobile terminal is close to the access point, a 64-QAM modulation is used and as mobile moves to the boundary of coverage of the access point BPSK modulation is used that requires substantially lower received signal strength to operate.

(a) If the data rate for the BPSK system is 12 Mbps, what is the data rate of the 64-QAM modem?

(b) What is the difference between the received signal strength requirement of the 64-QAM and BPSK modulation techniques?

(c) If the coverage with 64-QAM is D meters, what is the coverage with BPSK modem when we operate in a large indoor open area with a distance-power gradient of $\alpha = 2$?

(d) Repeat (b) for an indoor office area with a distance-power gradient of $\alpha = 3$.

Problem 14:

(a) Use Eq. (3.15) to calculate the bandwidth efficiency C/W of the 1 × 1, 2 × 2, and 4 × 4 MIMO systems when the signal to noise ration per bit is $E_b/N_0 = 10$ dB.

(b) Repeat (a) for $E_b/N_0 = 6$ dB.

Problem 15:

(a) Use Eq. (3.16) to calculate the signal to noise requirement in dB for the bandwidth efficiency of C/W = 10 and a 1 × 1, 2 × 2, and 4 × 4 MIMO systems.

(b) Repeat (a) for C/W = 20.

PROJECTS

Project 1: Error Rate and Phase Jitter in QPSK Modulation

(a) Sketch a typical QPSK signal constellation and assign 2-bit binary codes to each point in the constellation. Define the decision lines for the received signal constellation so that the receiver can detect received noisy symbols from each other.

(b) We discussed in the chapter that the probability of symbol error for a multi-amplitude, multi-phase modem with coherent detection can be approximated by $P_s = 0.5\, erfc(d/2\sqrt{N_0})$, where d is the minimum distance between the points in the constellation and N_0 is the variance of the additive Gaussian noise. Use this equation to calculate the probability of error of the QPSK modems. Observe that if we consider the signal constellation and the decision lines of part (a), this equation can be modified to $P_s = 0.5\, erfc(\delta\, /8\sqrt{N_0})$, where δ is the minimum distance of a point in the constellation from a decision line.

(c) Use MatLAB or an alternative computation tool to plot the probability of symbol error versus signal to noise ratio in dB. What are the signal to noise ratios (in dB) for the probability of symbol error of 10^{-2} and 10^{-3}? Let's refer to these two SNR's as SNR-2 and SNR-3.

(d) Simulate transmission of the QPSK signal corrupted by additive Gaussian noise for 10,000 transmitted bits. Generate random binary bits and use two bits to select a symbol in the constellation of part (a), add complex additive white Gaussian noise to the symbol so that the signal to noise in dB is SNR-2, and use the decision lines to detect the symbols. Find the number of erroneous symbol decisions and divide it by the total number of symbols to calculate the symbol error rate. Compare the error rate with the expected error rate of 10^{-2}.

(e) Repeat (d) for SNR-3 and error rate of 10^{-3}.

(f) Assume that a channel produces a fixed phase error θ. Give an equation for calculation of the probability of error for a QPSK modem operating over this channel. Use the minimum distance from the decision line, δ, and $0.5\,erfc(\delta/8\sqrt{N_0})$ for the calculation.

(g) Assume that the received signal-to-noise ratio is 10 dB, and sketch the probability of error versus the phase error $0 < \theta < \pi/4$.

(h) Repeat (c), (d) and (e) for a channel with a phase shift of $\theta = \pi/8$.

Project 2: Error Rate and Phase Jitter in 16-QAM Modulation

(a) For a 16-QAM Modem, use MatLAB to plot the probability of symbol error, P_S, versus E_S/N_0, where N_0 is the variance of the noise and E_S, is the average energy per transmitted symbol.

(b) Repeat (a) for 10° phase error at the receiver, and compare the results with those of part (a). What are the signal to noise ratios (in dB) for the probability of symbol error of 10^{-2} and 10^{-3}?

(c) Repeat (a) and (b) for 64-QAM.

4

CODING AND RELIABLE PACKET TRANSMISSION

4.1 INTRODUCTION

In Chapter 3 we showed that bit streams used for information transfer are transmitted over the medium as digital communication symbols, each carrying one or more bits of information. The data stream or the information bits are either generated from analog audio and video sources or are carrying computer data bits. The analog audio or video sources are converted to a set of digital bits using *source coding* techniques. This way, regardless of the source of the information, we have digital bits coming from an application to be transferred by the information network to a destination. Figure 4.1 shows a broad overview of an information network. The information network is a shared transmission facility consisting of switches/

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

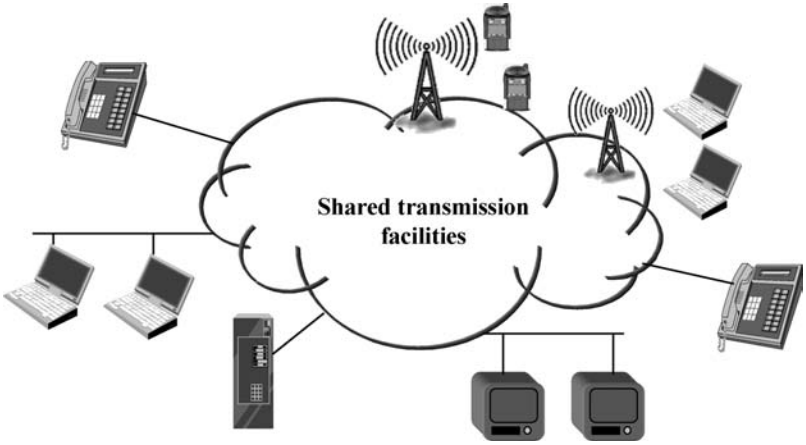
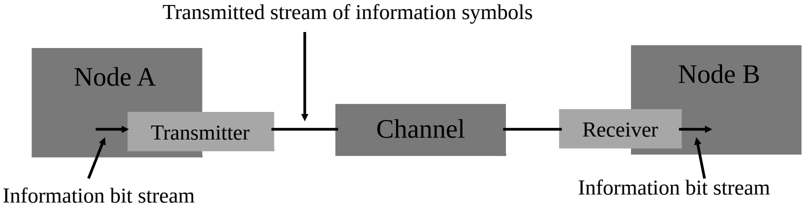


FIGURE 4.1 Broad overview of an information network as a shared medium connecting a huge number of users to one another.

routers and point-to-point trunks connecting a huge number of users through shared or point-to-point access virtual or physical links to provide circuit-switched or packet-switched services. The core of today's transmission in the networks is digital; however, analog POTS and CTV services are still playing a major role in the worldwide networking market. Figure 4.2 illustrates fundamental aspects of point-to-point communications. The bit stream generated by the information source is transmitted through the channel using information symbols, each carrying a few numbers of bits. All modern information networks divide the information bit stream into packets or frames to communicate through the shared medium. Packets or frames are formed to ensure the integrity of the received data. Each packet or frame carries a block of application data plus several overhead bits used for coding and reliable transmission through the point-to-point and the end-to-end connections.

A number of coding techniques and protocols are used to ensure reliable transmission. Error correcting techniques are used to protect the information bits against the disturbances caused by the medium. Coding techniques used for error correcting are referred to as *channel coding* techniques. Channel coding can be done such that the receiver can correct errors in the received stream or it can send feedback to the transmitter if it detects errors. Therefore, we have forward error correction (FEC) coding as well as automatic repeat request (ARQ) schemes.

Modern information networks are based on packet or frame transmission. Each packet carries a block of information bits plus additional bits used for error control and signaling to ensure reliable transmission of the packet from the source terminal to the appropriate



To ensure the integrity of the received data we need to create a frame or packet.

FIGURE 4.2 Elements of point-to-point connections.

destination. To ensure reliable transmission of packets we first need to frame a packet so that at the receiver we can detect the start and the end of the packet. In addition, we may use ARQ schemes that ask for retransmission of corrupted packets. Channel coding techniques are applied to transmitted packets to help the receiver correct or detect erroneous bits in the received packet. If the FEC codes cannot correct the erroneous bits in a packet, then the receiver can use an ARQ scheme.

Spread-spectrum systems and CDMA techniques have been used in a variety of forms in modern information networks. The legacy IEEE 802.11 and IEEE 802.11b, the Bluetooth, ZigBee and UWB IEEE 802.15 WPANs, and the 2G and 3G cellular networks are based on these technologies. Spread-spectrum and CDMA technology basically use PN as well as *M*-ary orthogonal codes extensively. For these reasons it can be said that these techniques are a branch of coding techniques used for reliable transmission over unreliable channels.

Another important problem in packet transmission is congestion control. Congestion occurs when the receiver cannot handle the flow of packets arriving from the transmitter. Congestion control protocols have evolved to handle this situation. Congestion control protocols are either used in the data link layer (DLL) to ensure reliable point-to-point packet transmission over a physical link or at the transport layer to ensure reliable end-to-end connections over the information network.

In this chapter we provide an overview of the most popular coding and reliability techniques applied to modem information networks. Full details of these techniques are beyond the scope of this book. We intend to provide a variety of techniques and introduce each of them with simple examples. We start our technical discussions in Section 4.2 by giving an overview of source coding and framing techniques. In Section 4.3 we present some examples of popular error detection and correction techniques commonly used in modern information networks. Error correcting codes are usually used in real-time applications such as cellular telephones, where the delay constraints on the transmission do not allow retransmission of corrupted packets. Error detection algorithms are very popular, in particular, in data applications to allow the receiver to find a corrupted packet so that it can request retransmission if it is needed. Section 4.4 is devoted to coding techniques used in spread-spectrum and CDMA networks. Section 4.5 describes examples of ARQ methods used in the DLL to provide a mechanism for retransmission of erroneous packets. Examples of congestion control protocols used for reliable packet transmission in point-to-point and end-to-end communications are given in Section 4.6.

4.2 SOURCE CODING AND FRAMING TECHNIQUES

In this section we first introduce an overview of applied source coding and compression techniques used for voice, audio, image, and video compression. These techniques transfer the information source to a bit stream so that, as far as the transmission system is concerned, they look similar to any other data file. In modern information networks, bit streams are divided into blocks to form frames or packets. Later in this section we describe basic techniques used for framing a block of bits in a data stream.

4.2.1 Information Source and Coding

The electrical signals representing information sources are either analog or digital. The amplitude of an analog signal changes continuously in time and it takes a continuum of values. The amplitude of a digital information source changes in discrete time and it accepts only a finite number of discrete values. Therefore, in an analog signal, any small variations of the amplitude in time are meaningful, whereas variations of time and amplitude are quantized into certain values in a digital signal.

The main sources of information which are communicated through a telecommunication network are voice/audio, data, and image/video. A voice signal is generated from the vibrations of the air transferred through a microphone to an analog signal. By data or computer data we refer to digitized information and its associated digitized signal. Analog video signals are generated by scanning still-image frames in sequence to represent motion.

Today, the main long-haul telecommunication networks, the PSTN and the Internet, operate based on digital transmission. Therefore, all analog sources of information are ultimately converted to digital in some place in the network. The analog to digital conversion techniques and their associated data rates vary in different networks. The conversion of analog signals to digital takes place in two steps: sampling and quantization. The minimum sampling rate for an analog signal is defined by Nyquist's theorem as twice the bandwidth of the information message. The quantization could be done by simple mapping of the amplitude to digital values or by using complex algorithms which compress the digital data representing the analog waveform into fewer bits translating into lower data rates. The technique used to encode the analog signal to a digital stream of data is referred to as source coding. To clarify this basic concept we proceed to a couple of simple examples.

Figure 4.3 shows the simple pulse code modulation (PCM) source coding technique. The samples of an analog waveform obtained at the Nyquist rate are quantized into 8 bits of information representing the amplitude of the signal at the sampling time. The quantized sampled signal is then transmitted using a PCM waveform. Analog voice in POTS is transferred through twisted-pair wiring to the end office of the PSTN in analog form. In the end office, this is passed through amplitude compandors to restrict the amplitude fluctuation due to soft and loud voice signal levels and then through analog to digital conversion using PCM encoding. The bandwidth of the voice-band signal transmitted through POTS is roughly $B = 4$ kHz; therefore, the sampling rate is $R_s = 8$ ksamples/s. The PCM encoder encodes the companded amplitude samples of the signal into 8 bits, resulting in a stream of 8(ksamples/s) × 8(bits/sample) = 64 kb/s.

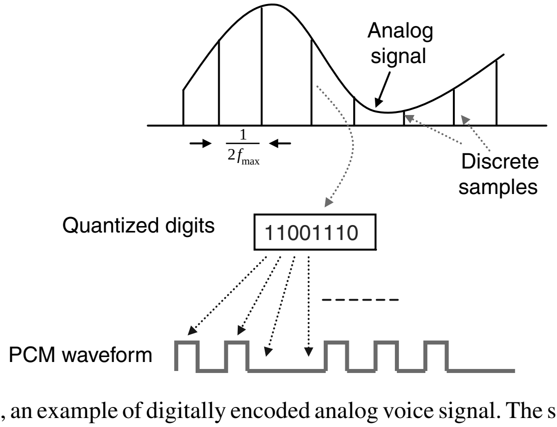


FIGURE 4.3 PCM, an example of digitally encoded analog voice signal. The sampled analog signal is mapped to a quantized digital number to be transmitted using a PCM waveform.

The analog voice in the digital cellular networks is digitized at the terminal and then the digitized data is transmitted through the wireless channel to the BS. Since frequency bands are very precious in wireless communications, the 64 kb/s PCM encoding is not appealing and cellular networks use speech coding techniques which encode the voice signal into around a 10 kb/s speech coded waveform. This way we have more than six times more efficient use of the bandwidth available for cellular networks. Two of the most popular speech encoding techniques used in 3G cellular networks are AMR-WB (used in 3G WCDMA networks, which encodes up to 7 kHz of the voice into data rates ranging from 6.6 to 23.85 kb/s) and VMR-WB (used in 3G CDMA-2000 cellular networks, which encodes the speech signal into variable rates from 1.2 to 9.6 kb/s).

Table 4.1 shows a collection of source coding techniques and their associated data rates which are applied to voice, audio, image, and video. This table provides the reader with quantitative examples for the range of the required transmission rates for transmission of different sources of information.

TABLE 4.1 Sample Source Coding Techniques and their Relative Data Rates

Voice (4 kHz)

G.711 (PCM) → 64 kb/s (inside PSTN)
ADPCM, 32 kb/s (PCS services, inside PSTN)
G.726 and G.727 (ADPCM) → 16, 24, 32, and 40 kb/s
APCO (IMBE) → 4.4 and 7.2 kb/s
IS-95 (QCELP) → 1.2–9.6 kb/s
IS-733 (QCELP-13) → 1–13.3 kb/s
G.728 (LD-CELP) → 16 kb/s
G.723.1, 6.3 kb/s using MPC-MLQ and 5.3 kb/s using ACELP
IS-54, JDC (VSELP) → 8 and 13 kb/s
GSM-EFR (ACELP) → 12.2 kb/s
GSM-AMR (ACELP) → 4.75 and 12.2 kb/s
GSM DCS-1800 (RPE-LTP) → 5.6, 13, and 22.8 kb/s
WCDMA (AMR-WB) → 6.60 to 23.85 kb/s – uses 50–7000 Hz speech
CDMA-2000 (VMR-WB) → 1.2 to 9.6 kb/s

Audio (22 kHz)

CD quality → 1.411 Mb/s
MP3 → mid-range 128 kb/s
ISO/IEC MPEG: MPEG-1 Layer III (known as MP3 → 32–320 kb/s), MPEG-1 Layer II (32–384 kb/s), AAC (64–320 kb/s), HE-AAC (48–128 kb/s)
ITU-T: G.711 (64 kb/s sampled at 8 kHz), G.723.1 (5.3 and 6.3 kb/s), G.726 (16–40 kb/s), G.728 (16 kb/s), G.729a (8 kb/s)

Image

ISO/IEC/ITU-T: JPEG
Others: PNG

Video, variable rates from 3 to 100 Mb/s

ISO/IEC: MPEG (1.5–80 Mb/s)
ITU-T: H.120 (1544 and 2048 kb/s), H.261 (40–2048 kb/s)

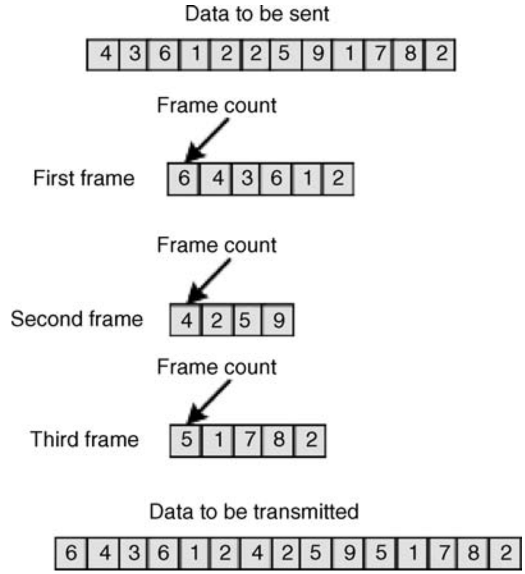


FIGURE 4.4 Character count for framing the data stream.

4.2.2 Framing Techniques

Framing techniques are used to identify the start and the end of a frame and they are also used for variable-length frame transmission. In computer communications, each character is an ASCII code of length 1 byte. The simplest approach for framing the stream of these codes is using a character count in which the first character is used to specify the number of characters in the frame. Figure 4.4 shows an example of this technique. The sequence of data characters is divided into a frame of characters each, for example, representing a word in a text. The character count approach adds an extra character to each word indicating the number of characters in that word. Although very simple to implement, this method results in erroneous detection of the characters when a transmitted bit is not received at the receiver side. This situation is shown in an example illustrated in Fig. 4.5. Owing to an erroneous

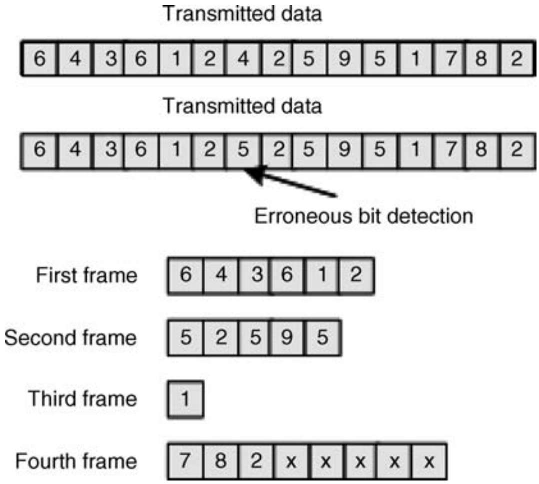


FIGURE 4.5 Erroneous detection using character count.

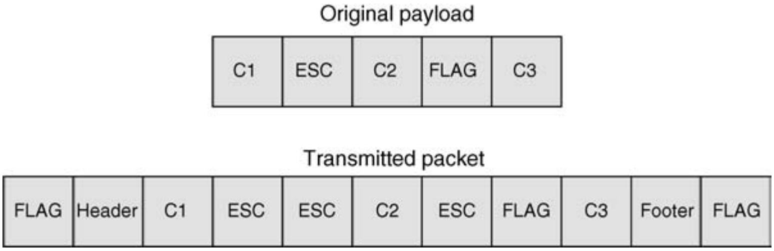


FIGURE 4.6 Example of byte stuffing to form a frame.

transmission, the length “4” of one of the frames is switched to “5,” resulting in erroneous detection of all the following frames. Character stuffing techniques handle the situation better than character counting.

Character-oriented Stuffing. In the character-oriented stuffing or byte-stuffing method the payload framing is implemented by appending a flag byte to the start and the end of each frame. The header and trailer flag bytes can be the same or different. In the case of occurrence of the flag byte in the original payload, the link layer protocol should be instructed to insert a special character, usually an escape byte or ESC character, immediately before the flag byte in the original payload. If the payload also includes an ESC character, then the same process is applied. Figure 4.6 provides an example of byte stuffing for framing. The original frame contains both the header and the ESC as part of the original payload. The data link protocol adds an ESC in front of each of them. At the receiver the data link protocol reads the header as an indicator of the start of the packet. Then it reads the payload byte by byte. After reading the first ESC it reads the next ESC character and it realizes that the first one was for protection of this ESC, so it removes the first ESC. When the protocol arrives at the second ESC and reads a header after that, it realizes that it was not a header and eliminates the ESC and keeps the header as a character in the payload. When the protocol reads the final header it eliminates it and recognizes that the payload is completed. If there is a sequence of ESC or headers in the payload, then one ESC is added to each ESC or each header. This way, each time that the protocol reads the received data and observes an ESC character it eliminates the ESC and keeps the next byte as payload data anyway. This approach is used in character-based transmission using point-to-point protocol (PPP), which is common in data terminal interfaces to the computer devices and dial-up Internet connections. Figure 4.7 shows an example of a PPP packet format using character stuffing. In addition to flags and payload, we have other fields defining address, control operations, version of the protocol, and checksum bits used for the integrity of the transmitted bits. As the knowledge of networks evolved and newer generations of networks were introduced, the byte-stuffing framing technique became less popular and more efficient bit-stuffing techniques replaced them.

Flag	Address	Control	Protocol	Variable length	Checksum	Flag
01111110	1-byte	1-byte	1 or 2-bytes	payload	2 or 4-bytes	01111110

FIGURE 4.7 Frame format of the PPP with start- and end-flag bytes.

Original bit stream	0101110111100111111111111111001001
Stuffed bit stream to be transmitted	0111111001011101111100111110111110111110010010111111101111110

FIGURE 4.8 Example of bit-stuffing technique for framing.

Bit Stuffing. In order to compensate the byte dependence of the byte-stuffing framing technique and to generalize the idea to arbitrary length of character, the bit-stuffing technique was introduced. Bit stuffing allowed the frames to have arbitrary numbers of bits and they are not restricted to the byte unit. In this approach, in a manner similar to byte stuffing, each frame starts and ends with a special byte, 01111110. However, the DLL protocol using bit stuffing inserts a 0 whenever five consecutive 1s are encountered in the payload. This way, the special byte 01111110 cannot occur in the framed payload and the receiver can detect the start and end of each frame and decode it correctly. Figure 4.8 provides an example of frame transmission using this approach to clarify the situation. In the case of occurrence of the flag byte in the original payload, the protocol inserts a 0 and sends 011111010. At the receiver side, after detection of 011111010, the protocol deletes the extra 0 and passes 01111110 to the higher layers, keeping link layer protocol operation transparent to the higher layers. The high-level data link control (HDLC) protocol is an example of a data link protocol using the bit-stuffing technique. Figure 4.9 shows a typical frame format for the HDLC protocol which is basically very similar to that of the PPP shown in Fig. 4.7. The difference is in the program written for the transmitter and the receiver to implement the protocol using bit stuffing rather than byte stuffing.

4.3 FEC CODING

Error control coding techniques are channel coding techniques used in framing the data packets that can be used to either correct the erroneous bits or detect their occurrence in a data frame. In this section, we first address the fundamentals of error control coding by defining the idea of *distance* and relating it to the bandwidth and ability to detect or correct errors. Then we provide a survey of coding techniques used in the design of popular information networks. We give three examples of block codes, an example of convolution codes, and a general description of TCM, and block interleaving codes.

4.3.1 Fundamentals of Coding

Block coding, as the name suggests, involves encoding a *block* of bits into another block of bits, with some redundancy to combat errors. Block coding, in its simplest form, consists of a *parity check* bit. An extra bit is added to each block of k bits and the extra bit is selected so that each new block of $k+\ 1$ bits has either an even number of 1s or an odd number of 1s. The

Flag	Address	Control	Variable length	Checksum	Flag
01111110	1-byte	1-byte	payload	2-bytes	01111110

FIGURE 4.9 Frame format of the HDLC protocol with star- and end-flag bytes.

extra bit is called the parity check bit. The result is that, if the channel introduces a single bit error in a block of $k+1$ bits, the number of 1s in the block will no longer be even (or odd) and the receiver can *detect* the error. The simplicity of this code is clear, because if there is an even number of errors in the block the errors cannot be detected, since the number of 1s is maintained even or odd.

Block codes use finite-field arithmetic (modern algebraic techniques) properties to encode and decode blocks of bits or symbols. Most operations are based on linear feedback shift registers (LFSRs), which are easy and inexpensive to implement. Most block codes are created in *systematic* form where the k data bits are retained as is and the $n-k$ parity check bits are either prepended or appended to them. The parity check bits are generated via a generator matrix or generator polynomial. The encoded block of n bits is called the *codeword*, and this is transmitted over the channel. Codes generated by a polynomial are called cyclic codes, and block codes of this nature are called *cyclic redundancy check* (CRC) codes and are employed in a variety of data transmission schemes for error correction and detection.

Coding and Data Rate. Using a variety of algebraic techniques, efficient encoding rules have been obtained that calculate a set of $n-k$ parity check bits that apply parity checks to a group of bits in the block of k bits. Together with these parity check bits, the size of the encoded block is $k+(n-k)=n$(bits). The block code is called an (n,k) block code and the code rate is $R_c = k/n$. This means that if the raw data rate is R_b (bits/s), then only kR/n bits/s correspond to actual data. The factor k/n is called the code rate. The rest of the bits do not contain useful information and are included only for error control purposes.

The received word may be identical to the codeword in the case of error-free transmission and may have been modified due to channel errors. The modifications may result in another valid codeword, in which case it is not possible either to detect or to correct the errors. The probability of such a false detection is upper bounded [Wol82] by

$$P_{FD} \leq 2^{-(n-k)} \tag{4.1}$$

Example 4.1: Block Coding and Effective Data Rate In a wireless cellular network operating at a transmission rate of 270 kb/s a block of 184 bits is encoded into 224 bits that form a codeword. In this system the number of parity check bits for this block encoder is 40, the code rate is $R = 184/224 = 0.82$, the effective information transmission rate is $270 \times 0.82 = 221.4$ kb/s.

The block coding schemes, in general, break the transmitting signal into blocks of bits and then each block is mapped to a different bit pattern. The key in block codes is to define the output patterns in a way that, at the receiver side, the blocks can be decoded and errors can be detected, if they have occurred. Coding of each block is independent from coding of the previous blocks; hence, block codes are memoryless. The following example discusses a simple block coding scheme and its impact on the data rate of the system.

Example 4.2: A Simple Block Code and its Relation to Data Rate of the System Table 4.2 provides a coding scheme for 2-bit blocks by mapping each block to a 4-bit block.

Since the coding scheme works on 2-bit blocks, the mapping consists of $2^2 = 4$ rows to cover all the pattern possibilities. The codeword patterns, however, can be chosen

TABLE 4.2 A Simple Coding Scheme with Two Parities

Message pattern	Codeword pattern
00	0011
01	0101
11	1010
10	1100

from $2^4 = 16$, which gives

$$\binom{16}{4} = 1820$$

different codes. The codewords for a good block code are always chosen to have the maximum difference between 0s and 1s in them, which maximizes the effectiveness of the code to detect errors if they occur.

Assuming the communication system is using binary phase-shift keying (BPSK), i.e. each transmitted bit represents a symbol, then, with a data rate of 4800 symbols/s, before applying the coding scheme, 3-bit blocks are transmitted sequentially, which gives the same 4800 bits/s. However, after applying the coding scheme, the codewords are transmitted at 4800 symbols/s. But within the 4800 transmitted symbols, in each second, only 2400 symbols contain data information and the rest are redundant bits added for error detection. It can be concluded that the effective data rate of the system has then dropped to 2400 bits/s. This example shows the trade-off between the data rate and ability to detect erroneous bit patterns (extra security) that can be applied to raw data. Consequently, the drop in effective data rate of the system changes the bandwidth requirement of the system.

Hamming Distance and Performance. As discussed in Example 4.2, a block code maps a block of input stream into a block of output stream referred as a codeword. A good set of codewords are those codewords with maximum differences in their bits. For example, if instead of 1 the transmitter sends 111 and 0 is represented by 000, then, at the receiver side, a received set of bits corresponding to 110 is very likely to be in reality 111 and the receiver can make this judgment. The idea behind the design of block codes is thus to have a large *distance* between any pair of codewords. This distance is measured in terms of the number of positions (bits or symbols) in which the codewords differ and is called the Hamming distance. The *minimum Hamming distance* between the set of all codewords of a block code determines its error detection or correction capability. A block code with a minimum Hamming distance of $d_{\min}$ can detect blocks of errors that have a "weight" of less than $d_{\min}$ and can correct blocks of errors that have a weight up to $t_{\max}$, where

$$t_{\max} = \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \tag{4.2}$$

Here, $\lfloor x \rfloor$ refers to the largest integer less than or equal to x. An error block is also represented in a manner similar to a block of data bits. Those bits that are not changed are represented by 0s and those that are changed are represented by 1s. The *weight* of the error block corresponds to the number of 1s in the block, or the number of bits that are changed and thus in error. Intuitively, we can see why a block code with a minimum distance of $d_{\min}$ can correct up to $t_{\max}$ errors. Given two codewords in the set, the distance between them is greater than or equal to $d_{\min}$. An error block modifies a codeword into the received word. If its distance from the correct codeword is less than half the distance between the correct codeword and any other codeword, then we can associate the received word as being closest to the original codeword and correct it accordingly. If, however, the error block modifies the received word to make it closer to some other codeword, then the error correction procedure will not work.

For sensitive data transfer, it is still possible to *detect* errors with weights larger than $t_{\max}$. The detection of errors is performed by determining whether or not the received word is a valid codeword. This can be done by computing the parity check bits again from the k data bits and comparing them with the parity check bits received over the channel. It is possible that the data bits were received correctly but that the parity check bits were in error, but the two cases are indistinguishable.

Example 4.3: Error Detection in (4, 2) Coding and its Hamming Distance In the previous example, the Hamming distance of the proposed coding scheme is 2 because each codeword is different from other codewords at least in two bits. Therefore, this coding scheme has the ability to detect one bit error, but it cannot correct the error. Assume the first codeword, 0011, has been transmitted and due to channel errors it is received as 0010. Since 0010 is not a codeword in the table, the receiver can detect the occurrence of the error. But if the receiver compares this codeword with all codewords in the table, then it finds 0011 and 1010 each with 1 bit distance and it has no way to correct the error. Any single-bit error within a codeword is detectable using this code. If a 0000 was received with two bits of errors, then the receiver still can detect the error but it cannot correct it because it cannot decide between 0011 and 1100. If the two bits that are in error had occurred at the second and third bits then we would have received a 0101, which is a codeword in the table, and we would have had no means to correct or detect the error. Therefore, this code cannot detect occurrence of all 2-bit errors.

If we know the transmission error rate, then we can calculate the probability of occurrence of different erroneous codewords. Assume the transmission takes place with a bit transmission error rate of $p = 10^{-2}$. Since the received codewords contains 4 bits and the coding scheme can detect a single bit error in a codeword, the probability of occurrence of a single bit error in a 4-bit codeword is

$$\binom{4}{1} p(1-p)^3 = 0.038$$

However, if 2 bits (or more) are in error in a received codeword, then the receiver might not be able to detect the error correctly. The probability of occurrence of such an event is

$$P = 1 - \binom{4}{0}(1-p)^4 - \binom{4}{1} p(1-p)^3 = 5.92 \times 10^{-4}$$

which is much more unlikely than the occurrence of single bit errors.

It can be shown that, to detect n bit errors, the encoder has to have a Hamming distance of $n + 1$ between the codewords. Similar steps can be taken to design a coding scheme which is able to correct the erroneous received bit, which we will discuss later in this chapter. It can also be shown that, to design a block code that can correct up to n bit errors, a Hamming distance of $2n + 1$ is desirable. The other important parameter in block codes is the

Hamming weight of the code, which is simply the maximum number of 1s representing the distance of the code from all-zero code of the same length.

Example 4.4: Error Correction in (4, 2) Code All received bit patterns with single errors can be detected and corrected, as shown in Example 4.3. However, 2-bit errors in a received codeword might be detected but the correction cannot be performed to identify the erroneous bit.

In the case of occurrence of 3-bit errors in a received signal, the received codeword can easily be interpreted as another codeword. For example, if the transmitted codeword is 0011 and the corresponding codeword is 1101, then the receiver interprets the received codeword as single bit error detection of 1100 and declares 1100. Consequently, the transmitted symbol is interpreted as 10 instead of 00. However, the probability of occurrence of such case, assuming $p = 10^{-2}$, is

$$\binom{4}{3} p^3 (1-p) = 3.96 \times 10^{-6}$$

Similar erroneous detection of the received signal takes place when the codeword is corrupted completely and all 4 bits are received in error. Detection of 1100 clearly indicates that the transmission codeword is 10, while the transmitted codeword could be 0011 with all 4 bits in error. The probability of occurrence of all 4 bits in error is

$$\binom{4}{4} p^4 = 10^{-8}$$

4.3.2 Block Codes

In this section we review simple and practical block codes used in communication systems to detect or correct erroneous received bit patterns. The simplest and yet effective mechanism for block coding is inserting the parity bit at the end of each block. Appending or prepending the parity bit to the input block increases the Hamming distance of the block of data by one and instantly allows the receiver to detect one bit error at receiver side.

Example 4.5: One-Bit Parity Code and Detection of Single Errors In a coding scheme, the encoder adds a parity bit to 7-bit blocks of a bit stream such that total number of 1s in the 8-bit block is even. For example, the encoder transforms 00110001 into 01100011. The decoder can compare the parity bit to check the number of 1s in an 8-bit block to verify its valid detection. However, in the case of a single bit error the decoder can detect the erroneous detection of the block and ask for retransmission. The probability of correct detection of a block, assuming $p = 10^{-2}$, is

$$P_0 + P_1 = \binom{8}{0} (1-p)^8 + \binom{8}{1} p (1-p)^7 = 0.9973$$

In a system without the parity bit, in order to receive the entire block correctly all the bits have to be receiver unaltered, which occurs with the probability of

$$P_0 = \binom{7}{0} (1-p)^7 = 0.9321$$

The coded bit stream shows 7% improvement in reliability.

There are a number of more effective block codes used in information networks; here we present examples of Hamming, CRC, and non-binary codes and their applications.

Hamming Codes. Hamming codes were invented by Richard Hamming in the late 1940s to correct errors caused by optical readers used for data entry to computers with punch cards. Today, these codes are still used in computer, communication, and data compression applications. There are different ways to represent a coding technique. Hamming codes are defined as linear block codes with error detection and correction capabilities. Common sizes of Hamming codes are (7, 4), (15, 11), and (31, 26), with generalization to ($2^n - 1$, $2^n - n - 1$) codes. A Hamming code can be described by a systematic approach that codes each input block to a codeword using the properties of matrices. Decoding the received codeword is similarly done by taking the proper steps to define appropriate matrices.

The systematic approach is based on three matrices, the generator matrix **G**, the parity check matrix **H**, and the parity matrix **P**. The generator matrix is used at the transmitter side to map the block of data into the coded word. The transpose of the parity check matrix is used at the receiver to map the received code word to the so-called syndrome matrix **S**. The syndrome is then used to identify the erroneous bit and correct it. The parity matrix **P** is an auxiliary matrix used for the formation of the **G** and **H** matrices. Table 4.3 summarizes the matrix operations used for systematic description of the Hamming codes. There are simple methods to find the generator and parity check matrices of the Hamming codes. Other *linear* block codes can also be represented using similar matrices. The approach is best illustrated with an example.

Example 4.6: A (7, 4) Hamming Code To design a (7, 4) Hamming code, we first start by defining the generator matrix **G** by appending the parity matrix to an identity matrix of size k. The parity matrix is simply produced by using all combinations with 3-bits with two or more 1s in the case of the Hamming code:

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad \mathbf{I}_{4\times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \mathbf{G} = [\mathbf{I}_{4\times 4} \quad \mathbf{P}] = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

TABLE 4.3 Summary of Systematic Representation of Hamming Codes

Systematic notations for an (n, k) code	
Code rate	k/n
Identity matrix	$\mathbf{I}$
Parity matrix	$\mathbf{P}$
Information message matrix	$\mathbf{M} = [m_1 \quad m_2 \quad \ldots \quad m_k]$
Coded word matrix	$\mathbf{C} = \mathbf{M} \times \mathbf{G} = [m_1 \quad m_2 \quad \ldots \quad m_k \,\vert\, p_1 \quad p_2 \quad \ldots \quad p_{n-k}]$
Code generation matrix	$\mathbf{G} = [\mathbf{I} \quad \mathbf{P}] : \quad n \times k$
Received code word	$\mathbf{R} = \mathbf{C} + \mathbf{e}$
Syndrome matrix	$\mathbf{S} = \mathbf{R} \times \mathbf{H} = \mathbf{C} \times \mathbf{H} + \mathbf{e} \times \mathbf{H} = \mathbf{e} \times \mathbf{H}$
Parity check matrix	$\mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I} \end{bmatrix} : \quad (n-k) \times n$
(**P** matrix for Hamming codes is generated using nonzero combinations)	

The second step is to construct the parity check matrix **H** by appending the **I** matrix of size $(n-\mathrm{k})$ to the parity matrix:

$$\mathbf{P}=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\\0&1&1\end{bmatrix},\quad \mathbf{I}_{3\times 3}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\Rightarrow \mathbf{H}=\begin{bmatrix}\mathbf{P}\\ \mathbf{I}_{3\times 3}\end{bmatrix}=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\\0&1&1\\1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

For each block of 4 bits of information message, e.g. $\mathbf{M}=[0\quad 0\quad 0\quad 1]$, the codeword C can be obtained by multiplying the block by the **G** matrix:

$$\mathbf{C}=\mathbf{M}\times\mathbf{G}=[0\quad 0\quad 0\quad 1]\times\begin{bmatrix}1&0&0&0&1&1&1\\0&1&0&0&1&1&0\\0&0&1&0&1&0&1\\0&0&0&1&0&1&1\end{bmatrix}=[0\quad 0\quad 0\quad 1\quad 0\quad 1\quad 1]$$

In a noiseless environment, the codeword C is received unaltered at the receiver side. The system then multiplies the $R=C$ by the parity check matrix:

$$\mathbf{S}=\mathbf{R}\times\mathbf{H}=[0\quad 0\quad 0\quad 1\quad 0\quad 1\quad 1]\times\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\\0&1&1\\1&0&0\\0&1&0\\0&0&1\end{bmatrix}=[0\quad 0\quad 0]$$

The product is referred to as the syndrome of the codeword, and for unaltered received bits the syndrome is $[0\quad 0\quad 0]$. In the case of erroneous detection of one bit, e.g. the third bit:

$$\mathbf{R}=\mathbf{C}+\mathbf{e}\Rightarrow\mathbf{R}=[0\quad 0\quad 0\quad 1\quad 0\quad 1\quad 1]$$
$$+[0\quad 0\quad 1\quad 0\quad 0\quad 0\quad 0]=[0\quad 0\quad 1\quad 1\quad 0\quad 1\quad 1]$$

$$\mathbf{S}=\mathbf{R}\times\mathbf{H}=[0\quad 0\quad 1\quad 1\quad 0\quad 1\quad 1]\times\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\\0&1&1\\1&0&0\\0&1&0\\0&0&1\end{bmatrix}=[1\quad 0\quad 1]$$

The syndrome pattern of $[1 \quad 0 \quad 1]$ is observed as the third row of the parity check matrix and, hence, indicates that indeed the third bit has been received in error. The correct codeword can then be identified as $[0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1]$ and the 4-bit block can be reconstructed from the first four bits as $\mathbf{M} = [0 \quad 0 \quad 0 \quad 1]$.

Example 4.7: Probability of Erroneous Detection of the Codeword Using Hamming Code It has been proven that (7, 4) Hamming code is capable of detecting and correcting single-bit error patterns in the received codewords. The (7, 4) Hamming code is unable to detect more than single-bit errors. The probability of erroneous detection of a codeword can then be calculated as

$$\begin{aligned} P_{\mathrm{e}} &= P_2 + P_3 + P_4 + P_5 + P_6 + P_7 + P_8 = 1 - P_0 - P_1 \\ &= 1 - \binom{8}{0}(1-p)^8 - \binom{8}{1}p(1-p)^7 = 0.0027 \end{aligned}$$

Furthermore, you can use the (7, 4) Hamming code for detection of 3-bit errors in a codeword. In this case erroneous detection of the codeword is

$$\begin{aligned} P_{\mathrm{e}} &= P_4 + P_5 + P_6 + P_7 + P_8 \\ &= 1 - P_0 - P_1 - P_2 - P_3 = 1 - \sum_{i=0}^{3} \binom{8}{i} p^i (1-p)^{8-i} = 6.7788 \times 10^{-7} \end{aligned}$$

Cyclic Redundancy Check Codes. CRC codes were introduced by Peterson and Brown in 1961 [Pet61]. These codes are commonly used to create the checksum for variable-length and long blocks of data. These codes can be implemented with small overhead and simple hardware. The basic idea of a CRC code is that it interprets the long block of data as a binary number, divides it by a prime number of length $n + 1$ bits, and uses the n-bit remainder as the checksum for the coding. At the receiver, the main data is divided by the same prime number and the remainder is checked to detect the occurrence of an error in transmission of the block. As an example, the IEEE 802.3 Ethernet uses CRC codes of length 33 bits with a 32-bit remainder, which is shown in Fig. 4.10. The variable-length data of up to 1500 bytes and an overhead of up to 22 bytes are divided by a 33-bit prime number to form a CRC-32 remainder as the checksum.

CRC codes are also referred to as polynomial codes because they can be expressed using polynomial expressions. CRC codes use a set of polynomials whose coefficients are either "0" or "1" using the so-called finite field arithmetic. A summary of polynomial expressions describing CRC codes is given in Table 4.4. In this approach we define an rth-generator polynomial $G(x)$ representing the prime number used for division. The message or information polynomial of any length is then multiplied by x^r and divided by $G(x)$ using long division. The remainder of such division then defines the CRC checksum bits corresponding to the message. Again, an example helps clarify the situation.

Preamble	Start delimiter	DA	SA	Length of data	Data	Pad	Checksum
(7)	(1)	(2 or 6)	(2 or 6)	(2)	(0-1500)	(0-46)	(4)

FIGURE 4.10 Frame format of the IEEE 802.3 Ethernet variable-length packets with a 4-byte (32-bit) checksum using CRC codes.

TABLE 4.4 Summary of Polynomial Expressions Used for Describing CRC Codes

Generator polynomial	$G(x)$ order r; example: $1001 \Rightarrow G(x) = x^3 + 1$, $r = 3$
Information polynomial	$I(x)$ order $> r$
Parity polynomial	$p(x) = x^r I(x)/G(x)$
Coded polynomial	$C(x) = x^r I(x) + p(x)$
Received coded polynomial	$R(x) = C(x) + e(x)$
Parity check division	$R(x)/G(x) = C(x)/G(x) + e(x)/G(x) = e(x)/G(x)$

If $e(x)$ is not a multiple of $G(x)$ then we can detect the error

Example 4.8: CRC Code Using $G(x) = x^3 + x + 1$ In this example, we describe CRC codes using the prime polynomial $G(x) = x^3 + x + 1$ associated with the prime number [1011]. The checksum is the remainder after dividing the original block of bits by this prime number. Assume that our information bit stream is

$$I = [1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1]$$

associated with the information polynomial represented by

$$I(x) = x^6 + x^3 + x^2 + 1$$

The parity polynomial is then generated by

$$P(x) = x^3 \times I(x)/G(x) = \frac{x^9 + x^6 + x^5 + x^3}{x^3 + x + 1} = x^2 + 1$$

This division can be found using simple binary division of [1001101] by the prime number [1011] as shown below:

```
        1010011
1011)1001101000
     1011
      0101
      0000
       1010
       1011
        0011
        0000
         0110
         0000
          1100
          1011
           1110
           1011
            101
```

The resulting remainder checksum is [101] with polynomial expression $P(x) = x^2 + 1$. The coded transmitted word is then defined as

$$C(x) = x^3 \times I(x) + P(x) \Rightarrow C = [1001101101]$$

At the receiver side, the noisy version of the transmitted codeword is received as $R(x) = C(x) + e(x)$. The received codeword is then checked for errors by

$$\frac{R(x)}{G(x)} = \frac{C(x)}{G(x)} + \frac{e(x)}{G(x)}$$

Apparently, the received codeword is acceptable if $e(x)$ is a multiple of $G(x)$.

In practice, generator polynomials can be more complex; for example, the generator polynomial for IEEE 802.3 Ethernet is given by

$$G(x) = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1$$

Nonbinary Codes. In multilevel modulation schemes, it is possible to encode *symbols* (nonbinary alphabets) in a similar manner to simple block codes. Bose–Chaudhuri–Hocquenghem (BCH) codes are a popular class of nonbinary codes. Reed–Solomon (RS) codes, a subset of the BCH codes, are another good example of nonbinary block codes employed in a variety of wireless networks. The symbols are commonly drawn from a set of 2^m alphabets (each alphabet represents m bits). An RS codeword has length $n = 2^m - 1$. The number of symbols being encoded can vary from 1 to $n - 1$. Depending on the number of symbols encoded, the minimum distance between the codes, which in the case of RS codes is $n - k + 1$, changes.

Example 4.9: An RS Block Encoder In a wireless data network, (63, 47) RS encoders form a set of 64 symbols, each representing 6 bits. This means that a block of 47 symbols is encoded into a block of 63 symbols for a symbol code rate of 47/63 = 0.746. The minimum distance between the codes is $d_{\min} = 63 - 47 + 1 = 17$. It can thus correct up to eight symbol errors.

4.3.3 Convolutional Codes

Convolutional codes are used in variety of LANs and WANs, such as IEEE 802.11g or GSM, when a strong channel coding technique is needed. Convolutional codes, unlike block codes, do not map individual blocks of bits into blocks of codewords. Instead, they accept a continuous stream of bits and map them into an output stream, introducing redundancies in the process. Usually, a code rate can be defined for convolutional codes as well. If there are k bits per second input to the convolutional encoder and the output is n bits per second, then the code rate is once again k/n. The redundancy, however, is dependent not only on the incoming k bits, but also several of the preceding k bits. The number of preceding k bits used in the encoding process is called the constraint length m and is similar to the *memory* in the system. Simple examples can further clarify the situation.

Example 4.10: Data Rate Calculation for a Convolutional Encoder A WLAN uses a rate 9/16 convolutional encoder on an incoming data stream of rate 27 Mb/s. The transmission facilities needs to provide for a 27/(9/16) = 48 Mbps to carry the encoded data stream.

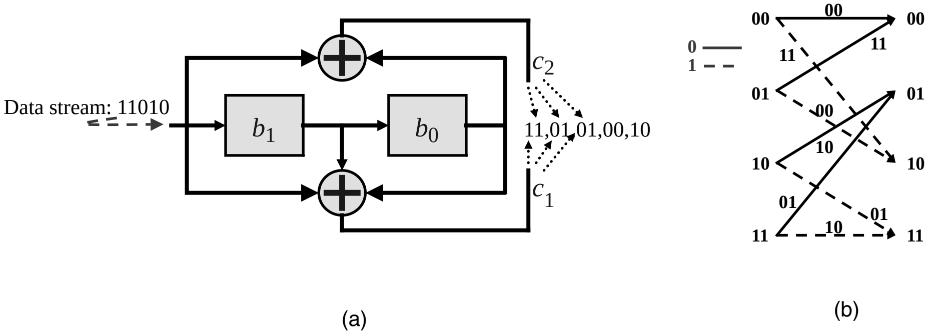


FIGURE 4.11 Convolutional coding with its respective state machine.

Convolutional encoders are implemented using simple feedback shift register structures. The input data stream and the memory of the past bits are used to generate several streams of coded data. For each arriving bit of information we have a block of information formed by encoded bits from different streams.

Example 4.11: Encoder for a ½ Rate Convolutional Code Figure 4.11*a* shows a simple convolutional code that utilizes two shift registers along with the stream of input bits. When the system is at rest, all the shift registers and outputs are set to 0. When the first input bit is 1 we have $D_1 = 0$ and $D_2 = 0$. Adding (XOR-ing) the outputs of the shift registers and input bit will result in $C_1 = D_1 \oplus D_2 \oplus b_i = 0 \oplus 0 \oplus 1 = 1$ and $C_2 = D_2 \oplus b_i = 0 \oplus 1 = 1$, resulting in a transmitted output of 11. In the next clock time, another 1 arrives at the input of the system. The values of the shift registers are now shifted to $D_1 = 1$ because of the previous input bit and $D_2 = 0$ because of the previous value of D_1, which was 0. The outputs are then computed as $C_1 = D_1 \oplus D_2 \oplus b_i = 1 \oplus 0 \oplus 1 = 0$ and $C_2 = D_2 \oplus b_i = 0 \oplus 1 = 1$. The transmitted bits are then 01. Repeating the steps for all the input bits results in transmitted bits being $T = [1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0]$, with the leftmost bit transmitted first.

Figure 4.11*b* illustrates a different representation of the coding process in a convolutional encoder using a state or trellis diagram. The state machine of Fig. 4.11*a* has two bits of memory which can form four different states shown in Fig. 4.11*b*. Depending on the arriving bit of information and the current state of the machine, the next state is determined. A “0” information bit is denoted by a solid line and a “1” information bit by a dashed line. The encoded two bits associated with each of the information bits are marked on the line representing that information bit. This diagram gives a different view of the encoder operation. Decoding the received convolutional coded data stream is performed using the *Viterbi algorithm* and trellis decoding which exploits this visualization of the process. A Viterbi decoder uses the trellis structure and the received coded stream of data to find the maximum likely transmitted information bits. This is best described with an example.

Example 4.12: Decoding a Convoluted Message Using a Viterbi Decoder In Example 4.11 the sequence of information bits was [11010], which was encoded into the transmitted bits $T = [11\,01\,01\,00\,10]$ with two encoded bits per information bit. Assume that during transmission the second and fifth bits from the left get corrupted so that the receiver detects them erroneously. Then the detected received sequence would be $R = [1\boldsymbol{\mathit{0}}\,01\,\boldsymbol{\mathit{1}}1\,00\,10]$. Figure 4.12 shows how a Viterbi decoder works on this received data to detect the original information bits regardless of the two errors which have occurred during the transmission. When the first two

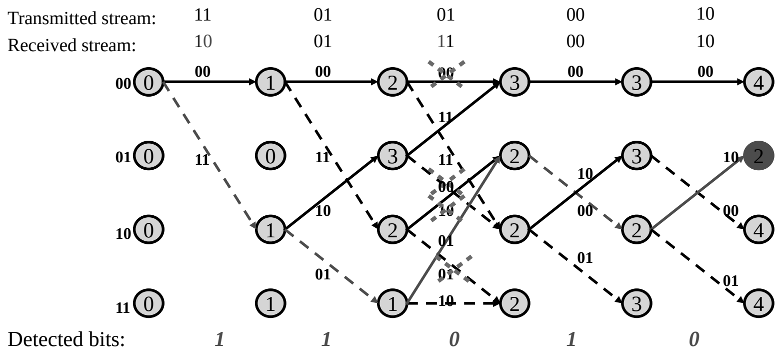

Transmitted stream: 11 01 01 00 10
Received stream: 10 01 11 00 10
00
01
10
11
00
11
10
01
Detected bits: 1 1 0 1 0

FIGURE 4.12 Viterbi decoder for convolutional coding.

bits "10" arrive we start from state "00"; if the information bit associated with the received "10" bits was a "0" then we should have received "00" and we will have a one-bit error that is kept inside the circle representing the next state. Similarly, if the information bit was "1" then we should have received a "11," which has a one-bit error from what we actually received, so we keep that in the next state.

When we process the second received pair of bits, "01", we continue counting the number of errors for all four possible paths. For example, if from state "00" we had a "0" information bit then we expect another "00" received rather than "01" that we have actually received. So the next state has one more error on top of the one we already had in this path and we keep track of the two errors inside the next state. Similarly, we get associated errors for all other three possibilities.

In the next stage of the trellis we have two routes to arriving to a state. In these situations we simply eliminate the route which causes a larger number of errors. For example, arriving at "00" in the fourth stage we have one path with two new errors and two from the last stage and another with no new and three past errors; so we take the second route, which causes fewer errors. We continue the same way for the next two stages to complete the trellis for all received coded bits. In this stage we select the state with minimum errors, which is state "01" with two errors, and we trace the arriving routes to detect the transmitted information bits. This state is connected to the previous state "10" with a solid line, indicating a "0" information bit for that stage. The previous states are connected by dashed–solid–dashed–dashed lines, indicating detection of [11011], which was the actual information sequence, which is now detected correctly in spite of the two errors in transmission of the coded bits.

In a manner similar to block codes, convolutional codes can be compared based on their respective Hamming distances (sometimes called the free distance). Like block codes, the error correction capability of convolutional code is given by $\lfloor (d_{\min}-1)/2 \rfloor$. However, the distance in convolutional codes is defined based on collective distances across an entire path. For example, in Fig. 4.12, the distance between the decoded path [11 01 01 00 10] and the top path connecting zero states [00 00 00 00 00] is 5, which has allowed correction of two bit errors.

In practice, convolutional and block codes are often mixed to form a packet. For example, in IEEE 802.11g, the data stream is scrambled before convolutional coding is applied and, after convolutional coding, block interleaving takes place. The following provides another example of mixed coding used for formation of speech packets in GSM, a 2G cellular system.

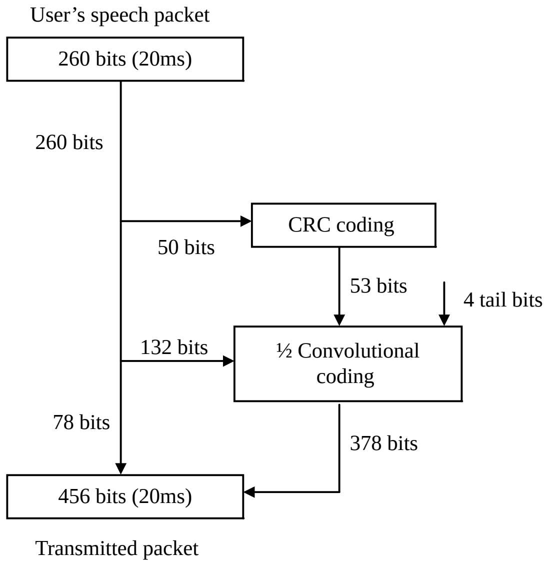


FIGURE 4.13 GSM speech packet.

Example 4.13: Mixed Coding for Preferred Data Integrity The speech coder in a GSM cellular telephone network, shown in Fig. 4.13, produces a data stream of 13 kb/s forming 260 bits of user speech packet every 20 ms. The speech encoder has three classes of bits with different protection requirements: class I are 50 bits, which are very important for the encoder; class II is 132 bits, demanding medium protection; and class III has 78 bits with no need for additional protection. The most significant 50 bits of the class I bits are enhanced with a block code that adds three parity bits. The sum total of 182 class I and class II bits plus the three parity bits, plus four tail bits (189 bits in all) are passed through a ½ rate convolutional coder to produce 378 bits. The 78 class III bits are added to these 378 bits without any encoding to produce 456 bits of encoded data. The effective transmission rate for this mixed coding scheme is 456 (bits)/(20 ms) = 22.8 kb/s and the overall rate is 260/ 456 = 0.57. High-rate and complex coding techniques are commonly used for error control in real-time speech transmission over long-haul facilities where having a feedback from the receiver is not practical.

Punctured Codes. In the case of punctured codes, some of the parity check bits are eliminated. This improves the efficiency (increases the code rate) but reduces the error detection/correction capability. Puncturing is an easy way of accomplishing a certain code rate from a given error control code. Consider an example of a punctured code that uses a basic ½ rate convolutional code to generate a k/n rate code. The input stream is first coded into two streams using the ½ basic convolutional codes. Parts of the encoded bits are then eliminated from transmission to change the rate of the code. The bit elimination process is implemented using a puncturing matrix which is applied to the two encoded data streams. For example, the puncturing matrix

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

is used produce a rate $^2/_3$ punctured code from the two output streams of the $^1/_2$ convolutional code. This matrix eliminates every other bit of the stream on the first branch and transmits all the bits in the second branch. This way, each block of 2 bits of information generates a block of 4 bits with $^1/_2$ convolutional code but only three of the $^1/_2$-coded bits are transmitted, resulting in an effective $^2/_3$-rate code. At the receiver the Viterbi decoder includes the elimination of the bits in calculation of the distances. Punctured convolutional codes are widely used in the satellite communications, digital video broadcasting, and wireless LANs and WANs.

Trellis-coded Modulation. The first application of QAM (see Chapter 3) was in the early 1970s in the voice-band data communication modems with the invention of the 9600 bits/s modems over four-wire commercial telephone lines. The next major invention first implemented for voice-band data communications was TCM, which was introduced in early 1980s [Ung87]. The importance of this technique is that it combines the coding and design of signal constellations to achieve higher data rates. Prior to the invention of TCM, it was a common belief that coding could not usefully increase the data rate over band-limited channels such as telephone lines. Today, QAM with TCM is used in broadband DSL and cable modems (802.14) and many other applications. QAM modulation has a two-dimensional signal constellation that doubles the data rate by running two orthogonal data streams over the same band. TCM doubles the *points in the constellation*, which calls for a 3 dB loss in the performance but correlates the sequence of transmitted symbols using a sequential convolutional code which provides large gains so that the overall performance can be 3–6 dB better than the original constellation with half of the points. TCM laid the ground for 14.4 and then 19.2 kb/s modems, which were later extended to higher data rates using data compression techniques.

Turbo Codes. The term turbo codes is generally used to describe a class of high-efficiency source coding schemes used for telecommunication applications in limited bandwidth scenarios. Turbo codes are considered one of the few coding schemes to approach the theoretical Shannon limit of transmission rate of communication over a noisy channel. The advantages of turbo codes over other source coding schemes lie mainly in the ability of increasing the data rate with a limited power transmitter and decreasing the transmitted power for a fixed data rate. However, very complex designs and implementation at transmitter and receiver sides and high latency introduced by turbo codes are disadvantages of these codes.

Fundamentally, at the transmitter side, the encoder forms three blocks of data. The first n bits contain the original data. The next $k/2$ bits represent the parity bits of the original data, which are computed using a recursive convolutional code. The last $k/2$ bits represent the parity bits of a known permutation of the data using the same recursive convolutional code. The total code rate of the turbo code is then calculated as $n/(n + k)$. At the other end of the communication link, the receiver forms two decoders corresponding to the two convolutional encoders, with an interleaver in between. Each of the decoders generates a line delay which synchronizes the received data with their corresponding parity bits. Various turbo codes are implemented in satellite communication, 3G cellular networks, and IEEE 802.16 (WMAN standard).

4.3.4 Codes for Manipulating Data

The data streams in modern networks are often manipulated before transmission. The manipulations performed include scrambling or block interleaving. In the next two sections we provide short descriptions and simple examples of scrambling and block interleaving techniques.

Scrambling. Scramblers are used in wireless networks to provide some resilience to eavesdropping when data is broadcast over the air. This way, for example, a cellular network provider can protect the customer's data on the air and a broadcast satellite service provider can make sure that only its subscribers can use their channels. Another example application for scramblers is to avoid the transmission of large sequences of similar bits, either 0s or 1s. A scrambler in this context is different from encrypting the data to make it unintelligible by unwanted users. In this context, a scrambler replaces the data stream by another stream, removing the possibility of occurrence of strings of similar bits not suitable for a synchronization process at the receiver. However, they are very useful in randomizing the data, particularly when nothing is transmitted and the transmission medium is idle. To ensure reception of data with adequate transitions in the stream, as we explained in Chapter 3, line coding techniques are used.

A scrambler is a pseudo randomizer device that manipulates a data stream before transmission. The manipulations are reversed by a descrambler at the receiver to recover the original data stream. Simple scramblers are implemented using LFSR codes and the implementation of the scrambler and the descramblers are very similar. An example will further clarify the basic implementation of a scrambler and a descrambler.

Example 4.14: Scrambler and Descrambler Implementation Figure 4.14 presents a typical scrambler and descrambler implemented with shift registers. The polynomial representing the scrambler is $G(x) = 1 + x^4 + x^7$. Figure 4.15 presents the general block diagram of the IEEE 802.11g transmitter with the above scrambler placed before the convolutional encoder. The block interleaving is performed after FEC and before data is sent to the OFDM modem.

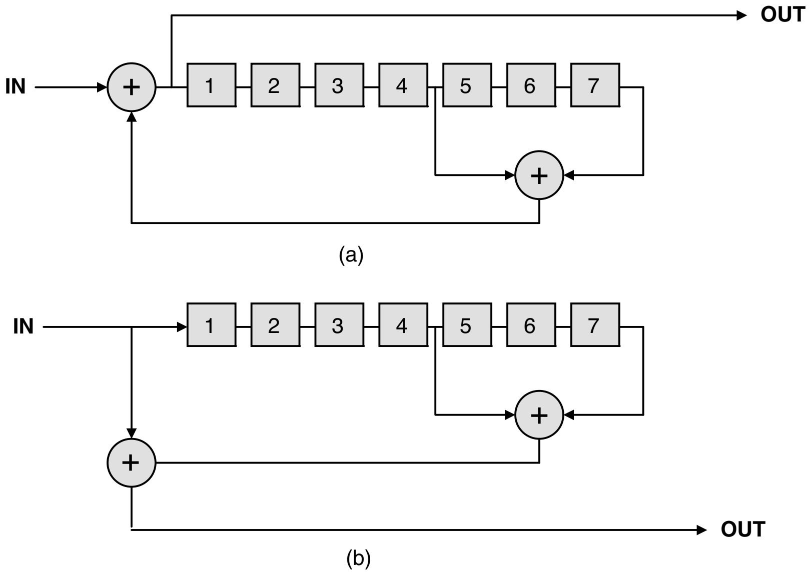


FIGURE 4.14 Implementation of a typical (*a*) scrambler and (*b*) descrambler, using shift register codes.

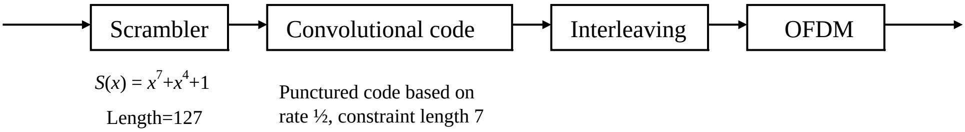


FIGURE 4.15 Overall block diagram of IEEE 802.11g using a scrambler to randomize data, convolutional codes for FEC, and block interleaving to spread blocks of errors.

Block Interleaving. Block interleaving is very popular in wireless transmissions. In transmissions over wireless media, errors often happen in blocks of bits (burst errors) which are difficult to correct. Recall that codes considered in this chapter typically can detect small numbers of bit errors within a set of received bits. Block interleaving is used to spread the error over a larger interval of time so that the number of bit errors within a given set of bits is small, thereby allowing error control codes to detect or correct errors and increase the data integrity. For instance, consider a Hamming code that can correct single-bit errors over codewords of size 7 bits. This means that if there is a single error over a block of 7 bits, the coding scheme can correct it. On the other hand, a burst of five errors cannot be corrected by this code. If, however, we can spread these errors over five codewords so that each codewords "sees" only one error, then it is possible to correct each of the errors. One way of accomplishing this is as follows. Codewords are arranged one below the other and bits are transmitted vertically. At the receiver, the codewords are reconstructed and the bits are decoded horizontally. Since the burst of errors affects the serially transmitted vertical bits that are spread over several codewords, the errors can be corrected. Block interleaving introduces delay because several codewords have to be received first before the voice packet can be reconstructed. There is only so much delay that is acceptable for normal voice conversations, and the interleaving process should not create an unacceptable value of delay in the process.

Example 4.15: Scrambling and Block Interleaving For the GSM packets shown in Fig. 4.13, the output of the convolutional encoder consists of 456 bits for each speech stream input of 260 bits. If the 456 bits are split into eight blocks of 57 bits each, then the 57 bits are spread over eight frames of duration 4.6 ms carrying eight user packets so that, even if one frame out of eight is lost, the speech quality is not affected. The delay in reconstructing the codewords corresponds to the reception of eight frames, which takes $4.6 \times 8 = 37$ ms, which is less than the 50 ms usually tolerable for processing of voice conversations.

4.4 CODING FOR SPREAD-SPECTRUM AND CODE-DIVISION MULTIPLE ACCESS SYSTEMS

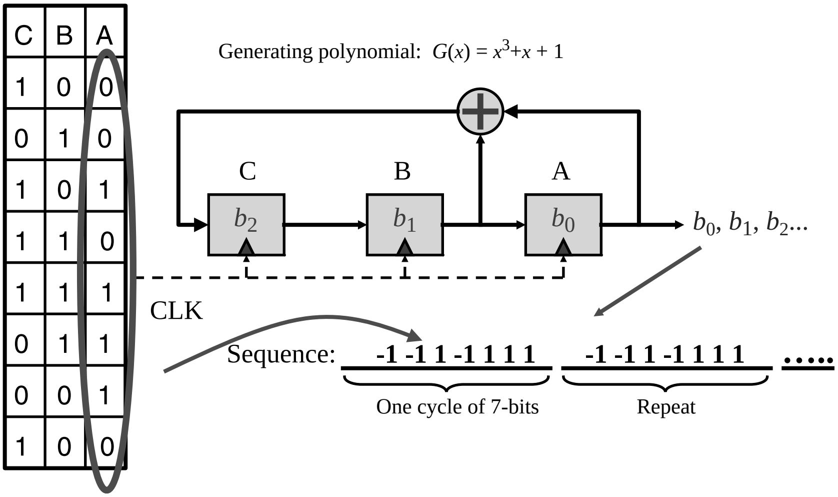
In this section we introduce two fundamental codes used in spread-spectrum and CDMA systems. These are the PN sequences and M-ary orthogonal codes. PN sequences are used for spread-spectrum systems, which were described in Chapter 3, as well as a variety of other applications in security, authentication, and localization. M-ary orthogonal codes are used either to address a specific user in the CDMA network or to increase the reliability of transmission over a disturbed channel. Examples of systems using these coding techniques are given in Chapters 7, 9 and 10 when we discuss CDMA cellular, IEEE 802.11 WLAN, and IEEE 802.15 WPANs in the networks.

4.4.1 Pseudo Noise Codes

PN codes are also called pseudorandom sequences and they are used as codes for implementation of spread-spectrum systems. Since the sequenced pattern is only coded to appear random, the sequences are referred to as pseudorandom or PN sequences or codes. Even though they are called pseudorandom, the sequences are not random, but appear random since they contain almost equal numbers of 0s and 1s. The most common PN sequences are the maximal-length sequences or M-sequences that are created using maximal-length LFSRs of length m. The name arises from the fact that the M-sequences are the longest sequences that can be generated by an m-stage LFSR. The contents of the LFSRs repeat after a cycle of $2^m - 1$ shifts. The length of the PN sequence before it repeats itself is also $2^m - 1$. The maximal-length shift registers are represented by polynomials that denote the connections that are active in the LFSR.

Example 4.16: PN Sequence of Length 7 Figure 4.16 shows an LFSR of length $m = 3$ and its 7-bit maximum-length PN sequence. The feedback connections from the first (exponent 0) and second (exponent 1) shift registers (but not the third) are represented by $G(x) = 1 + x + x^3$. The output sequence of 7 bits {0,0,1,0,1,1,1} repeats itself every 7 bits. The states of the LFSR and the outputs are also shown in Fig. 4.16. If we consider the state of the machine as a 3-bit binary number, as shown in Fig. 4.16, the sequence of states is {4,2,5,6,7,3,1}, which also repeats every seven clock times.

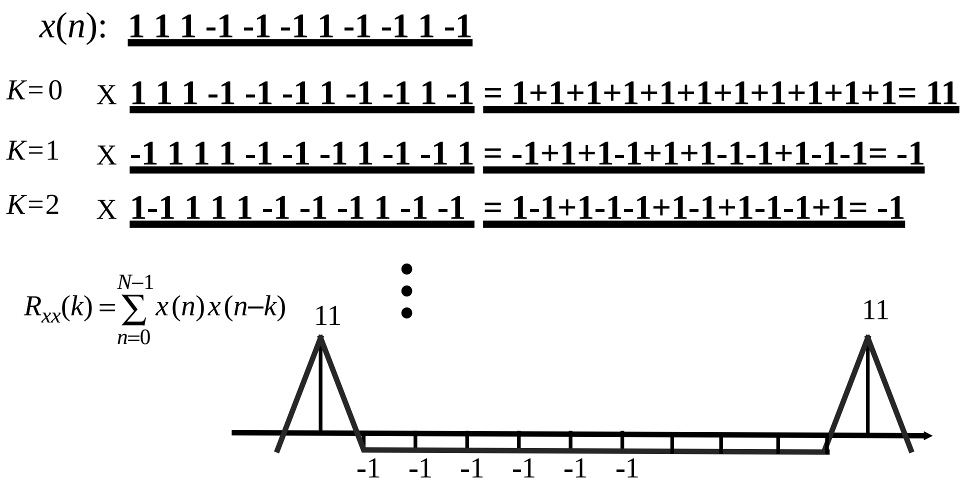
PN sequences are widely employed because of their nice properties. Some of the important properties include the fact that they have nearly equal numbers of 0s and 1s and their autocorrelation exhibits a strong peak. The periodic autocorrelation has a peak value of $2^m - 1$ for zero lag and a value of -1 for all other lags. Consequently, it is possible to differentiate between users by computing the correlation between their allocated sequences and a replica of the sequence of the user under consideration. Another popular PN sequence in information networking is the Barker sequence used in IEEE 802.11 and 802.11b standards. The common Barker code used in 802.11 has a sequence of length 11 and it is given by $\{1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \quad 1 \quad -1 \quad -1 \quad 1 \quad -1\}$. The autocorrelation property of the Barker codes is similar to the LFSR codes, but the length is not restricted to $2^m - 1$. The following example shows the calculation of the autocorrelation property of the Barker codes.



C	B	A
1	0	0
0	1	0
1	0	1
1	1	0
1	1	1
0	1	1
0	0	1
1	0	0

Generating polynomial: $G(x) = x^3 + x + 1$
C
B
A
b_2
b_1
b_0
$b_0, b_1, b_2 \ldots$
CLK
Sequence: -1 -1 1 -1 1 1 1 -1 -1 1 -1 1 1 1 .....
One cycle of 7-bits
Repeat

FIGURE 4.16 LFSR codes of length 7.



$x(n)$: 1 1 1 -1 -1 -1 1 -1 -1 1 -1

$K=0$ X 1 1 1 -1 -1 -1 1 -1 -1 1 -1 = 1+1+1+1+1+1+1+1+1+1+1= 11

$K=1$ X -1 1 1 1 -1 -1 -1 1 -1 -1 1 = -1+1+1-1+1+1-1-1+1-1-1= -1

$K=2$ X 1-1 1 1 1 -1 -1 -1 1 -1 -1 = 1-1+1-1-1+1-1+1-1-1+1= -1

This circular ACF follows the same pattern as the ACF of the LFSR codes

FIGURE 4.17 Autocorrelation of the Barker code.

Example 4.17: Autocorrelation of the Barker Code of Length 11 The autocorrelation of a sequence is defined by

$$R_{xx}(k) = \sum_{n=0}^{N-1} x(n)x(n-k)$$

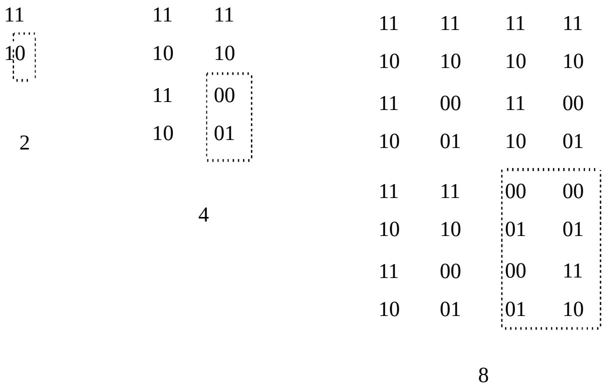
For the 11-bit Barker code, the sequence is $x(n) = \{1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \quad 1 \quad -1 \quad -1 \quad 1 \quad -1\}$. Figure 4.17 shows the first few steps in calculation of the ACF of this code. For $k = 0$ the sequence is multiplied by itself, resulting in all $+$ 1 terms, which sum up to 11, representing the peak of the ACF. For $k = 1$ the sequence is shifted one time in a circular manner, resulting in a new sequence $x(n - 1)$. Multiplication of $x(n)x(n - 1)$ and addition of all terms results in a $-$ 1. The same thing happens for $x(n)x(n - 2)$ and all other values of the ACF until we shift the sequence 11 times, which brings us to the starting point. The ACF of the Barker code follows the same pattern as PN sequences and it has a peak every 11 chips of value 11 with all other values fixed at -1.

4.4.2 *M*-ary Orthogonal Codes

M-ary orthogonal codes, as the name indicates, provide a set of codewords which are orthogonal.[1] The set of orthogonal sequences, or *orthogonal-codes*, will typically be derived from a Hadamard or Walsh code. Walsh codes are based on the Hadamard matrix that is defined by the recursion

$$\mathbf{H}_{2N} = \begin{bmatrix} \mathbf{H}_N & \mathbf{H}_N \\ \mathbf{H}_N & \overline{\mathbf{H}_N} \end{bmatrix}$$

[1] Two codewords $\mathbf{W}_1 = \{a_1, a_2, \ldots, a_N\}$ and $\mathbf{W}_2 = \{b_1, b_2, \ldots, b_N\}$ are orthogonal if $\sum_{i-1}^{N} ai \oplus bi = 0$, where the summation sign means exclusive-OR sum.



$$\begin{array}{c} 11 \\ 10 \end{array}$$

2

$$\begin{array}{cc} 11 & 11 \\ 10 & 10 \\ 11 & 00 \\ 10 & 01 \end{array}$$

4

$$\begin{array}{cccc} 11 & 11 & 11 & 11 \\ 10 & 10 & 10 & 10 \\ 11 & 00 & 11 & 00 \\ 10 & 01 & 10 & 01 \\ 11 & 11 & 00 & 00 \\ 10 & 10 & 01 & 01 \\ 11 & 00 & 00 & 11 \\ 10 & 01 & 01 & 10 \end{array}$$

8

FIGURE 4.18 Walsh codes of orders two, four, and eight.

Figure 4.18 shows the Hadamard matrix of orders two, four, and eight. The rows of this matrix provide for a set of orthogonal codes referred to as Walsh codes. Therefore, the first 8-bit Walsh code is $\mathbf{W_0} = [0\,0\,0\,0\,0\,0\,0\,0]$ and the last 8-bit Walsh code is $\mathbf{W_7} = [0\;1\;1\;0\;1\;0\;0\;1]$. In general, Walsh codes of any order are orthogonal:

$$\mathbf{W}_i \times \mathbf{W}_j = \begin{cases} N, \; i = j \\ 0, \; i \neq j \end{cases}$$

because, for each pair, half of the bits are different and the other half of the bits are the same. M-ary orthogonal codes are used in different contexts in wireless networks. In CDMA links, the orthogonal codes are used to separate different user channels. The problem with orthogonal sequences is that the users must be synchronized, since orthogonal sequences do not have good correlation properties outside of the zero-lag case. Consequently, in some CDMA networks, orthogonal sequences are employed on the downlink channel where the BS is able to synchronize transmissions. For the less reliable uplink, synchronization sometimes employs PN sequences over orthogonal sequences. M-ary orthogonal codes are also used by individual users to improve the quality of the received signal at the BSs. In the uplink of some of the CDMA networks and some WLANs, M-ary orthogonal codes are used by individual users to improve the quality of transmission. In this approach, a user has a set of orthogonal sequences representing a set of coded symbols for transmission.

Example 4.18: 3-Bit to 8-Bit* M*-ary Orthogonal Codes Consider the 8 × 8 Walsh codes shown in Fig. 4.18. Then assume that we form blocks of 3 bits from the incoming data stream and use each 3 bits as an address to select one of the eight orthogonal codes and transmit that 8-bit code instead of the original 3 bits. Then, at the receiver, when we receive the 8-bit code we compare it with all possible 8-bit codewords and select the code which has the minimum distance with the received coded symbol. Since all codes are 4 bits different from one another, if we have a 1-bit error we can correct it and if we have up to a 3-bit error we can detect it. This is a very strong code with the coding rate of $^3/_8$ and will substantially improve the quality of the received data stream. A 64-ary orthogonal code is used in the upstream of QUALCOMM's first CDMA cellular network.

Another popular example of M-ary orthogonal codes used in information networking is the complementary code (complementary code keying is used in the IEEE 802.11b) which is described in Chapter 9. This orthogonal code uses four complex symbols of a quadrature phase-shift keying (QPSK) constellation to form the orthogonal set of transmitted symbols.

4.5 ARQ SCHEMES

In the PSTN, if a voice packet is received in error, then it is either dropped or replaced with an attenuated version of the previous packet to provide a semblance of continuity. Retransmissions of damaged packets are not considered because of the sensitivity of voice to delay. ARQ schemes essentially are used in data networks where reliability of received information is of paramount importance and delay is less of a problem compared with real-time multimedia applications. Using ARQ schemes, if a block of data is received in error then the receiver requests retransmission of the block of data. This request may be explicit or built into several protocols already operational in the system for flow control or other purposes. Usually, an acknowledgement packet is employed to indicate correct reception of one or more transmitted packets. If an acknowledgement is not received in a certain time frame, or a negative acknowledgement is received, then the transmitter will retransmit the packet.

There are three basic ARQ schemes. The stop-and-wait ARQ scheme waits for an acknowledgement for each individual packet before sending the next one. This is especially inefficient if the round-trip times are large, because the transmitter spends a lot of time waiting for the acknowledgement. In order to improve upon this scheme, the go-back-N ARQ scheme transmits up to N packets at a time and waits for acknowledgements. Multiple packets can be acknowledged with one response. Depending on the receipt of acknowledgements, the transmitter will back up to the last correctly received packet and retransmit the following ones (N or less packets). It is possible that some of the subsequent packets are received correctly, but they will be discarded. In order to eliminate this inefficiency, a selective-repeat ARQ scheme can be employed. Here, only those packets that are received in error are retransmitted.

Example 4.19: Acknowledgements and Retransmissions in a WLAN In most WLAN applications, every transmitted packet is acknowledged because the channel is unreliable. It is similar to a stop-and-wait protocol in that sense. Because round-trip times are small, and both the AP and the mobile stations share the channel, the inefficiency is limited. It is often possible to piggyback the acknowledgements over data packets transmitted in the other direction.

4.5.1 Stop and Wait

In the simplest form of ARQ scheme, the stop-and-wait scheme transmits the encoded data, waits for acknowledgement of correct detection and reception of the symbol (or frame) at the receiver side, and then transmits the next symbol (or frame). If a symbol (frame) is detected erroneously at the receiver, then an acknowledgement is not sent to the transmitter and the receiver simply expects a retransmission. The process is repeated until the symbol (frame) is correctly received. If the transmitter does not receive the acknowledgement from the receiver within a certain time limit, the transmission times out and the symbol (frame) is retransmitted. An extra altering bit, known as sequence bit, is prepended to the frame in order to avoid problems of losing the acknowledgement bit and/or latency concerns raised in such schemes.

Example 4.19 Figure 4.19 illustrates an example of the basics of stop-and-wait schemes. Frames are numbered 0, 1, 0, 1, and so on. The first frame 0 is sent and acknowledged. Frame 1 is lost and resent after the time-out. The resent frame 1 is acknowledged and the timer stops. Frame 0 is sent and acknowledged, but the acknowledgment is lost. The sender has no idea if the frame or the acknowledgment is lost; so, after the time-out, it resends frame 0, which is acknowledged.

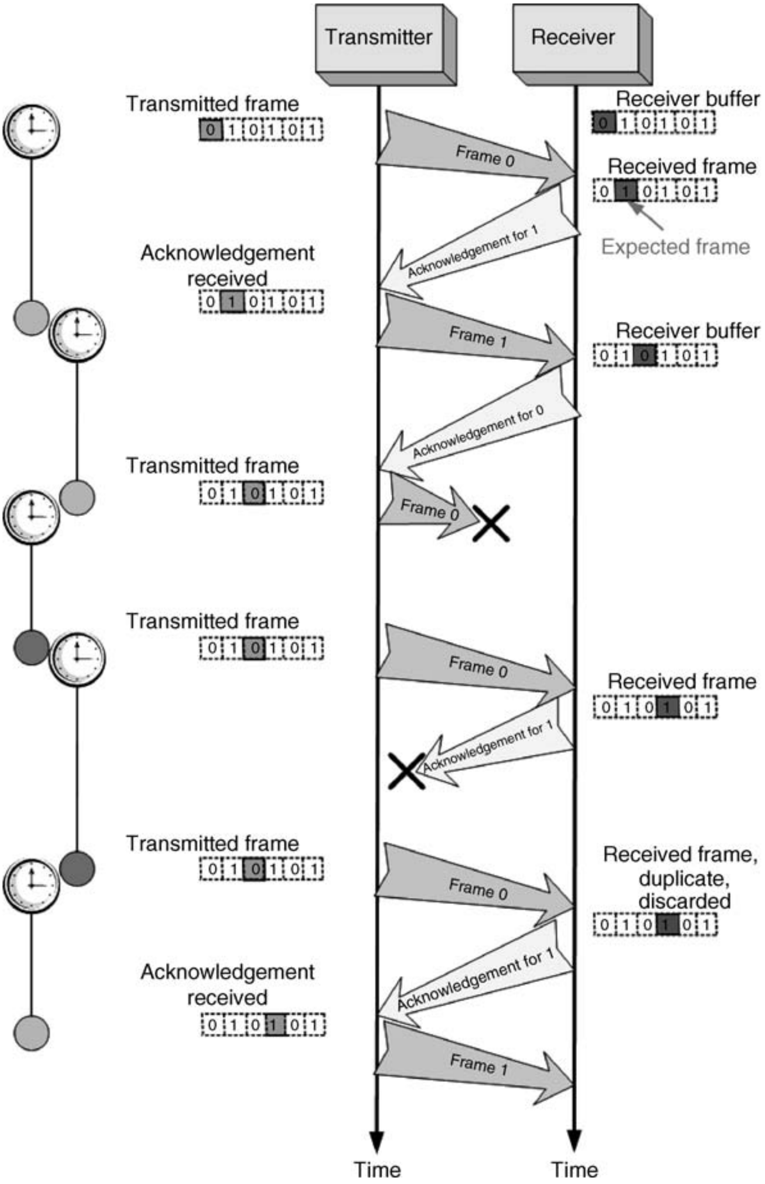


FIGURE 4.19 Stop-and-wait scheme in ARQ.

Example 4.20: Efficiency of Stop-and-Wait Scheme Assume that in a stop-and-wait ARQ system the bandwidth of the line is 1 Mb/s and it takes 20 ms of propagation delay to make a round trip between the two terminals using the scheme. The bandwidth-delay product of system is $(1 \times 10^6) \times (20 \times 10^{-3}) = 20000$ bits, which implies that the propagation delay is equivalent to the length of so many bits. Assuming that the system data packets are 1000 bits in length and the length of the acknowledgement packet is negligible, the link utilization is only 1000/20000 or 5%. For this reason, for a link with a high bandwidth or long delay, the use of stop-and-wait ARQ wastes the capacity of the link.

4.5.2 Go-Back-*N*

In a manner similar to the stop-and-wait approach, go-back-*N* sends frames of data within its window size and waits for their acknowledgement. The difference between the two approaches is the number of frames during the time interval of waiting for acknowledgement. In the go-back-*N* ARQ scheme, several frames are sent by the transmitter. The receiver keeps track of their sequence number. The receiver, however, sends acknowledgment for each frame and the transmitter keeps track of the acknowledgements. In this scheme, the channel is still being used by the transmitter and receiver while acknowledgements are being waited on and the throughput of the network increases. In the case of erroneous detection of a frame (or a lost frame), the transmitter does not receive its corresponding acknowledgement and goes back *N* steps and retransmits the frame onwards.

Example 4.21: Go-Back-*N *Scheme The go-back-*N* scheme can be considered as a generalization of the stop-and-wait scheme in which the size of the window is larger than one packet. However, as illustrated in Fig. 4.20, the receiver acknowledges each packet separately and will not accept any packets with a sequence number after the current window's sequence number.

4.5.3 Selective-Repeat Automatic Repeat Request

One disadvantage of the go-back-*N* scheme is the need to retransmit several frames when a frame is lost. In order to avoid this issue, the selective-repeat ARQ scheme is used. This scheme allows transmission and reception of frames even after the occurrence of an error. The receiver, on the other hand, keeps track of consecutive frames and when a frame is lost or damaged it sends the sequence number corresponding to it with every acknowledgement it sends. Eventually, the transmitter retransmits the lost frame after emptying its buffer.

The size of the buffer is generally selected with respect to the size of the sequence number to avoid confusion. In communication protocols with variable lengths of frames, a variant of this scheme is used to transmit and acknowledge reception of the blocks of the original frame.

Example 4.22: Selective-Repeat ARQ Figure 4.21 illustrates the basic operation of selective-repeat ARQ. In the selective-repeat ARQ scheme, the transmitter sends the packet with the lowest sequence number. The transmitter then sends consecutive packets regularly until it receives an acknowledgement (or acknowledgement of no reception of a packet with a specific sequence number) or it is timed out for the first transmitted packet. If timed out, it transmits the first packet again, and continues by sending the consecutive packets. If, during the transmission, it receives acknowledgement for some of the packets but not all, it

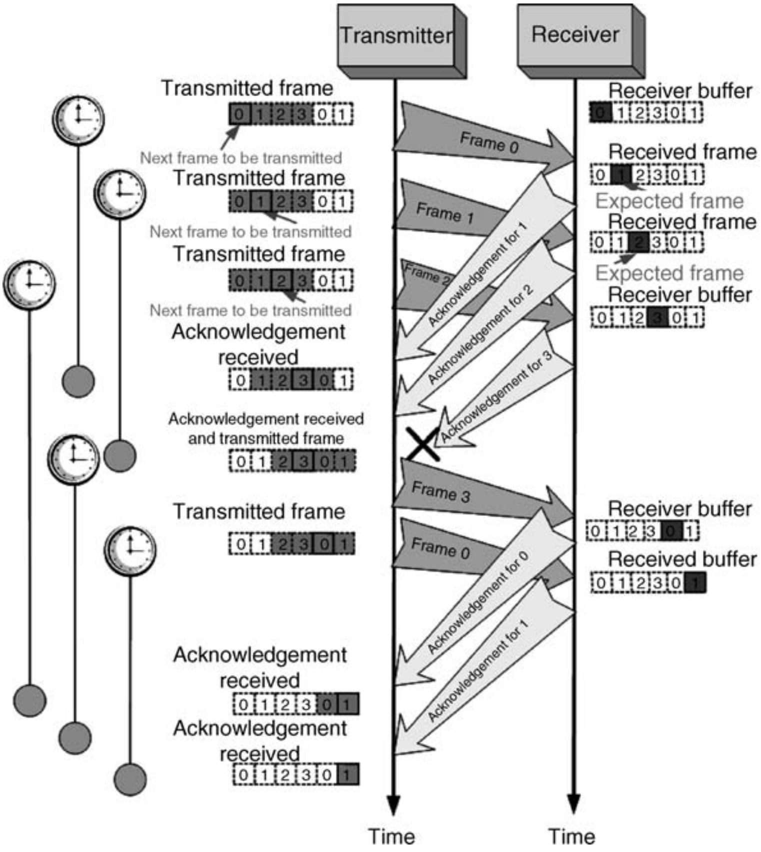


FIGURE 4.20 Go-back-*N* scheme in ARQ.

transmits the lost packets again. It is a common practice to keep the size of the buffer as half of the sequence number size.

4.5.4 Hybrid Automatic Repeat Request

In addition to the above methods, it is possible to combine the ARQ schemes with FEC schemes and form a hybrid ARQ scheme. Adding the redundancy information of the FEC schemes increases the probability of receiving a valid frame in very noisy environments but decreases the overall throughput of the network. Typical FEC codes which are utilized in hybrid ARQ are RS, BCH, and turbo codes. At the receiver side, when a frame is received it is first decoded and its validity is checked by checking the FEC scheme output. If the frame is received without problems, or the errors are correctable, the receiver sends an acknowledgement to the transmitter. On the hand, if the frame is lost or its errors cannot be corrected, a negative acknowledgement is sent to the transmitter, which then performs a retransmission.

Example 4.23: Hybrid ARQ Figure 4.22 shows the basic operation of the hybrid ARQ scheme. The transmitter sends the packets to different receivers. Each receiver either

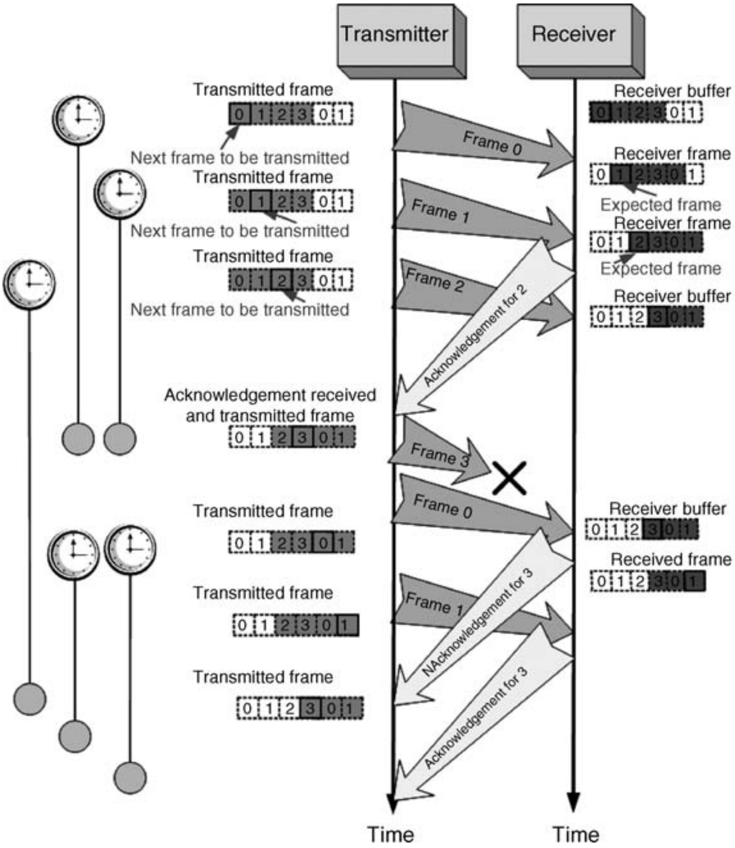


FIGURE 4.21 Selective repeat scheme in ARQ.

completely retrieves the stream of the packets or partially receives the packets stream. In the case of erroneous detection of a packet or loss of a packet at each receiver, the receiver sends a signal to the transmitter. The transmitter then calculates the FEC block code of the packets and sends them to the receivers so that each receiver can retrieve the entire bit stream.

4.6 FLOW CONTROL PROTOCOLS

Managing and controlling the rate of data transmission between two nodes in order to avoid overflowing at the receiver side is referred to as flow control. If a transmitter node transmits the data at a faster rate than the receiver can process, then eventually the input buffer at the receiver side will be full and no packets can be accepted by the buffer. In such conditions, packets are lost and their retransmission decreases the efficiency of the system. Flow control mechanisms are designed to avoid the occurrence of such situations. The two basic approaches to flow control depend on the type of messaging between the nodes, particularly the acknowledgement message sent back to the transmitter from the receiver. Mechanisms can be designed to employ different rates in different directions of a communication link,

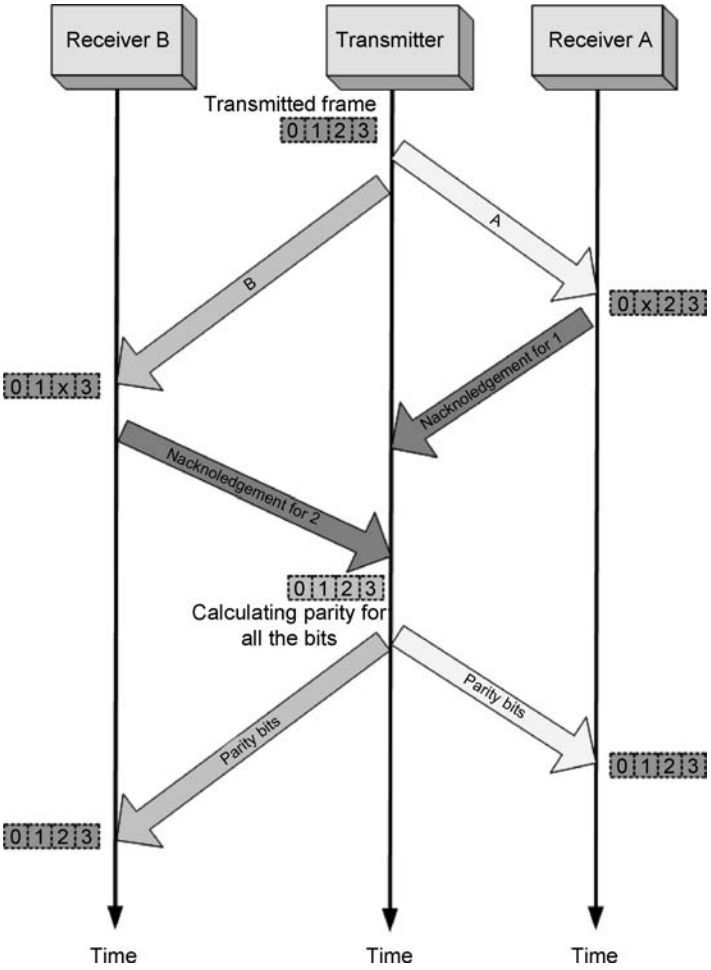


FIGURE 4.22 Hybrid ARQ scheme.

permitting one direction to transfer with a higher data rate. Both the network and transmitter can employ flow control mechanisms. Like error control ARQ mechanisms, the stop-and-wait mechanism and sliding-window mechanism are practical implementations of flow control protocols.

4.6.1 Stop and Wait

In the simplest form of flow control, a stop-and-wait mechanism is adopted between the nodes to avoid packet loss due to buffer overflow. A transmitter transmits the frame. If the frame was received in a valid format at the receiver side, the receiver sends back an acknowledgement response to the transmitter indicating its preparedness to accept another frame. The transmitter, on the other hand, will hold off transmitting until the time it receives the acknowledgement (with indication of preparedness) packet. The rate of transmission is then decided by the processing power of the receiver.

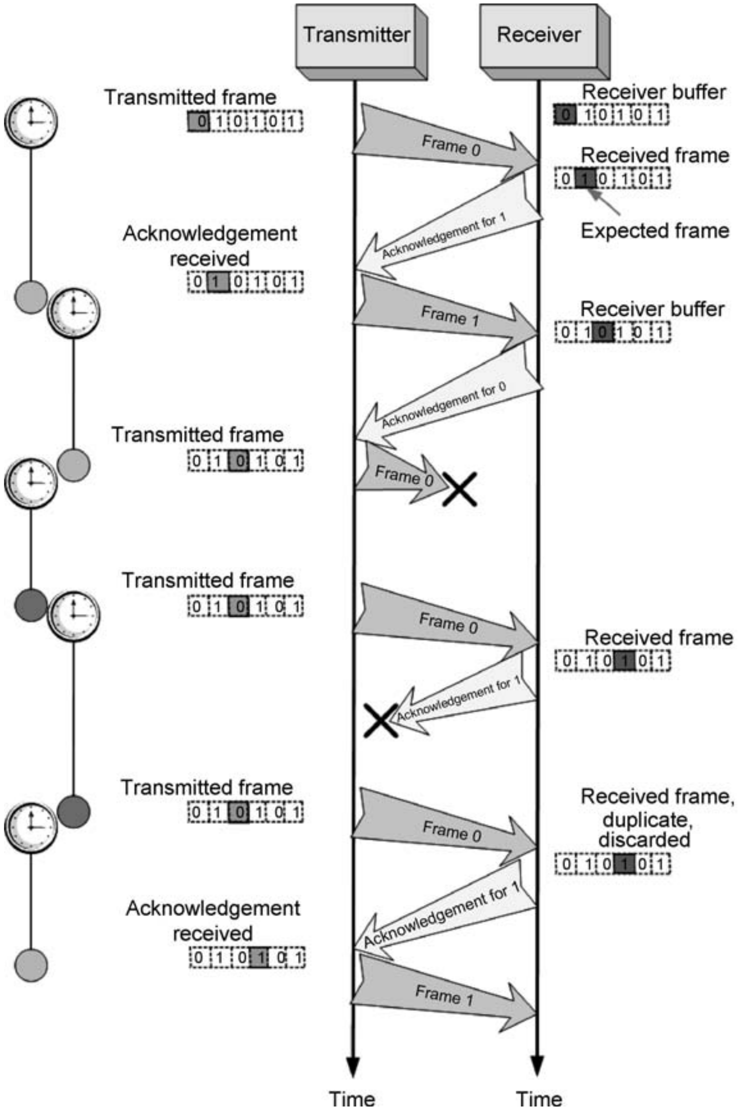


FIGURE 4.23 Stop-and-wait scheme in ARQ for flow control.

In practice, this method is ineffective when the size of the frame is large. To avoid the problem, the transmitter divides large messages into smaller frames and sends them to the receiver. The inefficiency of this approach in terms of utilizing the channel is also another practical issue. The issue becomes very problematic when a message is extremely short but the nodes are located at a far distance from one another.

Example 4.24: Stop-and-Wait Flow Control Protocol Figure 4.23 illustrates the basic stop-and-wait algorithm for flow control. In a noiseless environment, the transmitter transmits the first packet. Upon receiving the packet at the receiver side, the receiver sends an acknowledgement to the transmitter indicating that the first packet has been received flawlessly. The transmitter consequently transmits the next packet.

In a noisy environment, the first frame or its acknowledgement from the receiver gets lost in transmission and the transmitter times out before receiving the acknowledgement. In such cases, the transmitter resends the packet whose acknowledgement has not been received. A similar procedure takes place when the transmission delay is too long and the transmitter times out before receiving the acknowledgement, although the acknowledgement is received later on.

4.6.2 Sliding Window

The efficiency of the previous scheme can be improved by allowing multiple frames to be sent consecutively before receiving the acknowledgement packet from the receiver. Like the go-back-*N* scheme for error control, in the sliding-window scheme, a transmitter marks the frame with its corresponding sequence number and then sends multiple frames. The receiver acknowledges reception of a frame and sends an acknowledgement packet which contains the sequence number of the next expected frame. It is possible for the receiver to send a single acknowledgement packet containing acknowledgements for several frames and indicating the sequence number of the next frame. The transmitter assigns a window with length of allowable number of transmissions before receiving acknowledgement. Upon sending each extra frame, the window clears its leftmost content from the sliding-window buffer and receiving each acknowledgement packet extends the window and adds a valid sequence number to be sent to the rightmost place of the window. A similar approach takes place at the receiver side.

Example 4.25: Sliding-Window Protocol Figure 4.24 describes the flow of the sliding-window algorithm. The transmitter and receiver, initially, set their respective buffers to seven packets. The transmitter transmits the first packet, the second packet, and finally the third packet. The transmitter buffer then marks the transmitted packets as it transmits them. At the receiver side, as soon as the receiver receives the third packet, it first clears the packets from its buffer,

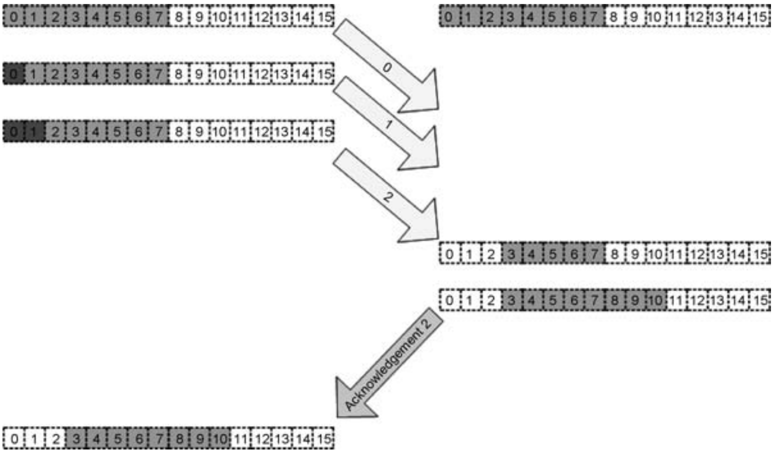


FIGURE 4.24 Sliding window scheme in flow control.

resizes the buffers, and sends back an acknowledgement to the transmitter to verify that three packets have been received in order. When the transmitter receives the acknowledgement of reception of three packets, it restarts its buffer starting from the fourth packet. This time, the transmitter sends four packets and clears the packets from its buffer. The receiver receives the four packets. If the order of the packets differs from the transmitted order, it sends a message back to the transmitter and indicates the packet sequence number that the receiver is currently waiting for.

QUESTIONS

1. What are the differences among source coding, channel coding and line coding techniques?
2. What are the most popular voice coding techniques used in cellular networks and how their rate compare with the PCM coding techniques commonly used in the PST?
3. Give the range of data rates needed for video coding techniques used in popular Internet applications.
4. What the difference between the FEC coding and ARQ coding? Given an example for application of these codes and the reasoning why it has been selected for that application.
5. Why do we need to frame the data stream?
6. What are the differences among bit- and character-stuffing techniques? Give an example application in which each of these coding techniques is applied.
7. What is Hamming distance of a block code and how does it relate to error detection and error correction capabilities of a coding technique?
8. Explain why Hamming codes were used in punch card application.
9. Explain why and how CRC codes are used in the IEEE 802.3 Ethernet frames.
10. What are non-binary codes and what is their advantage over binary codes?
11. What are the differences among convolutional codes and block codes in term of complexity of the implementation, ability in error detection and correction, and typical applications they are used in?
12. What is a punctured convolutional code?
13. What are the differences among traditional convolutional codes and TCM? Give application examples in information networking for each of the codes.
14. What are the pros and cons of the turbo codes?
15. What are the practical differences among scrambling and block interleaving techniques? Give an example information networking application for each technique.
16. What are the differences among LFSR and Barker PN-sequences?
17. How do we find out that a received packet is erroneous?
18. What is the difference between error control and flow control mechanisms?

PROBLEMS

Problem 1:

Assuming the following character encoding scheme:

A: 01000111; B: 11100011; FLAG: 01111110; ESC: 11100000

Give the transmitted bit sequence for the five character frame: A A B ESC FLAG

(a) If character counting technique is used.
(b) If flag byte with byte stuffing is used.
(c) If starting and ending flag bytes and bit stuffing is applied.

Problem 2:

Use bit stuffing technique on the following string of data bits "00110111110 111111011110".

Problem 3:

A CRC code has the generator polynomial $G(x) = x^4 + x + 1$.

(a) Give the coded stream if we apply the code to the following bit string "10001110".
(b) Assuming the second bit of the encoded digits is received in error, show that this error can be detected at the receiver.

Problem 4:

A 1 Mbit frame of binary data is transmitted over a transmission facility with 10 Mbps speed and a 250 msec end-to-end delay.

(a) Determine the channel utilization if the stop-and-wait protocol is used.
(b) Repeat (a) if Go-Back-8 protocol is used.

Problem 5:

(a) If the probability of occurrence of a bit error is p, what is the probability of having a single bit error in a block of five bits?
(b) Repeat (a) for probability of having exactly two bits of error.
(c) Repeat (a) for probability of having at most two errors.
(d) For $p = 10^{-3}$ calculate the numerical values for parts (a), (b), and (c), and give an intuitive explanation of why these differences exists.

Problem 6:

(a) Give the generator and parity check matrices for the Hamming (7,4) code.
(b) For the data block (0111) give the transmitted code word.
(c) If the third bit of the coded word is erroneously received determine the syndrome matrix and show that it can correct the error.

Problem 7:

(a) Give the code generation function for the systematic Hamming (11,7) code used for ASCII parity generation.
(b) Use the code generation matrix to generate parities for "H" [1001000].

(c) Assuming the third received bit is in error and show how you can use the parity check matrix to detect the error.

(d) Repeat (c) if we have the third and fifth bits in error.

Problem 8:

(a) For the rate ½ convolutional code shown in Fig. 4.11 find the output sequence if the input sequence is 11001.

(b) Assuming third and fifth received bits are in error and show how the trellis diagram can detect them.

Problem 9:

Determine, sketch, and label the ACF of the LFSR code with generation matrix $G(x) = x^3 + x + 1$.

Problem 10:

Give all the 8 × 8 Walsh codes and show the first two codes are orthogonal.

Problem 11:

The signal constellation of a voice band modem using QAM modulation with trellis coded modulation has 128 points.

(a) If the symbol transmission rate of the modem was 1800 symbol per second what would be the data transmission rate in bits per second.

(b) If the symbol transmission rate of the modem was 2400 symbol per second what would be the data transmission rate in bits per second.

Problem 12:

(a) If we use 5 × 5 PAM constellation over two TP lines with bandwidth of 100 MHz what would be the effective bit rate of the modem?

(b) If we were using a 32-QAM constellation with trellis coded modulation, instead of 5 × 5 PAM, what would be the effective bit rate of the modem?

(c) Assuming that 32-QAM with trellis coded modulation has 3 dB gain over the 16-QAM, what would be the difference in power requirement between 5 × 5 PAM and this modulation.

Problem 13:

In the IEEE 802.11g OFDM modulation the symbol transmission rate for each carrier is 250 KSps, and 48 of the 64 carriers are used for data transmission. The 16-QAM modulation option of this standard uses rate $^3/_4$ convolutional code. Determine the user data rates when this modulation/coding is used by the modem.

Problem 14:

This problem illustrates the concept of using orthogonal waveforms in CDMA. Figure P4.1 shows the waveforms used by three users: Al, Bill, and George to send messages from a

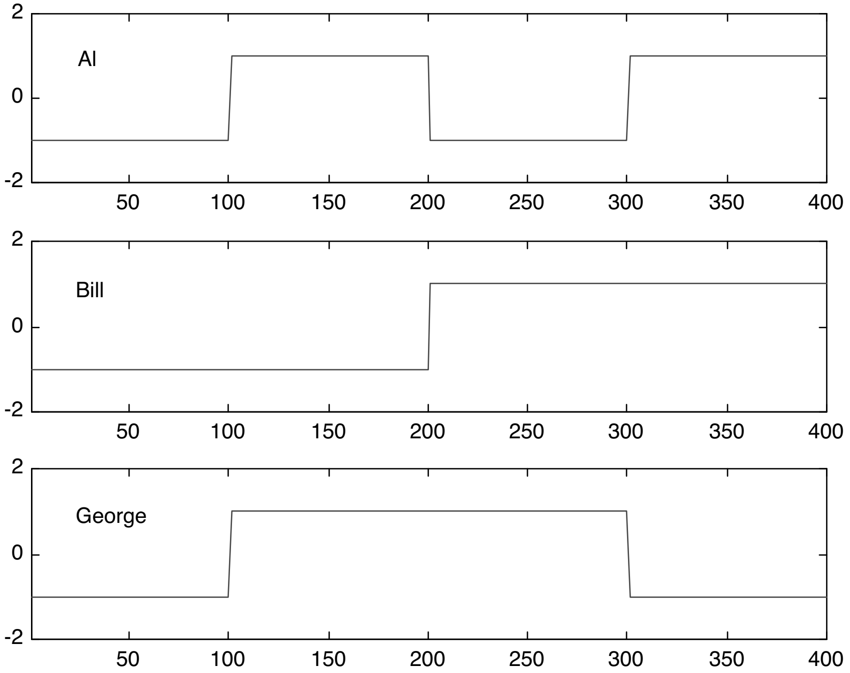


FIGURE P4.1 Orthogonal waveforms transmitted by three users.

wireless text transmitter device to a base station. The chip duration is 100 ns and the bit duration is 400 ns. Each of them is transmitting eight bits each, corresponding to the ASCII codes for A, B and G respectively. They transmit their waveform if the ASCII bit is a zero, and the negative of the waveform if it is a one.

(a) Draw the signals corresponding to their ASCII code transmitted by each of them.

(b) Draw the composite transmitted signal that will be the sum of the individual signals if they are transmitting at the same time and are synchronized perfectly.

(c) Let us refer to the transmitted signal in (b). We shall call the part of the signal during the durations 0–400 ns, 400 ns–800 ns, etc. as "symbols". Compute the cross-correlation at zero lag of each of the symbols in the transmitted waveform with the waveforms of Al, Bill and George. Comment on the results.

Problem 15:

Suppose you are asked to design a 16.8-kbit/sec constellation for a wireless voiceband modem. The symbol rate is 2400 symbols/sec and the received signal-to-noise ratio is 28 dB.

(a) For an uncoded constellation, find the alphabet size and select a constellation.

(b) Give your reasoning for the selection. Using the asymptotic bound, give the symbol error rate for the modem.

(c) Repeat part (a) for a trellis-coded constellation with 3-dB coding gain.

Problem 16:

The maximum length of the cable for a LAN is 2,500 m and the data rate is 10 Mbps. If we implement a stop-and-wait protocol on the LLC and the propagation speed on the cable is 200,000 Km/s

(a) What the maximum propagation time for a packet?
(b) What is the minimum PDU size that guarantees a minimum efficiency of 70%?

Problem 17:

(a) The multi-carrier modem of the IEEE 802.11a uses 48 carriers for data transmission. If each carrier uses 16-QAM and the transmission rate of each carrier is 250 Ksymbols/sec, what is the overall transmission rate of the modem in bits/second?
(b) If a rate ½ coding was applied to the data what would be the overall data rate in bits/second?

Problem 18:

In Matlab, the function `conv` performs the convolution of two vectors. Samples of a signal can be represented as vectors (as in the case of spread spectrum pulses). Suppose that the M-sequence of problem 1 is used as the basic waveform for transmitting a zero and its negative is used to transmit a one. The matched filter will have the flipped version of the M-sequence as its impulse response, i.e. [1 −1 −1 −1 −1 1 1 1 −1 1 1 1 −1 −1 1]. You convolve the input to the matched filter with this vector to get the output. Let us suppose we are transmitting four bits 0, 1, 1, 0. Assume also that the channel is a three-path channel with inter-path delays of $5T_c$ and $8T_c$ respectively. Plot the output of the matched filter. What will be the output if you are using NRZ with a matched filter? (Note that in this case, you need to replace the M-sequence by all ones. What will the MF impulse response be?).

Problem 19:

Block interleaving is a solution to enable simple error correcting codes to correct long bursts of errors. For both the situations of Problem 11 in Chapter 3, determine the number of codewords over which interleaving has to be performed if the length of the codeword is 7 bits and a single bit error can be corrected.

Problem 20:

If codewords have to be received in sequence for message delivery, determine the delay encountered by the block interleaving scheme of Problem 20. How does this impact voice transmission?

5

MEDIUM ACCESS METHODS

5.1 INTRODUCTION

This chapter presents an overview of the medium access methods commonly used in networks. Access methods form a part of layer 2 of the TCP/IP protocol stack and layer 3 of the IEEE 802 standard for LANs that is responsible for interacting with the medium to coordinate successful operation of multiple terminals over a shared channel. Most multiple access methods were originally developed for wired networks and later on adapted to the wireless medium. However, requirements on the wired and wireless media are different, thereby demanding modifications in the original protocols to make them suitable for the wireless medium. Today, the main differences between wireless and wired channels are availability of bandwidth and reliability of transmissions. The wired medium includes optical media with enormous bandwidth and very reliable transmission (with error rates very close to zero all the time). Bandwidth in wireless systems is always limited because the medium (air) cannot be duplicated and also the medium is shared between all wireless systems that include multichannel broadcast television, and a number of other bandwidth demanding applications and services. In the case of wired operation, we can always lay

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

additional cables to increase the capacity as needed, even if it is an expensive proposition. In a wireless environment, we can reduce the size of *cells* to increase capacity, as will be discussed in Chapter 7. With the reduction of the size of the cells, the number of cells increases; and with this, the need for improvements in the wired infrastructure to connect these cells increases. Also, the complexity of the network for handling additional handoffs and mobility management increases, posing a practical limitation upon the maximum capacity of the network. As far as transmission reliability is concerned, as we saw in Chapter 2, the wireless medium suffers from multipath and fading that causes a serious threat to reliable data transmission over the communication link. Since the wireless channel is so unreliable, as discussed in Chapters 3 and 4, people have developed a number of signal processing and coding techniques to improve transmission reliability over the wireless channel. In spite of these techniques, the reliability of the wireless medium is below that of the wired medium used as the backbone of the wireless networks.

Although in practice we prefer to have the same access method and the same frame structure for the wired backbone and wireless access, wireless networks often use different packet sizes and a modified access method to optimize the performance to the specifics of the unreliable wireless medium.

Example 5.1: IEEE 802.3 and IEEE 802.11 The IEEE 802.3 standard (based on Ethernet) is the successful and dominant standard for local wired communications (see Chapter 8). Consequently, the IEEE 802.11 WLAN standard (see Chapter 9), under ideal circumstances, desired the use of a similar access method when its development started in the early 1990s. CSMA with collision detection (CSMA/CD) is the protocol used in the Ethernet. However, collision detection in wireless channels is difficult for several reasons, and the IEEE 802.11 standard had to resort to CSMA/CA, which can be viewed as a wireless adaptation of IEEE 802.3.

Example 5.2: ATM and Wireless ATM In the mid–late 1990s ATM was perceived to be the transmission scheme for all future networking. In the mid 1990s, when wireless solutions were considered, a wireless ATM working group was formed to extend the ATM short packet solution with QoS naturally for wireless access. The group had to make significant changes to the wireless version because ATM was designed for error free and reliable transmission over optical channels.

To avoid substantial overlap with existing literature, we use as examples the access methods used in wireless networks with justification of why and how they are employed in different wireless networks. We allude to wired networks where appropriate. Ethernet, the major medium access scheme for wired networks, is discussed briefly in this chapter and in Chapter 8 in more detail.

The access methods adopted by voice-oriented and data-oriented networks were traditionally quite different. Voice-oriented networks were designed for relatively long telephone conversations as the major application, exchanging several megabytes of information in a *full duplex*[1] mode. A signaling channel that exchanges short messages between two

[1] Duplexing refers to how communication is possible between two parties in both directions. In simplex communications, the information flow is unidirectional, in half-duplex communications it is bidirectional, but only in one direction at a time (time-division duplexing – TDD), and the communication is bidirectional simultaneously in full duplex. For full duplex operation, two separate physical channels are usually required, like two frequency bands (frequency-division duplexing – FDD) or two separate cables.

calling components sets up the call by obtaining resources (such as the link, switches, etc.) in the telephone network at the beginning of the conversation and terminates these arrangements by releasing the resources at the end of the call. The wireless access methods evolved for interaction with these networks assign a slot of time, a portion of frequency, or a specific code to a user, preferably for the entire length of the conversation. We refer to these techniques as centralized assigned access methods. Data networks were originally designed for bursts of data for which the supporting network does not have a separate signaling channel. In packet communications, each packet carries some "signaling information" related to the address of the destination and the source. We refer to the access methods used in these networks as random access methods accommodating randomly arriving packets of data. Certain local area data networks also *take turns* in accessing the medium, as in the case of token passing and polling schemes. In some other cases, the random access mechanisms are used to temporarily *reserve* the medium for transmitting the packet. In recent years, the use of VoIP for telephone conversations has blurred the distinction between voice-oriented and data-oriented networks. However, differentiation between the types of access scheme, namely assigned access and random access, still exists.

In the next two sections of this chapter we provide a short description of assigned access and random access methods.

5.2 CENTRALIZED ASSIGNED ACCESS SCHEMES

Circuit-switched networks, as well as legacy voice-oriented wireless networks, such as cellular telephony or PCS services, use assigned access schemes or channel partitioning techniques. In assigned access schemes, a fixed allocation of channel resources, frequency, time, or a spread-spectrum code is made available on a predetermined basis to a single user for the duration of the communication session. The three basic assigned multiple access methods are FDMA, TDMA, and CDMA. Circuit-switched wired telephone networks started with FDMA and evolved into using TDMA in the mid-twentieth century. The choice of the multiple access method has a great impact on the capacity and quality of the service provided by a cellular telephone network. The impact of multiple access schemes is so important in this case that we commonly refer to various voice-oriented wireless systems by their channel access method, which is only a part of the layer 2 specification of the air interface of the network.

Example 5.3: Common Terminology for Digital Cellular Systems The GSM and the North American IS-136 digital cellular standards are commonly referred to as digital TDMA cellular systems and the IS-95/IMT-2000 are called digital CDMA cellular systems.

In reality, different systems use different modulation techniques as well. However, as we will see in the rest of this book, the impact of the choice of access method on the capacity and overall performance of the network is much more profound in cellular wireless systems. Consequently, the system is really distinguished by its access method. As we will see in our examples of cellular networks, a network that is identified with an access technique often uses other random or fixed assignment techniques as a part of its overall operation. However, it is identified by the access techniques employed for transferring the main information source for which the network is designed to carry.

Example 5.4: Random Access Techniques in Cellular Networks GSM uses slotted-ALOHA (a random access method) to establish a link between the mobile terminal and the BS. It also has an optional frequency-hopping pattern that improves the system performance when there is fading of the radio signal. However, the GSM network is built for voice communications and each session uses TDMA as the access method.

Another important design parameter related to the access method is the differentiation between the carrier frequencies of the forward (downlink – communication between the BS and mobile terminals) and reverse (uplink – communication between the mobile terminal and the BS) channels. If both forward and reverse channels use the same frequency band for communications, but the forward and reverse channels employ alternating time slots, then the system is referred to as employing TDD. If the forward and reverse channels use different carrier frequencies that are sufficiently separated, then the duplexing scheme is referred to as FDD. With TDD, since only one frequency carrier is needed for duplex operation, we can share more of the RF circuitry between the forward and the reverse channels. The reciprocity of the channel in TDD allows for exact open-loop power control and simultaneous synchronization of the forward and reverse channels. TDD techniques are used in systems intended for low-power local-area communications where interference must be carefully controlled and where low complexity and low power consumption are very important. Thus, TDD systems are often used in local-area pico- or micro-cellular systems deployed by PCS networks. FDD is mostly used in macrocellular systems designed for coverage of several tens of kilometers where implementation of TDD is more challenging (see Fig. 5.1).

5.2.1 Frequency-Division Multiple Access

In an FDMA environment all users can transmit signals simultaneously and they are separated from one another by their frequency of operation. The FDMA technique is built upon FDM, which is the oldest and still a commonly used multiplexing technique in the trunks connecting switches in the PSTN. It is also the choice of radio and TV broadcast, as well as of cable TV distribution. FDM is more suitable for analog technology, since it is easier to implement. When FDM is used for channel access it is referred to as FDMA.

Example 5.5: FDMA in AMPS with FDD Figure 5.2*a* shows the FDMA/FDD system commonly used in 1G analog cellular systems and a number of early cordless telephones. In FDMA/FDD systems, forward and reverse channels use different carrier frequencies and a fixed sub-channel pair is assigned to a user terminal during the communication session. At the receiving end, the mobile terminal filters the designated channel out of the composite signal. The early analog cellular telephone system called AMPS allocated 30 kHz of bandwidth for each of the forward and reverse channels. The result is a total of 421 channels in 25 MHz of spectrum assigned to each direction –395 of these channels were used for voice traffic and the rest for signaling.

Example 5.6: FDMA in CT-2 with TDD Fig. 5.2*b* shows an FDMA/TDD system used in the CT-2 digital cordless telephony standard. Each user employs a single carrier frequency for all communications. The forward and reverse transmissions take turns via alternating time slots. This system was designed for distances of up to 100 m and a voice conversation is based on 32 kb/s adaptive differential PCM (ADPCM) voice coding. The total allocated bandwidth for CT-2 is 4 MHz supporting 40 carriers each using 100 kHz of bandwidth.

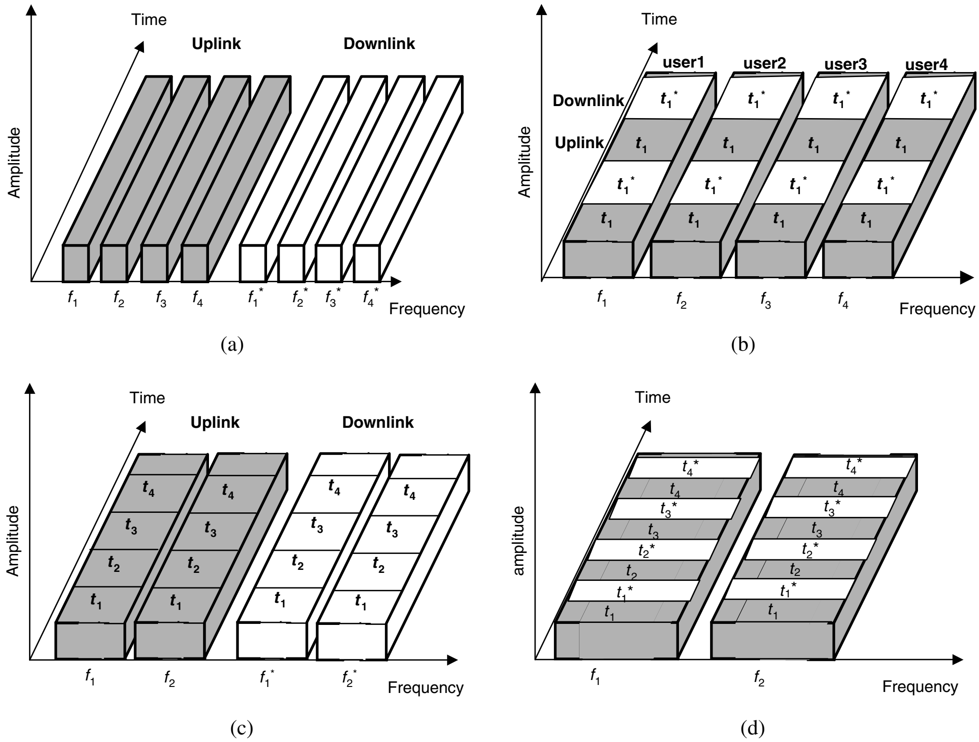


FIGURE 5.1 (*a*) FDMA/FDD; (*b*) FDMA/TDD; (*c*) TDMA/FDD with multiple carriers; (*d*) TDMA/TDD with multiple carriers.

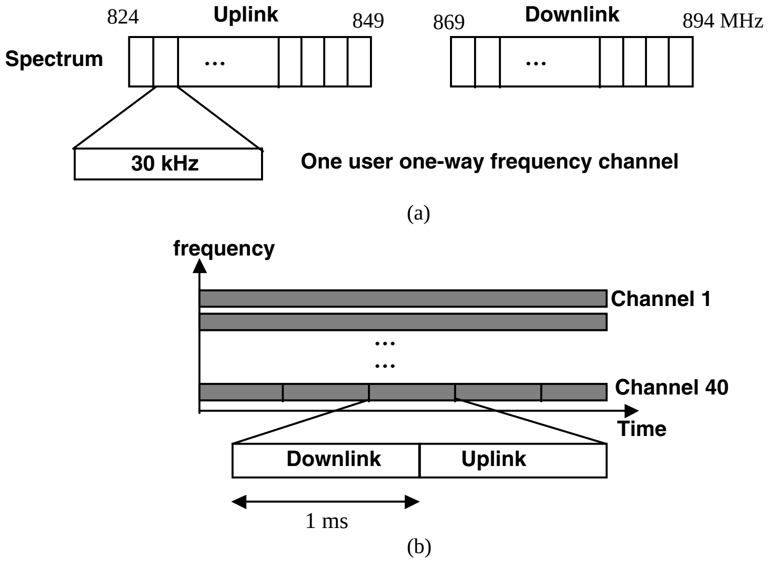


FIGURE 5.2 (*a*) FDMA/FDD in AMPS and (*b*) FDMA/TDD in CT-2.

The designer of an FDMA system must pay special attention to adjacent channel interference, particularly in the reverse channel. In both forward and reverse channels, the signal transmitted must be kept confined within its assigned band, at least to the extent that the out-of-band energy causes negligible interference to users employing adjacent channels. Operation of the forward channel in wireless FDMA networks is very similar to wired FDM networks. In forward wireless channels, in a manner similar to that of wired FDM systems, the signal received by all mobile terminals has the same received power and interference is controlled by adjusting the sharpness of the transmitter and receiver filters for the separate carrier frequencies. The problem of adjacent channel interference is much more challenging on the reverse channel. On the reverse channel, mobile terminals will be operating at different distances from the BS. The RSS at the BS of a signal transmitted by a mobile terminal close to the BS and the RSS at the BS of a transmission by a mobile terminal at the edges of the cell are often substantially different, causing problems in detecting the weaker signal. This problem is usually referred to as the near–far problem. If the out-of-band emissions are large, then they may swamp the actual information-carrying signal.

Example 5.7: Near–Far Problem

1. What is the difference between the RSS of two terminals located 10 m and 1 km from a BS in an open area?
2. Explain the effects of shadow fading on the difference in the RSSs.
3. What would be the impact if the two terminals were operating in two adjacent channels? Assume out-of-band radiation that is 40 dB below the main lobe.

Solution

1. As we saw in Chapter 2, the RSS falls by around 40 dB per decade of distance in open areas. Therefore, the received powers from a mobile terminal that is 10 m from a BS and another that is at a distance of 1 km are 80 dB apart.
2. In addition to the fall of the RSS with distance, we also discussed the issues of multipath and shadow fading in radio channels that cause power fluctuations on order of several tens of decibels. Therefore, the difference in the received powers due to the near–far problem may exceed even 100 dB.
3. If the out-of-band emission is only 40 dB below that of the transmitted power, then it may exceed the strength of the information-bearing signal by almost 60 dB.

To handle the near–far problem, FDMA cellular systems adopt two different measures. First, as we discuss in Chapter 7, when frequencies are assigned to a cell, they are grouped such that the frequencies in each cell are as far apart as possible. The second measure employed is power control, which will also be discussed briefly in Chapter 7. In addition, whenever FDMA is employed, *guard bands*[2] are also used in between frequency channels to reduce adjacent channel interference further. However, this has the effect of reducing the overall spectrum efficiency.

[2] When we say each AMPS channel is 30 kHz wide in Example 5.5, this also includes guard bands.

5.2.2 Time-Division Multiple Access

In TDMA systems, a number of users share the same frequency band by taking assigned turns in using the channel. The TDMA technique is built upon the TDM scheme commonly used in the trunks for telephones systems. The major advantage of TDMA over FDMA is its format flexibility. Because the format is completely digital and provides flexibility of buffering and multiplexing functions, time-slot assignments among multiple users are readily adjustable to provide different access rates for different users. This feature is particularly adopted in the PSTN, and the TDM scheme forms the backbone of all digital connections in the heart of the PSTN. The hierarchy of digital transmission trunks used in North America is the so-called T-carrier system that has an equivalent European system (the E-carriers) approved by the ITU. In the hierarchy of digital transmission rates standardized throughout North America, the basic building block is the 1.544 Mb/s link known as the T-1 carrier. A T-1 transmission frame is formed by TDM 24 PCM-encoded voice channels each carrying 64 kb/s of users' data. Service providers often lease T-carriers to interconnect their own switches and routers and for forming their own networks.

Example 5.8: The Use of T-Carriers in Cellular Networks Cellular networks often lease T-carriers from the long-haul telephone companies to interconnect their own switches referred to as mobile switching centers (MSCs). The difference between the MSC and a regular switch in the PSTN is that the MSC can support mobility of the terminal. The details of these differences are discussed in later chapters, when we provide examples of cellular networks. The end user subscribes to the cellular service provider.

Example 5.9: The Use of T-1 Lines in the Internet The routers in the Internet are sometimes connected through leased T-carrier telephone lines to form part of the Internet. The difference between a router and a PSTN switch is that the router can handle packet switching whereas the PSTN switch uses circuit switching. The end user subscribes to an Internet service provider (ISP) in this case.

With TDMA, a transmit controller assigns time slots to users, and an assigned time slot is held by a user until the user releases it. At the receiving end, a receiver station synchronizes to the TDMA signal frame and extracts the time slot designated for that user. The heart of this operation is synchronization that was not needed for FDMA systems. The TDMA concept was developed in the 1960s for use in digital satellite communication systems and first became operational commercially in telephone networks in the mid 1970s [Pah95].

In cellular and cordless systems, the migration to TDMA from FDMA took place in the 2G systems. The first cellular standard adopting TDMA was GSM. The GSM standard was initiated to support international roaming among Scandinavian countries in particular and the rest of Europe in general. The digital voice adoption in the TDMA format facilitated the network implementation, resulted in improvements in the quality of the voice, and provided a flexible format to integrate data services in the cellular network. The FDMA systems in the USA very quickly observed a capacity crunch in major cities; among the options for increasing capacity, TDMA was adopted initially through the IS-54 system, which was later replaced by IS-136. TDMA was adopted in 2G cordless telephones, such as CT-2 and digital enhanced cordless telephony (DECT), to provide format flexibility and to allow more compact and low-power terminals.

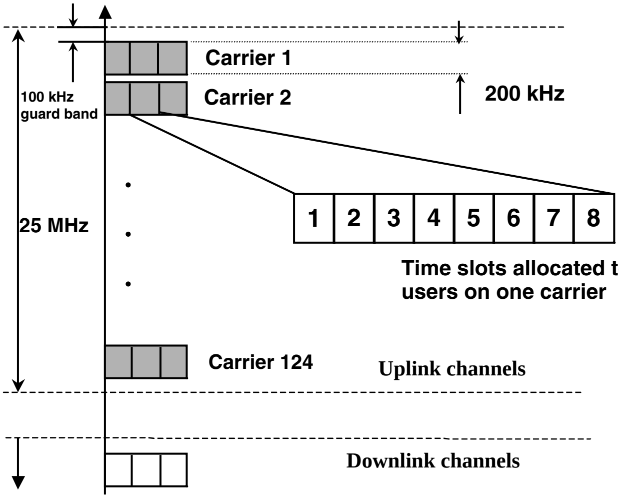


FIGURE 5.3 FDMA/TDMA/FDD in GSM.

Example 5.10: TDMA in GSM Figure 5.3 shows an FDMA/TDMA/FDD channel used in 2G digital cellular in Europe (GSM). The particular example shows the eight-slot TDMA scheme used in the GSM system. Forward and reverse channels use separate carrier frequencies (FDD). Each carrier can support up to eight simultaneous users via TDMA, each using 13 kb/s encoded digital speech, within a 200 KHz carrier bandwidth. A total of 124 frequency carriers (FDMA) are available in the 25 MHz allocated band in each direction. 100 kHz of band is allocated as a guard band at each edge of the overall allocated band.

Example 5.11: TDMA in DECT Figure 5.4 shows an FDMA/TDMA/TDD system used in the pan-European digital PCS standard DECT. Since distances are short, a TDD format allows use of the same frequency for forward and reverse operations. The bandwidth per

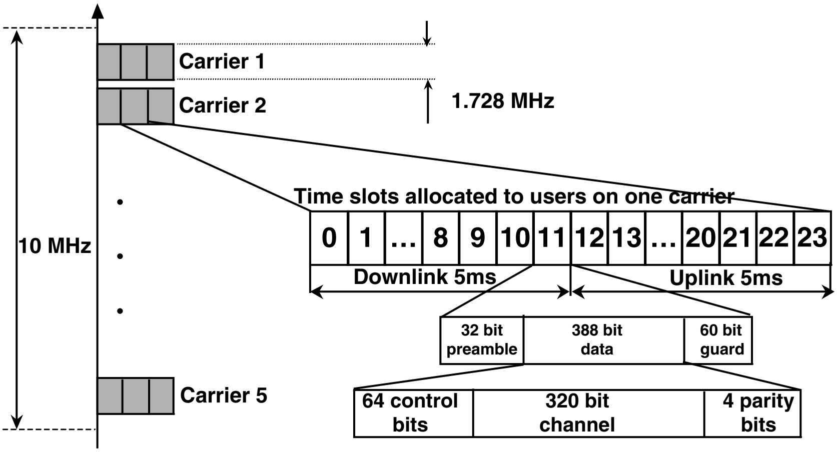


FIGURE 5.4 FDMA/TDMA/TDD in DECT.

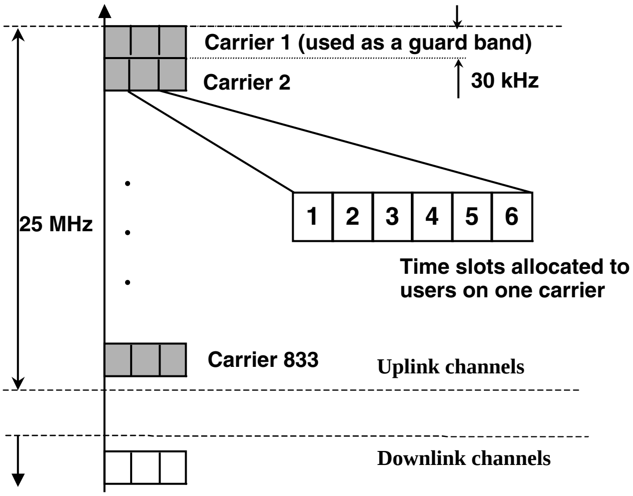

Carrier 1 (used as a guard band)
Carrier 2
30 kHz
25 MHz
1
2
3
4
5
6
Time slots allocated to users on one carrier
Carrier 833
Uplink channels
Downlink channels

FIGURE 5.5 FDMA/TDMA/FDD in IS-136 standard.

carrier is 1.728 MHz, which can support up 12 ADPCM-coded speech channels via TDMA. The total allocated band in Europe is 10 MHz, which can support five carriers (FDMA). Figure 5.4 shows the details of the TDMA/TDD time slots use in the DECT system. The frame duration is 10 ms, with 5 ms for portable-to-fixed station and 5 ms for fixed-to-portable. The transmitter transfers information in signal bursts which it transmits in slots of duration 10/24 = 0.417 ms. With 480 bits per slot (including a 60-bit guard time), the total bit rate is 1.152 Mb/s. Each slot contains 64 bits for system control (C, P, Q and M channels) and 320 bits for user information (I channel).

Example 5.12: TDMA in IS-136 Figure 5.5 shows the frame format for the TDMA/FDD with six slots considered for IS-136 both for the forward (base to mobile) and reverse (mobile to base) channels. In IS-136, each 30 kHz digital channel has a channel transmission rate of 48.6 kb/s. The 48.6 kb/s stream is divided into six TDMA channels of 8.1 kb/s each. The IS-136 slot and frame format, shown in Fig. 5.5, is much simpler than that of the GSM standard. The 40 ms frame is composed of six 6.67 ms time slots. Each slot contains 324 bits, including 260 bits of user data, and 12 bits of system control information in a slow associated control channel (SACCH). There is also a 28-bit synchronization sequence and a 12-bit digital verification color code (DVCC) used to identify the frequency channel to which the mobile terminal is tuned. In the mobile-to-base direction, the slot also contains a guard time interval of 6-bit duration, when no signal is transmitted, and a 6-bit ramp interval to allow the transmitter to reach its full output power level.

Owing to the near–far problem, the received signal on the reverse channel from a user occupying a time slot can be much larger than the received power from the terminal using the adjacent time slot. In such a case, the receiver will have difficulty in distinguishing the weaker signal from the background noise. In a manner similar to FDMA systems, TDMA systems also use power control to handle this near–far problem.

Capacity of Time-/Frequency-Division Multiple Access Cellular Systems. The capacity of a cellular system in the case of TDMA or FDMA depends primarily on the number of channels available per cell, which in turn depends on the reuse factor (how many cells apart can the same frequencies be reused). Let us suppose there is only one cell and the system employs TDMA. Then, the number of simultaneous users that can be supported will be simply the number of users per carrier n multiplied by the total number of carriers (in the case of FDMA, one user is supported per carrier). The total number of carriers is given by the total bandwidth W divided by the bandwidth of one carrier B. So, the number of simultaneous users supported will be $M = nW/B$. If the frequency reuse factor is N_f, then the number of simultaneous users per cell will be $M = n(W/N_f)/B$. We will revisit this in our discussion of CDMA next.

5.2.3 Code-Division Multiple Access

With the growing interest in the integration of voice, data, and video traffic in telecommunication networks, CDMA appears increasingly attractive as the wireless access method of choice. Fundamentally, integration of various types of traffic is readily accomplished in a CDMA environment, since coexistence in such an environment does not require any specific coordination among user terminals. In principle, CDMA can accommodate various wireless users with different bandwidth requirements, switching methods, and technical characteristics without any need for coordination. Of course, since each user signal contributes to the interference seen by other users, power control techniques are essential in the efficient operation of a CDMA system.

To illustrate CDMA and how it is related to FDMA and TDMA, it is useful to think of the available band and time as resources we use to share among multiple users. In FDMA, the frequency band is divided into slots and each user occupies that frequency throughout the communication session. In TDMA, a larger frequency band is shared among the terminals and each user uses a slot of time during the communication session. As shown in Fig. 5.6, in a CDMA environment, multiple users use the same band at the same time and the users are differentiated by a code that acts as the key to identify those users. These codes are selected so that when they are used at the same time in the same band a receiver knowing the code of a particular user can detect that user among all the received signals. In CDMA/FDD (Fig. 5.7*a*) the forward and reverse channels use different carrier frequencies. If both

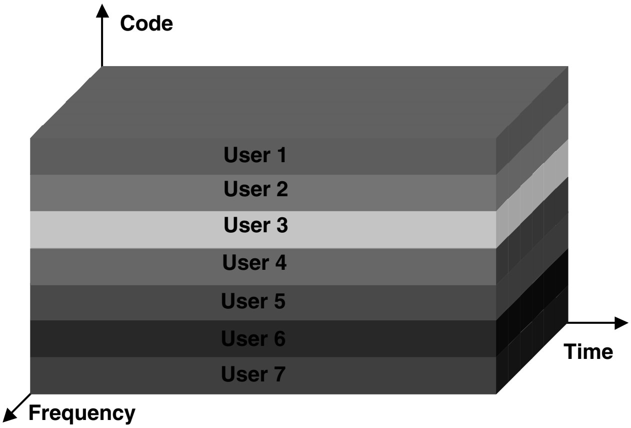


FIGURE 5.6 Simple illustration of CDMA.

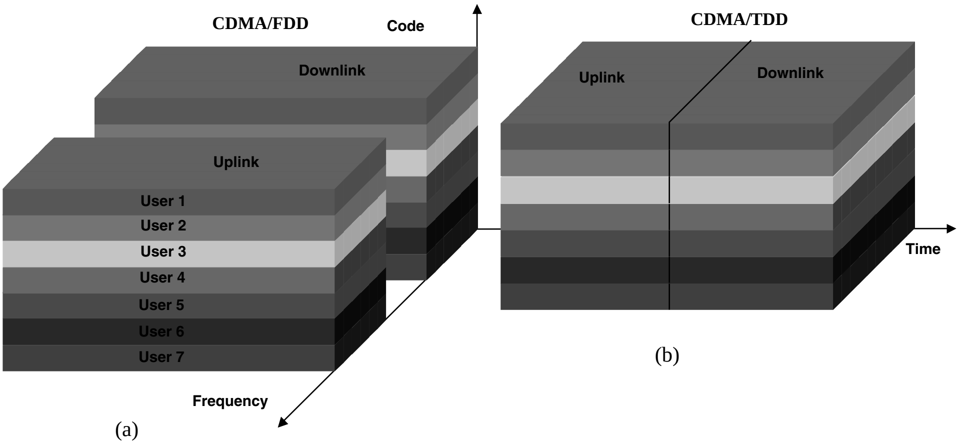


FIGURE 5.7 (*a*) FDD and (*b*) TDD with CDMA.

transmitter and receiver use the same carrier frequency (Fig. 5.7*b*), the system is CDMA/TDD.

In CDMA, each user is a source of noise to the receiver of other users, and if we increase the number of users beyond a certain value, then the entire system collapses because the signal received in each specific receiver will be buried under the noise caused by many other users. An important question is how many users can simultaneously use a CDMA system before the system collapses. We investigate the answer to this below.

Capacity of Code-Division Multiple Access. CDMA systems are implemented based on spread-spectrum technology that was presented in Chapter 3. In its most simplified form, a spread-spectrum transmitter spreads the signal power over a spectrum N times wider than the spectrum of the message signal. In other words, an information bandwidth of R_b occupies a transmission bandwidth of W, where

$$W = N_p R_b \tag{5.1}$$

The spread-spectrum receiver processes the received signal with a *processing gain* of N_p. This means that, during the processing at the receiver, the power of the received signal having the code of that particular receiver will be increased N times beyond the value before processing.

Let us consider the situation of a *single cell* in a cellular system employing CDMA. Assume that we have M simultaneous users on the reverse channel of a CDMA network. Further, let us assume that we have an ideal power control enforced on the channel so that the received power of signals from all terminals has the same value P. Then, the received power from the target user after processing at the receiver is NP and the received interference from $M - 1$ other terminals is $(M - 1)P$. If we also assume that a cellular system is interference limited and the background noise is dominated by the interference noise from other users, then the received SNR for the target receiver will be

$$S_r = \frac{N_p P}{(M-1)P} = \frac{N_p}{M-1} \tag{5.2}$$

All users always have a requirement for the acceptable error rate of the received data stream. For a given modulation and coding specification of the system, that error rate requirement will be supported by a minimum S_r requirement that can be used in Eq. (5.2) to solve for the number of simultaneous users. Then, solving Eqs (5.1) and (5.2) for M, we have

$$M = \frac{W}{R_b}\frac{1}{S_r} + 1 \cong \frac{W}{R_b}\frac{1}{S_r} \tag{5.3}$$

Example 5.13: Capacity of One Carrier in a Single-Cell CDMA System Using QPSK modulation and convolutional coding, the IS-95 digital cellular systems require $3 < S_r < 9$ dB. The bandwidth of the channel is 1.25 MHz and transmission rate is $R = 9600$ bits/s. Find the capacity of a single IS-95 cell.

Solution Using Eq. (5.3) we can support from

$$M = \frac{1.25\text{MHz}}{9600\text{ bps}}\frac{1}{8} \approx 16 \text{ up to } M = \frac{1.25\text{MHz}}{9600\text{ bps}}\frac{1}{2} \approx 65 \text{ users}$$

Practical Considerations. In the practical design of digital cellular systems, in addition to the bandwidth efficiency of the system, three other parameters affect the number of users that can be supported by the system. These are the number of sectors in each BS antenna, the voice activity factor, and the interference increase factor. These parameters are quantified as factors used in the calculation of the number of simultaneous users that the CDMA system can support. The use of sectored antennas is an important factor in maximizing bandwidth efficiency. Cell sectorization using directional antennas reduces the overall interference, increasing the allowable number of simultaneous users by a s*ectorization gain factor*, which we denote by G_A. With ideal sectorization the users in one sector of a BS antenna do not interfere with the users operating in other sectors, and $G_A = N_{sec}$, where N_{sec} is the number of sectors in the cell. In practice, antenna patterns cannot be designed to have ideal characteristics and, owing to multipath reflections, users in general communicate with more than one sector. Three-sector BS antennas are commonly used in cellular systems, and a typical value of the sectorization gain factor is $G_A = 2.5$ (4 dB). The voice activity interference reduction factor G_v is the ratio of the total connection time to the active talkspurt time. On average, in a two-way conversation, each user talks roughly 50% of the time. The short pauses in the flow of natural speech reduce the activity factor further to about 40% of the connection time in each direction. As a result, the typical number used for G_v is 2.5 (4 dB). The interference increase factor H_0 accounts for users in other cells in the CDMA system. Since all neighboring cells in a CDMA cellular network operate at the same frequency they will cause additional interference. This interference is relatively small due to the processing gain of the system and the distances involved; a value of $H_0 = 1.6$ (2 dB) is commonly used in industry.

Incorporating these three factors as a correction to Eq. (5.3), the number of simultaneous users that can be supported in a CDMA cell can be approximated by

$$M = \frac{W}{R_b}\frac{1}{S_r} + 1 \cong \frac{W}{R_b}\frac{1}{S_r}\frac{G_A G_v}{H_0} \tag{5.4}$$

If we define the *performance improvement factor* in a digital cellular system as

$$K_{\mathrm{p}} = \frac{G_{\mathrm{A}} G_{\mathrm{v}}}{H_0} \tag{5.5}$$

then, assuming the typical parameter values given earlier, the performance improvement factor is $K_{\mathrm{p}} = 4$ (6 dB).

Example 5.14: Capacity of One Carrier in a Multicell CDMA System with Correction Factors Determine the multicell IS-95 CDMA capacity with correction for sectorization and voice activity. Use the numbers from Example 5.13.

Solution If we continue the previous example with the new correction factor included, then the range for the number of simultaneous users becomes $64 < M < 260$.

5.2.4 Comparison of Code-, Time-, and Frequency-Division Multiple Access

CDMA was by far the most successful multiple access scheme in 2G cellular wireless systems in the USA. With wideband CDMA adopted as the multiple access scheme of choice in 3G cellular networks, one wonders why CDMA has become the favorite choice for wireless access in voice-oriented networks. Spread-spectrum technology became the favorite technology for military applications because of its capability to provide a low probability of interception and strong resistance to interference from jamming. In the cellular industry, CDMA was introduced as an alternative to TDMA to improve the capacity of 2G cellular systems in the USA. As a result, much of the early debates in this area were focused on calculation of the capacity of CDMA as it is compared with TDMA. However, capacity is not the only reason for the success of the CDMA technology. As a matter of fact, calculation of the capacity of CDMA using the simple approach provided above is *not* very conclusive and is subject to a number of assumptions, such as perfect power control, that cannot be practically met. The first CDMA service providers in the USA were using slogans such as "you can not believe your ears!" to address the superior quality of voice for the CDMA. However, the superiority of voice is partially dependent on the speech coder and it is not a CDMA versus TDMA issue. In order to provide a good explanation for the success of a complex and multidisciplinary technology, such as a cellular network, addressing consumer market issues has always been very important. Those of us involved in this debate for the past decade have seen the discussion of the ups and downs of CDMA in a variety of forums. One of the most interesting events that the principal author remembers was in 1997 in a major wireless conference in Taipei where one of the most famous figures in this debate in his keynote speech at the opening of the conference declared that "we have seen in the past that the VHS which was not a better technology defeated BETA." In his perception, at that time, CDMA was similar to BETA. In less than a year or so after that, CDMA was selected by a number of different communities around the world as the technology of choice for 3G and IMT-2000.

In the rest of this section we bring out a number of issues that may enlighten the reader towards a deeper understanding of the technical aspects of CDMA systems as they are compared with TDMA and FDMA networks. We hope that this may lead the reader to their own conclusion about the success of CDMA.

Format Flexibility. As we discussed before, telephone voice was the dominant source of income for the telecommunication industry up to the end of the last century. In the new millennium, the strong emergence of the Internet and cable TV industries has created a case for other popular multimedia applications. The cellular phones that were designed for telephony applications are now being used for other applications and need support for multimedia applications. To support a variety of data rates with different requirements, a network needs format flexibility. As we discussed earlier, one of the reasons for migrating from analog FDMA to digital TDMA was that TDMA provides a more flexible environment for integration of voice and data. The time slots of a TDMA network designed for voice transmission can be used individually or in a group format to transmit data from users and to support different data rates. However, all these users should be time synchronized, and the quality of the transmission channel is the same for all of them. The chief advantage of CDMA relative to TDMA is its flexibility in timing and the quality of transmission. In CDMA, users are separated by their codes, unaffected by the transmission time relative to other users. The power of the user can also be adjusted with respect to others to support a certain quality of transmission. In CDMA, each user is far more liberated from the other users, allowing a fertile setting to accommodate different service requirements to support a variety of transmission rates with different quality of transmission to support multimedia or any other emerging application.

Performance in Multipath Fading. As we saw in Chapter 2, multipath in wireless channels causes frequency-selective fading. In frequency-selective fading, when the transmission band of a narrowband system coincides with the location of the fade, no useful signal is received. As we increase the transmission bandwidth, fading will occupy only a portion of the transmission band, providing an opportunity for a wideband receiver to take advantage of the portion of the transmission band not under fade and provide a more reliable communication link. Technologies such as equalization, OFDM, sectored antennas, and spread spectrum can be employed in wideband systems to handle the impact of frequency-selective fading. The wider the bandwidth, the better is the opportunity for averaging out the faded frequency.

These technologies are not used in the first-generation analog cellular FDMA systems because they were analog systems and these techniques are digital. The pan-European GSM digital cellular system uses 200 kHz of band and the standard recommends using DFE. The North American digital cellular system, IS-136, uses digital transmission over the same analog band of 30 kHz of the North American AMPS system and does not recommend equalization because the bandwidth is not very large. An equalizer needs additional circuitry and some power budget at the receiver that was one of the drawbacks considered in IS-136. The bandwidth of IS-95 CDMA system is 1.25 MHz and W-CDMA systems for 3G networks use bandwidths that are as high 10 MHz. RAKE receivers are used to increase the benefits of wideband transmission by taking advantage of the so-called in-band or time diversity of the wideband signal. This is one of the reasons for having a better quality of voice in CDMA systems. As we mentioned earlier, quality of voice is also affected by the robustness of the speech-coding algorithm, coverage of service, methods to handle interference, handoffs, and power control as well.

System Capacity. Comparison of the capacity depends on a number of issues, including the frequency reuse factor, speech coding rate, and the type of antenna. Therefore, a fair comparison would be difficult unless we go to practical systems. The following simple example compares the capacity of FDMA, TDMA, and CDMA used in debates to evaluate alternatives for the 1G/2G North American cellular systems to replace the 1G analog.

Example 5.15: Comparison of the Capacity of Different 2G Systems Compare the capacity of 2G CDMA with 1G FDMA and 2G TDMA systems. For the CDMA system, assume an acceptable signal to interference ratio of 6 dB, data rate of 9600 bits/s, voice duty cycle of 50%, effective antenna separation factor of 2.75 (close to ideal three-sector antenna), and neighboring cell interference factor of 1.67.

Solution For the 2G CDMA system, using Eq. (5.4), for each carrier with $W = 1.25\,\text{MHz}$, $R_b = 9600$ bits/s, $S_r = 4$ (6 dB), $G_v = 2$ (50% voice activity), $G_A = 2.75$, and $H_0 = 1.67$ we have

$$M = \frac{W}{R_b}\frac{1}{S_r}\frac{G_A G_v}{H_0} = 108 \text{ users per cell}$$

For the 2G TDMA system with a carrier bandwidth of $B = 30\,\text{kHz}$, number of users per carrier of $m = 3$, and frequency reuse factor of $N_f = 4$ (commonly used in these systems), each $W = 1.25\,\text{MHz}$ of bandwidth provides for

$$M = \frac{W}{B}\frac{m}{N_f} = 31.25 \text{ users per cell}$$

For the 1G analog system with carrier bandwidth of $B = 30\,\text{kHz}$, number of users per carrier of $m = 1$, and frequency reuse factor of $N_f = 7$ (commonly used in these systems), each $W = 1.25\,\text{MHz}$ of bandwidth provides for

$$M = \frac{W}{B}\frac{1}{N_f} = 6 \text{ users per channel}$$

Another example of this form is instructive to compare these systems with the 2G TDMA system (GSM) that originated in Europe and since then has become a presence in the USA as well. We note that these are simple calculations that provide some insight into the system capacity.

Example 5.16: Comparison of North American Systems with GSM Determine the capacity of GSM for $N_f = 3$.

Solution For the GSM system with a carrier bandwidth of $B = 200\,\text{kHz}$, number of users per carrier of $m = 8$, and frequency reuse factor of $N_f = 3$ (commonly used in these systems), each $W = 1.25\,\text{MHz}$ of bandwidth provides for

$$M = \frac{W}{B}\frac{m}{N_f} = 16.7 \text{ users per cell}$$

Handoff. As we discuss in Chapter 7, handoff occurs when a received signal in a mobile station (MS) becomes weak and another BS can provide a stronger signal to the MS. The 1G FDMA cellular systems often used the so-called hard-decision-handoff in which the BS controller monitors the received signal from the BS and at the appropriate time switches the connection from one BS to another. TDMA systems use the so-called mobile-assisted handoff (MAHO) in which the MS monitors the received signal from available BSs and reports it to the BS controller which then makes a decision on the handoff. Since adjacent cells in both FDMA and TDMA use different frequencies, the MS has to disconnect from and reconnect to the network, which will appear as a click to the user. Handoffs occur at the edge of the cells when the received signals from both BSs are weak. The signals also fluctuate anyway because they are arriving over radio channels. As a result, decision-making for the handoff time is often complex and the user experiences a period of poor signal quality and possibly several clicks during the completion of the handoff process. Since adjacent cells in a CDMA network use the same frequency, a mobile moving from one cell to another can make a "seamless" handoff by the use of signal combining. When the MS approaches the boundary between cells, it communicates with both cells. A controller combines the signals from both links to form a better communication link. When a reliable link has been established with the new BS, the mobile stops communicating with the previous BS and communication is fully established with the new BS. This technique is referred to as soft handoff. Soft handoff provides a dual diversity for the received signal from two links, which improves the quality of reception and eliminates clicking as well as the ping-pong problem.

Power Control. As we discussed earlier in this chapter, power control is necessary for FDMA and TDMA systems to control adjacent channel interference and mitigate the unexpected interference caused by the near–far problem. In FDMA and TDMA systems, some sort of power control is needed to improve the quality of the voice delivered to the user. In CDMA, however, the capacity of the system depends *directly* on the power control, and an accurate power control mechanism is needed for proper operation of the network. With CDMA, power control is the key ingredient in maximizing the number of users that can operate simultaneously in the system. As a result, CDMA systems adjust the transmitted power more often and with smaller adjustment steps to support a more refined control of power. Better power control also saves on the transmission power of the MS, which increases the life of the battery. The more refined power control in CDMA systems also helps in power management of the MS, which is an extremely important practical issue for users of the mobile terminals.

Implementation Complexity. Spread spectrum is a two-layer modulation technique requiring greater circuit complexity than conventional modulation schemes. This, in turn, will lead to higher electronic power consumption and larger weight and cost for mobile terminals. Gradual improvements in battery and integrated circuit technologies, however, have made this issue transparent to the user.

5.2.5 Performance of Assigned Access Methods

Fixed assignment access methods are used with circuit-switched cellular and PCS telephone networks. In these networks, in a manner similar to the wired multichannel environments, the performance of the network is measured by the blockage rate of an initiated call. A call does not go through for two reasons: (1) when the calling number is not available; (2) when the telephone company is out of resources to provide a line for the communication session. In POTS, for both cases the user hears a busy tone signal and cannot distinguish between the two types of blockage. In most cellular systems, however, type (1) blockage results in a response that is a busy tone and type (2) with a message like "all the circuits are busy at this time, please try your call later." In the rest of this book we refer to blockage rate only as type (2) blockage rate. The statistical properties of the traffic offered to the network are also a function of time. The telephone service providers often design their networks so that the blockage rate at the peak traffic is always below a certain percentage. Cellular operators often try to keep this average blockage rate below 2%.

The blockage rate is a function of the number of subscribers, number of initiated calls, and the length of the conversations. In telephone networks, the Erlang equations are used to relate the probability of blockage to the average rate of the arriving calls and the average length of a call. In wired networks, the number of lines or subscribers that can connect to a multichannel switch is a fixed number. The telephone company monitors the statistics of the calls over a long period of time and upgrades the switches with the growth of subscribers so that the blockage rate during peak traffic times remains below the objective value. In cellular telephony and PCS networks, the number of subscribers operating in a cell is also a function of time. Everyone uses their cellular telephones during the day in downtown areas; in the evenings they use them in their residential areas, which are covered by a different cell. Therefore, traffic fluctuations in cellular telephone networks are much more than the traffic fluctuations in POTS. In addition, telephone companies can easily increase the capacity of their networks by increasing their investment on the number of transmission lines and quality of switches supporting network connections. In wireless networks, the overall number of available channels for communications is ultimately limited by the availability of the frequency bands assigned for network operation. To respond to the fluctuations of the traffic and cope with the bandwidth limitations, cellular operators use complex frequency assignment strategies to share the available resources in an optimal manner. Some of these issues are discussed in Chapter 5.

Traffic Engineering Using the Erlang Equations. The Erlang equations are the core of traffic engineering for telephony applications. The two basic equations used for traffic engineering are Erlang B and Erlang C equations. The Erlang B equation relates the probability of blockage $B(N, \rho)$ to the number of channels N_u and the normalized call density in units of channels ρ. The Erlang B formula is

$$B(N_u, \rho) = \frac{\rho^{N_u}/N_u!}{\sum_{i=0}^{N_u}(\rho^i/i!)} \tag{5.6}$$

where $\rho = \lambda/\mu$, λ is the call arrival rate, and μ is the service rate of the calls.[3]

Example 5.17: Call Blocking Using the Erlang B Formula We want to provide a wireless public phone service with five lines to a ferry crossing between Helsinki and Stockholm carrying 100 passengers where, on average, each passenger makes a 3 min telephone call every 2 h. What is the probability of a passenger approaching the telephones and none of the five lines being available?

Solution In practice, the probability of call blockage is often given and we need to calculate the number of subscribers. Here, we need an inverse function for the Erlang equation that is not available. As a result, a number of tables and graphs are available for this inverse mapping. Figure 5.8 shows a graph relating the probability of blockage $B(N_u, \rho)$ to the number of channels N_u and the normalized traffic per available channels ρ. From this graph, we can estimate the blocking probability. The traffic load is 100 users × 1 call/user × 3 min/call per 120 min = 2.5 Erlangs. Since there are five lines available and the traffic is 2.5 Erlangs, the blocking probability is roughly 0.07.

Example 5.18: Capacity Using Erlang B Formula An IS-136 cellular phone provider owns 50 cell sites and 19 traffic carriers per cell each with a bandwidth of 30 kHz. Assuming

[3] The equation assumes that the arrivals are Poisson and the service rate is exponential. For details, see [Ber87].

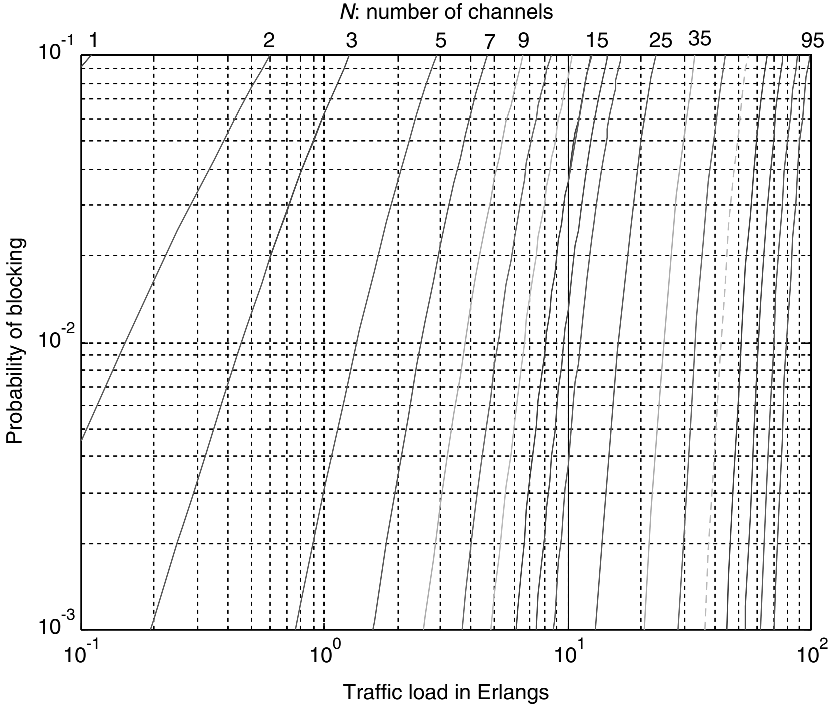


FIGURE 5.8 Erlang B chart showing the blocking probability as a function of offered traffic and number of channels.

each user makes three calls per hour and the average holding time per call of 5 min, determine the total number of subscribers that the service provider can support with a blocking rate of less than 2%.

Solution The total number of channels is $N_u = 19 \times 3 = 57$ per cell. For $B(N_u, \rho) = 0.02$ and $N_u = 57$, Fig. 5.8 shows that $\rho = 45$ Erlangs. With an average of five calls per minute the service rate is $\mu = 1/5$ min, and the acceptable arrival rate of the calls is $\lambda = \rho\mu = 1/5$ (min^{-1}) × 45 (Erlangs) = 9 (Erlangs/min). With an average of three calls per hour the system can accept 9 (Erlangs/min)/3 (Erlangs)/60 (min) = 180 subscribers per cell. Therefore, the total number of subscribers is 180 (subscribers/cell) × 50 (cells) = 8000 subscribers.

The Erlang C formula relates the waiting time in a queue if a call does not go through but it is buffered until a channel is available. These equations start with the probability that a call does not get processed immediately and gets delayed. The probability a call is delayed is given by

$$P(\text{delay} > 0) = \frac{\rho^{N_u}}{\rho^{N_u} + N_u!\left(1 - \frac{\rho}{N_u}\right)\sum_{k=0}^{N_u-1}\frac{\rho^k}{k!}} \tag{5.7}$$

Because of the complexity of the calculation, tables or graphs are again used to provide values for this probability based on normalized values of ρ. Figure 5.9 illustrates the

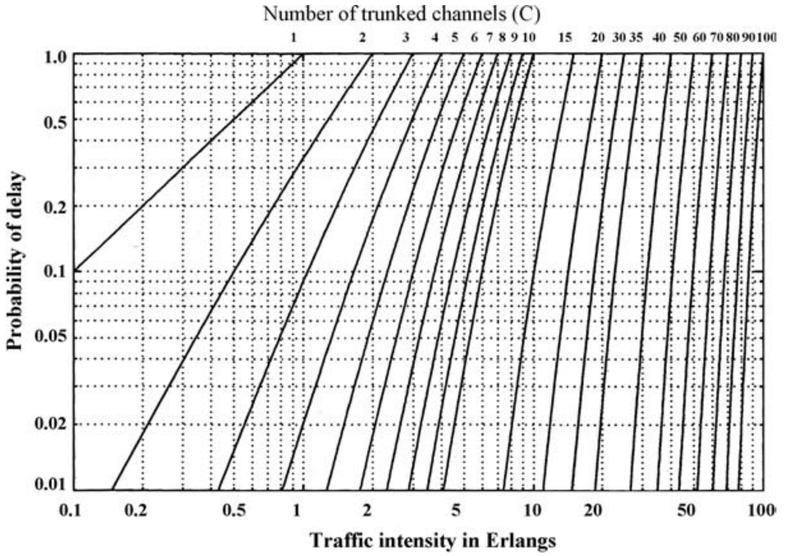


FIGURE 5.9 Erlang C chart relating the offered traffic to the number of channels and the probability of queuing.

relationship between probability of delay, number of channels N_u, and the normalized traffic per available channel ρ. The probability of having a delay that is more than a time t is given by

$$P[\text{delay} > t] = P[\text{delay} > 0]e^{-(N_u - \rho)\mu t} \tag{5.8}$$

This indicates the exponential distribution of the delay time. The average delay is then given by the average of the exponential distribution:

$$D = P[\text{delay} > 0]\frac{1}{\mu(N_u - \rho)} \tag{5.9}$$

Example 5.19: Call Delay Using Erlang C Formula For the ferry described in Example 5.17, answer the following questions:

1. What is the average delay for a passenger to get access to the telephone?
2. What is the probability of having a passenger waiting more than a minute for access to the telephone?

Solution

1. Using Eq. (5.7), for $N_u = 5$ and $\rho = 2.5$ we have $P[\text{delay} > 0] = 0.13$. Using Eq. (5.9), the average delay is $0.13/(5–2.5)/3 = 0.17$ min.
2. Using Eq. (5.8), $P[\text{delay} > 1 \text{ min}] = 0.13 \exp[-(5–2.5)1/3] = 0.13 \exp(-0.83) = 0.0565$.

5.3 DISTRIBUTED RANDOM ACCESS SCHEMES

Random access methods have evolved around bursty data applications for computer communications. In our discussion of fixed-assignment access methods, we noted that such methods make relatively efficient use of communications resources when each user has a steady flow of information to be transmitted. This would be the case, for example, with digitized voiced traffic, data file transfer, or facsimile transmission. However, if the information to be transmitted is intermittent or bursty in nature, then fixed-assignment access methods can result in communication resources being wasted for much of the duration of the connection. Furthermore, in wireless networks, where subscribers pay for service as a function of channel connection time, fixed-assignment access can be an expensive means of transmitting short messages and will also involve large call set-up times. *Random access* methods provide a more flexible and efficient way of managing channel access for communicating short bursty messages. In contrast to fixed-assignment access schemes, random access schemes provide each user station varying degrees of freedom in gaining access to the network whenever information is to be sent. A natural consequence of randomness of user access is that there is contention among the users of the network for access to a channel, and this is manifested in collisions of contending transmissions. Therefore, these access schemes are sometimes called contention-based schemes or simply *contention schemes.*

Random access techniques are widely used in wired LANs, and the literature in computer networking provides an adequate description of these techniques. When applied to wireless applications these techniques often are modified from their original wired version. The objective of the rest of this section is to describe the evolution of random access techniques that are used in wireless networks. We first discuss the random access methods used in wireless data networks and then we provide some details of the access methods used in WLAN applications.

5.3.1 Random Access Methods for Data Services

The random access methods used data networks can be divided into two groups. The first group consists of ALOHA-based access methods for which the terminals transmit their packets without any coordination between them (they contend for the medium). The second class is the carrier-sense-based random access techniques for which the terminal senses the availability of the channel before it transmits its packets.

ALOHA-based Random Access Techniques. The original *ALOHA protocol* is sometimes called *pure ALOHA* to distinguish it from subsequent enhancements of the original protocol. This protocol derives its name from the ALOHA system, a communications network developed by Norman Abramson and his colleagues at the University of Hawaii and first put into operation in 1971 [Abr70]. The initial system used ground-based UHF radios to connect computers on several of the island campuses with the university's main computer center on Oahu, by use of a random access protocol which has since then been known as the ALOHA protocol. The word ALOHA means hello in Hawaiian.

The basic concept of the ALOHA protocol is very simple. A terminal transmits an information packet upon when the packet arrives from the upper layers of the protocol stack. Simply put, terminals say "hello" to the medium interface as the packet arrives. Each packet is encoded with an error-detection code. The BS/AP checks the parity of the received packet. If the parity checks properly, then the BS sends a short acknowledgement packet to the MS.

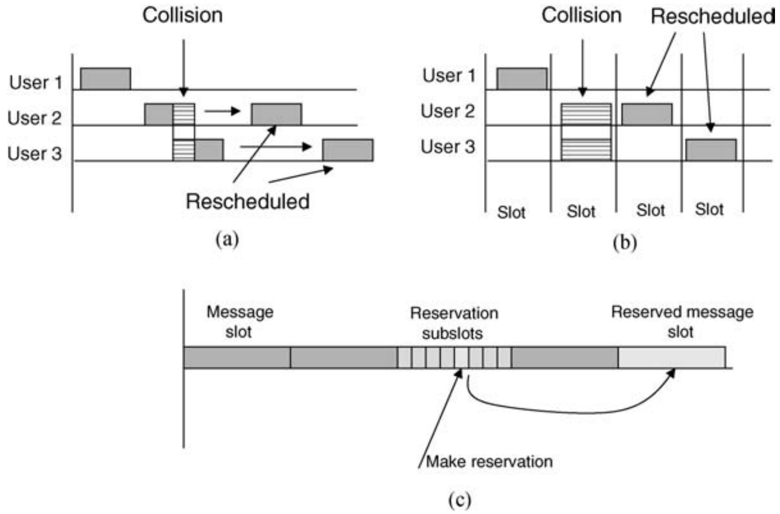


FIGURE 5.10 (*a*) Pure ALOHA protocol; (*b*) slotted ALOHA protocol; (*c*) R-ALOHA

Of course, since the MS packets are transmitted at arbitrary times, there will be collisions between packets whenever packet transmissions overlap by any amount of time, as indicated in Fig. 5.10*a*. Thus, after sending a packet, the user waits a length of time more than the round-trip delay for an acknowledgment (ACK) from the receiver. If no acknowledgment is received, then the packet is assumed lost in a collision and it is transmitted again with a randomly selected delay to avoid repeated collisions.

The advantage of the ALOHA protocol is that it is very simple and it does not impose any synchronization between mobile terminals. The terminals transmit their packets as they become ready for transmission, and they simply retransmit if there is a collision. The disadvantage of the ALOHA protocol is its low throughput under heavy load conditions. If we assume that packets arrive randomly, they have the same length, and are generated from a large population of terminals, then the maximum throughput of the pure ALOHA is 18%.

Example 5.20: Throughput of Pure ALOHA

1. What is the maximum throughput of a pure ALOHA network with a large number of users and a transmission rate of 1 Mb/s?
2. What is the throughput of a TDMA network with the same transmission rate?
3. What is the throughput of the ALOHA network if only one user was effective?

Solution

1. For a large number of mobile terminals each using a transmission rate of 1 Mb/s to access a BS/AP using the ALOHA protocol, the maximum data rate that successfully passes through to the BS is 180 kb/s.
2. If we have a TDMA system with negligible overhead (long packets), the throughput defined this way is 100% and throughput is 1 Mb/s.
3. The 1 Mb/s can be attained in an ALOHA system only if we have one user (no collision) who transmits all the time.

In wireless channels, where bandwidth limitations often impose serious concerns for data communications applications, this technique is often changed to its synchronized version, referred to as slotted ALOHA. The maximum throughput of a slotted ALOHA system under the conditions mentioned above is 36%, which is double the throughput of pure ALOHA.

In the slotted ALOHA protocol, shown in Fig. 5.10*b*, the transmission time is divided into time slots. The BS/AP transmits a beacon signal for timing and all MSs synchronize their time slots to this beacon signal. When a user terminal generates a packet of data, the packet is buffered and transmitted at the start of the next time slot. With this scheme we eliminate partial packet collision. Assuming equal-length packets, either we have a complete collision or we have no collisions. This doubles the throughput of the network. The report on collision and retransmission mechanisms remains the same as in pure ALOHA. Because of its simplicity, the slotted ALOHA protocol is commonly used in the early stages of registration of an MS to initiate a communication link with the BS.

Example 5.21: Slotted ALOHA in GSM In the GSM system, the initial contact between the MS and the BS tower to establish a traffic channel for TDMA voice communications is performed through a random access channel using the slotted ALOHA protocol. Other voice-oriented cellular systems adopt similar approaches as the first step in the registration process of a MS.

The throughput of the slotted ALOHA protocol is still very low for wireless data applications. This technique is sometimes combined with TDMA systems to form the so-called reservation ALOHA (R-ALOHA) protocol, shown in Fig. 5.10*c*. In R-ALOHA, time slots are divided into contention periods and contention-free periods. During the contention interval, MSs use very short packets to contend for the upcoming contention-free intervals that will be used for transmission of long information packets. The R-ALOHA protocol was used in the ALTAIR WLANs that were developed in the early 1990s to operate in licensed frequency bands around 18–19 GHz. The detailed implementation of R-ALOHA can take a variety of forms, and for that reason it is sometimes used under different names. The following example provides some details of the so-called dynamic slotted ALOHA protocol that is used in the Mobitex mobile data networks.

Example 5.22: Dynamic Slotted ALOHA Mobitex has a full-duplex communication capability (simultaneous transmissions on the uplink and downlink) and employs a *dynamic* slotted ALOHA protocol. Suppose that there are three MSs (MS_1, MS_2, and MS_3) in a cell. The situation is such that the BS has two messages to send to MS_3, MS_1 has a short status update that requires one slot, MS_2 has a long message to send, and MS_3 has nothing to transmit. An MS can transmit only during certain "free" cycles consisting of several slots of equal length that are periodically initiated by the BS using a *free* frame on the downlink. In this example, shown in Fig. 5.11, the BS indicates that there are six free slots for contention, each of a certain length. This can change depending on the traffic; hence the term "dynamic." Also note that MSs cannot transmit whenever they want, as in slotted ALOHA. The MSs with traffic to send, i.e. MS_1 and MS_2, select one of the six slots at random. In this case, MS_1 selects slot 1 and MS_2 selects slot 4. Hence, there is no collision. MS_1 is able to transmit its short status update in slot 1, after which it ceases transmission. MS_2 transmits in slot 4, requesting access to the channel using a message called ABD. Simultaneously, the BS would have transmitted its message to MS_3. Upon receipt of the message, MS_3 acknowledges it. The free slots are designed to be of the duration of the downlink message to MS_3 so

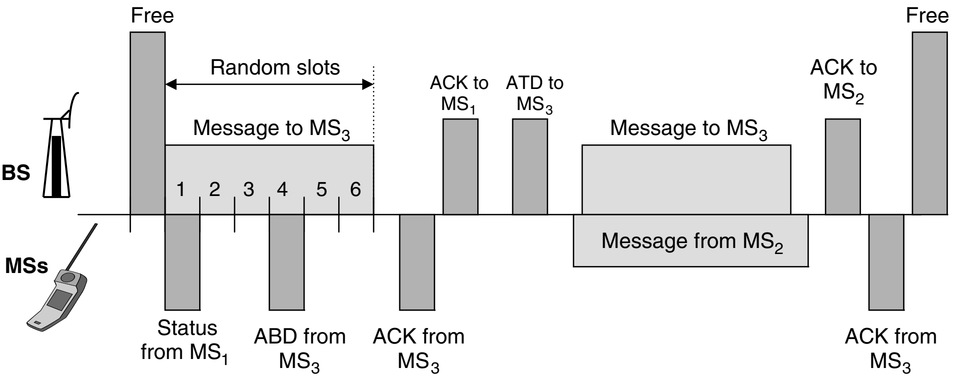


FIGURE 5.11 Dynamic slotted ALOHA used in Mobitex.

that the ACK from MS_3 can be received without contention. The BS also acknowledges the status report from MS_1 and sends an access grant (ATD) to MS_2. As MS_2 transmits its long message on the uplink, the BS can simultaneously send the second message to MS_3. After proper ACKs are transmitted and received, a new *free* cycle is started.

Example 5.23: Packet Reservation Multiple Access An example of a system that uses reservation for integrating voice and data services is the work done by David Goodman and his colleagues in developing the concept of packet reservation multiple access (PRMA) [Goo89]. PRMA is a method for transmitting, in a wireless environment, a variable mixture of voice packets and data packets. The PRMA system is closely related to R-ALOHA, in that it merges characteristics of slotted ALOHA and TDMA protocols. PRMA has been developed for use in centralized networks operating over short-range radio channels. Short propagation times are an important ingredient in providing acceptable delay characteristics for voice service. Our description here closely follows that of Goodman and coworkers [Goo89, Goo91].

The transmission format in PRMA is organized into frames, each containing a fixed number of time slots. The frame rate is identical to the arrival rate of speech packets. The terminals identify each slot as either “reserved” or “available,” in accordance with a feedback message received from the BS at the end of the slot. In the next frame, only the user terminal that reserved the slot can use a reserved slot. Any terminal not holding a reservation that has information to transmit can use an available slot.

Terminals can send two types of information, referred to as periodic and random. Speech packets are always periodic. Data packets can be random (if they are isolated) or periodic (if they are contained in a long unbroken stream of information). One bit in the packet header specifies the type of information in the packet. A terminal having periodic information to send starts transmitting in contention for the next available time slot. Upon successfully detecting the first packet in the information burst, the BS grants the sending terminal a reservation for exclusive use of the same time slot in the next frame. The terminal in effect “owns” that time slot in all succeeding frames as long as it has an unbroken stream of packets to send. After the end of the information burst, the terminal sends nothing in its reserved slot. This, in turn, causes the BS to transmit a negative acknowledgment (NACK) feedback message indicating that the slot is once again available.

To transmit a packet, a terminal must verify two conditions. The current time slot must be available and the terminal must have permission to transmit. Permission is granted according to the state of a pseudorandom number generator, permissions at different terminals being statistically independent. The terminal attempts to transmit the initial packet of a burst until the BS acknowledges successful reception of the packet or until the terminal discards the packet, because it has been held too long. The maximum holding time $D_{\max}$ (seconds) is determined by delay constraints on speech communication and is a design parameter of the PRMA system. If the terminal drops the first packet of a burst, it continues to contend for a reservation to send subsequent packets. It drops additional packets as their holding times exceed the limit $D_{\max}$. Terminals with periodic data (as opposed to voice) packets to send store packets indefinitely while they contend for slot reservations (equivalent to setting $D_{\max}$ to infinity). Thus, as a PRMA system becomes congested, both the speech packet dropping rate and the data packet delay increase.

Goodman and Wei [Goo91] analyze PRMA efficiency, which they quantify as the maximum number of conversations per channel that the system can support within a chosen constraint on packet-dropping probability. In their work they adopted a constraint of $P_{\text{drop}} < 0.01$. They used a speech source rate of 32 kb/s and a header length of 64 bits in each packet. Using computer simulation methods, they investigated the effects of six system variables on PRMA efficiency: (1) channel rate, (2) frame duration, (3) speech activity detector, (4) maximum delay, (5) permission probability, and (6) number of conversations. Over the range of conditions examined, they found many PRMA configurations capable of supporting about 1.6 conversations per channel and found that this level of efficiency could be maintained over a wide range of conditions. Their overall conclusion was that PRMA shows encouraging potential as a statistical multiplexer of speech packets. However, they judged that there are still many questions to be answered in order to verify that PRMA will perform properly in short-range radio systems. Issues requiring further investigation include the effects of mixing random information packets (data) and periodic information packets (speech) in PRMA and the effects of packet transmission errors on PRMA efficiency.

Example 5.24: Reservation in General Packet Radio Service A single 200 kHz carrier in GSM has eight time slots, each capable of carrying data at 9.6 kb/s (standard), 14.4 kb/s (enhanced) or 21.4 kb/s (if FEC is completely omitted). The raw data rate can thus be as high as $8 \times 21.4 = 171.2$ kb/s. The same time slots can be reserved for data access using slotted ALOHA. Medium access is based on a slotted ALOHA reservation protocol. In the contention phase a slotted ALOHA random access technique is used to transmit reservation requests, the BS then transmits a notification to the MS indicating the channel allocation for an uplink transmission, and finally the MS can transfer data on the allocated slots without contention. On the downlink, the BS transmits a notification to the MS indicating the channel allocation for downlink transmission of data to the MS. The MS will monitor the indicated channels and the transfer occurs without contention.

Carrier Sense Multiple Access-based Random Access Techniques. The main drawback of ALOHA-based contention protocols is the lack of efficiency caused by the collision and retransmission process. In ALOHA, users do not take into account what other users are doing when they attempt to transmit data packets, and there are no mechanisms to avoid collisions. A simple method to avoid collisions is to sense the channel before transmission of a packet. If there is another user transmitting on the channel, then it is obvious that a terminal should delay the transmission of the packet. Protocols employing this concept are referred to

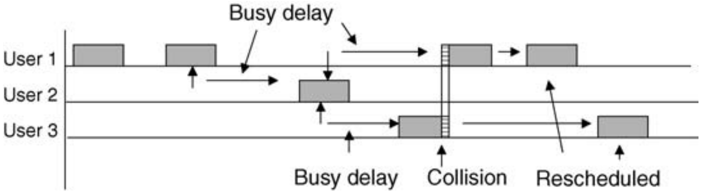


FIGURE 5.12 Basic operation of CSMA protocol.

as CSMA or listen-before-talk (LBT) protocols. Figure 5.12 shows the basic concept of the CSMA protocol. Terminal "1" will sense the channel first and then sends a packet. This is followed by a sensing and packet transmission by terminal "1" again. During the second transmission time of terminal "1", terminal "2" senses the channel and discovers that another terminal is using the medium. It then delays its transmission for a later time using a back-off algorithm. The CSMA protocol reduces the packet collision probability significantly compared with the ALOHA protocol. However, it cannot eliminate the collisions entirely. Sometimes, as shown in Fig. 5.12, two terminals sense the channel busy and reschedule their packets for a later time but their transmission times overlap with each other, causing a collision. Such situations do not cause a significant operational problem because the collisions can be handled in the same way as they were handled in ALOHA. However, if the propagation time between the terminals is very long, then such situations happen more frequently, thereby reducing the effectiveness of carrier sensing in preventing collisions. As a result, several variations of CSMA have been employed in local-area applications, whereas ALOHA protocols are preferred in wide-area applications.

Example 5.25: Examples of Wireless Networks that Employ ALOHA and CSMA As described earlier, Mobitex uses a variation of the ALOHA protocol while the IEEE 802.11 standard for WLANs employs a version of the CSMA protocol. The ALOHA protocol is also used in the random access logical channels in cellular telephone and satellite communication applications (WANs).

A number of strategies are used for the sensing procedure and retransmission mechanisms, which has resulted in a number of variations of the CSMA protocol for a variety of wired and wireless data networks. Figure 5.13 depicts the key elements of distinction among these protocols. If, after sensing the channel, the terminal attempts another sensing only after a random waiting period, then the carrier-sensing mechanism is called "non-persistent." After sensing a busy channel, if the terminal continues sensing the channel until the channel becomes free, the protocol is referred to as "persistent." In persistent operation, after the channel becomes free, if the terminal transmits its packet right away then it is referred to as "1-persistent" CSMA; and if it runs a random number generator and based on the outcome it transmits its packet with a probability p, then the protocol is called p-persistent CSMA.

In a wireless network, owing to multipath and shadow fading, as well as to the mobility of terminals, sensing the availability of the channel is not as simple as in the case of wired channels. Typically in a wireless network, two terminals can each be within range of some intended third terminal but out of range of each other, because they are separated by excessive distance or by some physical obstacle that makes direct communication between the two terminals impossible. This situation, where the two terminals cannot sense the

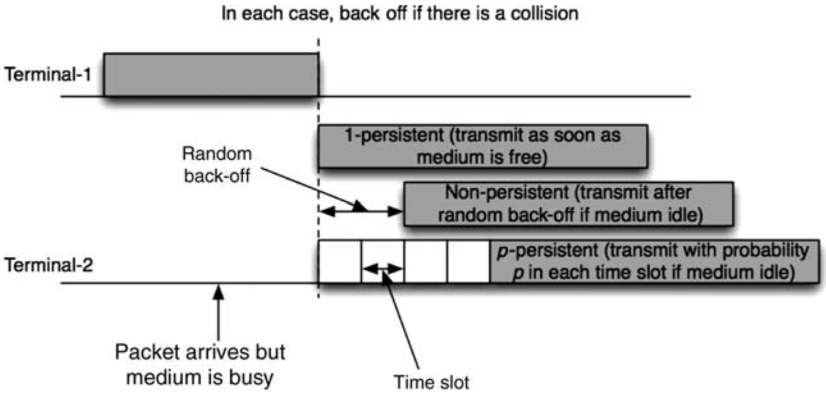

In each case, back off if there is a collision
Terminal-1
Random back-off
1-persistent (transmit as soon as medium is free)
Non-persistent (transmit after random back-off if medium idle)
Terminal-2
p-persistent (transmit with probability p in each time slot if medium idle)
Packet arrives but medium is busy
Time slot

FIGURE 5.13 Retransmission alternatives for CSMA.

transmission of each other but a third terminal can sense both of them, is referred to as the *hidden terminal problem*. This is a more likely situation in cases of radio networks covering wider geographic areas in which hilly terrain blocks some groups of user terminals from sensing other groups. In this situation the CSMA protocol will successfully prevent collisions among the users of one group but will fail to prevent collisions between users in groups hidden from one another.

To resolve the hidden terminal problem we need to facilitate the sensing procedure. In multi-hop adhoc networks, where there is no centralized station or infrastructure, a protocol called *busy-tone multiple-access* (BTMA) has been used in packet radio for military applications. A brief summary of BTMA is given in [Tob80], where a number of packet communication protocols are discussed and compared. In the BTMA scheme, the system bandwidth is divided into two channels: a *message channel* and a *busy-tone channel*. Whenever a station sends signal energy on the message channel, it transmits a simple busy-tone signal (e.g. a sinusoid) on its busy-tone channel. When any other terminal senses a busy-tone signal, it turns on its own busy tone. In other words, as a terminal detects that some user is on the message channel, it sounds the alarm on the busy-tone channel in an attempt to inform every user, including those hidden to the transmitting terminal. A user station with a packet ready to send first senses the busy-tone channel to determine whether the network is occupied.

Most cellular mobile data networks use different frequencies for forward (downlink) and reverse (uplink channels). The messages in the forward channel are transmitted from the mobile data BS, which is designed and deployed to provide a comprehensive and reliable coverage. In other words, the BSs are not hidden to the mobile terminals, whereas the mobile terminals may be hidden from one another. In this situation one may use the forward channel to announce the availability of the channel for the mobile terminals. This concept is used in a protocol referred to as *digital* or *data sense multiple-access* (DSMA). DSMA is very popular in mobile data networks and it is used in CDPD, ARDIS, and TETRA. In DSMA, the forward channel broadcasts a periodic busy-idle bit announcing the availability of the reverse channel for data transmission. A mobile terminal checks the busy-idle bit prior to transmission of its packet. As soon as the MS starts its transmission, the BS will change the busy-idle bit to the busy state to prevent other mobile terminals from transmission. Since the sensing process is performed after demodulation of data from the digital information, it is referred to as digital or data sense, rather than carrier sense, multiple access.

5.3.2 Access Methods for Local-Area Networks

Compared with WANs, a LAN operates over shorter distances with smaller propagation delays and, consequently, a transmission medium that is well suited for variations of the CSMA protocol. Low-speed WANs are developed for communicating shorter messages, whereas LANs are designed to facilitate large file transfers at high data rates. As a result, the length of the packets in LANs is much larger than the length of the packets in low-speed mobile data networks. When the length of the packets is long it would be very useful to pay further attention to packet collisions. LANs often employ variations of the CSMA protocol that either stop transmission as soon as a packet collision is detected or add additional features to avoid collisions.

Example 5.26: Packet Sizes in Wide- and Local-Area Wireless Networks

1. Determine the transfer time of a 20 kB file with a mobile data network with a transmission rate of 10 kb/s.
2. Repeat for an 802.11 WLAN operating at 2 Mb/s.
3. What is the length of the file that the WLAN of part (2) can carry in the time that mobile data service of part (1) carries its 20 kB file.

Solution

1. The early mobile data networks, such as ARDIS and Mobitex, limited the length of a file to around 20 kB. For a data rate of around 10 kb/s it would take $20 \times 8/10 = 16$ s to transfer such a file.
2. An IEEE 802.11 network operating at 2 Mb/s would transfer this file in 80 ms.
3. In a 16 s time interval the same WLAN transfers a 4 MB file.

The most popular version of the CSMA for wired LANs is CSMA/CD adopted in the IEEE 802.3 (Ethernet) standard, the dominant standard for wired LANs supporting data rates that can be up to several gigabits per second. The basic operation of CSMA/CD is the same as CSMA implementations discussed earlier. The defining feature of CSMA/CD is that it provides for detection of a collision shortly after its onset, and each transmitter involved in the collision stops transmission as soon as it senses a collision. In this way, colliding packets can be aborted promptly, minimizing the wastage of channel occupancy time by transmissions destined to be unsuccessful. More details of Ethernet are provided in Chapter 8. Other LAN multiple access protocols employed the idea of taking turns (as in polling or token-based schemes), but such schemes are not widely deployed. Ethernet pretty much dominates the wired LAN deployments of today.

Unlike plain CSMA, which requires an acknowledgment (or lack of an acknowledgment) to learn the status of a packet collision, CSMA/CD requires no such feedback information, since the collision-detection mechanism is built into the transmitter. When a collision is detected, the transmission is immediately aborted, a jamming signal is transmitted, and a retransmission back-off procedure is initiated, just as in CSMA [Pah95]. As is the case with any random access scheme, proper design of the back-off algorithm is an important element in assuring stable operation of the network.

Example 5.27: Binary Exponential Back-Off The back-off algorithm recommended by IEEE 802.3 Ethernet is referred to as the binary exponential back-off algorithm, which is combined with the 1-persistence CSMA protocol with collision detection. When a terminal senses a transmission it continues sensing (persistent) until the transmission is completed. After the channel becomes free, the terminal sends its own packet. If another terminal was also waiting, then a collision occurs because of the 1-persistence and the two terminals reattempt transmission with probability $^1/_2$ after a time slot that spans twice the maximum propagation delay allowed between the two terminals. A time slot that spans twice the maximum propagation delay is selected to ensure that in the worst-case scenario the terminal will be able to detect the collision. If a second collision occurs, then the terminals reattempt with probability $^1/_4$, which is half of the previous retransmission probability. If collision persists, then the terminal continues reducing its retransmission probability by half up to 10 times; after that it continues with the same probability six more times. If no transmission is possible after 16 attempts, it reports to the higher layers that the network is congested and transmission shall be stopped. This procedure exponentially increases the back-off time and gives the back-off strategy its name. The disadvantage of this procedure is that the packets arriving later have a higher chance of surviving the collision, which results in an unfair first-come-last-served environment. It can be shown that the average waiting time for the exponential back-off algorithm is $5.4T$, where T is the time slot used for waiting [Tan97, Sta00].

The CSMA/CD scheme is also used in many IR LANs, where both transmission and reception are inherently directional. In such an environment, a transmitting station can always compare the received signal from other terminals with its own transmitted signal to detect a collision. Radio propagation is not directional, posing a serious problem in determining other transmissions during your own transmission. As a result, collision detection mechanisms are not well suited for radio LANs. However, compatibility is very important for WLANs; therefore, designers of these networks have had to consider CSMA/CD for compatibility with the Ethernet backbone LANs that dominate the wired-LAN industry.

While collision detection is easily performed on a wired network, simply by sensing voltage levels against a threshold, such a simple scheme is not readily applicable to radio channels because of fading and other radio channel characteristics. The one approach that can be adopted for detecting collisions is to have the transmitting station demodulate the channel signal and compare the resulting information with its own transmitted information. Disagreements can be taken as an indicator of collisions, and the packet can be immediately aborted. However, on a wireless channel, the transmitting terminal's own signal dominates all other signals received in its vicinity and, thus, the receiver may fail to recognize the collision and simply retrieve its own signal. To avoid this situation the station's transmitting antenna pattern should be different from its receiving pattern. Arranging this situation is not convenient in radio terminals because it requires directional antennas and expensive front-end amplifiers for both transmitters and receivers.

The approach called CSMA/CA, shown in Fig. 5.14, is actually adopted by the IEEE 802.11 WLAN standard. The elements of CSMA/CA used in the IEEE 802.11 are interframe spacing (IFS), contention window (CW), and a back-off counter. The CW intervals are used for contention and transmission of the packet frames. The IFS is used as an interval between two CW intervals. The back-off counter is used to organize the back-off procedure for transmission of packets. The method of operation is best described by an example.

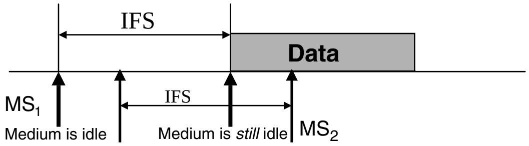


FIGURE 5.14 CSMA/CA adopted by the IEEE 802.11.

Example 5.28: Operation of Collision Avoidance in IEEE 802.11 Figure 5.15 provides an example for the operation of the CSMA/CA mechanism used in the IEEE 802.11 standard. Stations A, B, C, D, and E are engaged in contention for transmission of their packet frames. Station A has a frame in the air when stations B, C, and D sense the channel and find it busy. Each of the three stations will run its random number generator to get a back-off time by random. Station C followed by D and B draws the smallest number. All three terminals persist on sensing the channel and defer their transmission until the transmission of the frame from terminal A is completed. After completion, all three terminals wait for the IFS period and start their counters immediately after completion of this period. As soon as the first terminal, station C in this example, finishes counting its waiting time it starts transmission of its frame. The other two terminals, B and D, freeze their counter to the value that they have reached at the start of transmission for terminal C. During transmission of the frame from station C, station E senses the channel, runs its own random number generator, which in this case ends up with a number larger than remainder of D and smaller than the remainder of B, and defers its transmission for after the completion of station C's frame. In the same manner as the previous instance, all terminals wait for IFS and start their counters.

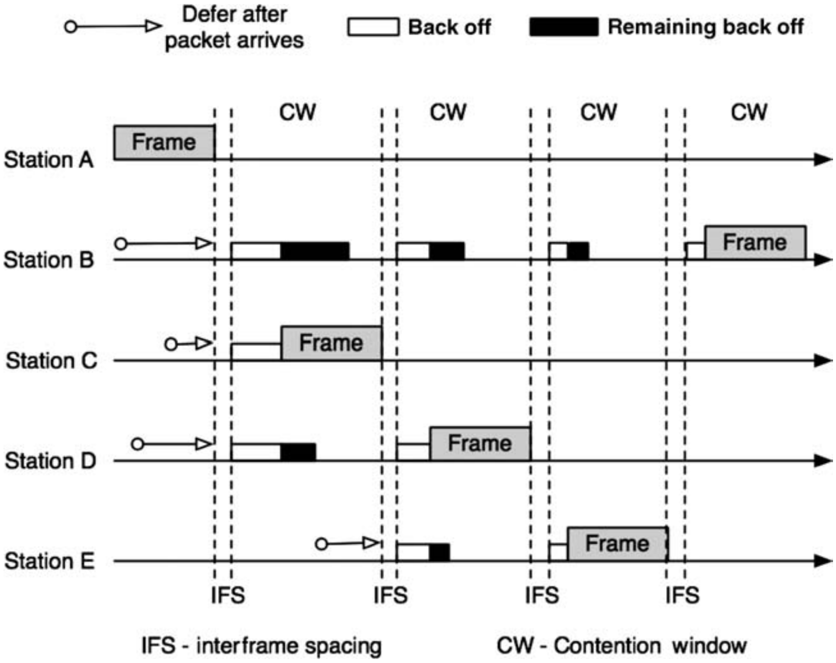


FIGURE 5.15 Illustration of CSMA/CA.

Station D runs out of its random waiting time earlier and transmits its own packet. Stations B and E freeze their counters and wait for the completion of the frame transmission from terminal D and the IFS period after that before they start running down their counters. The counter for terminal E runs down to zero earlier, and this terminal sends its frame while B freezes its counter. After the IFS period following completion of the frame from station E, the counter in station B counts down to zero before it sends its own frame. The advantage of this back-off strategy over the exponential back-off used in IEEE 802.3 is that the collision detection procedure is eliminated and the waiting time is fairly distributed in a way that, on average, a first come, first served policy is enforced.

Another related technique considered for collision avoidance in WLANs is the *combing* method [Wil95]. As shown in Fig. 5.16, the time is divided into comb and data transmission intervals. During the comb period each station alternates between transmission and listening periods according to a code assigned to the station. All stations will continue advancing in their code until they sense a carrier during their listening period. If they do not sense a carrier at the end of the code, then they transmit their packet. If they sense a carrier, then they postpone their transmission until the next comb interval. A simple example will further clarify this method.

Example 5.29: Combing for Collision Avoidance Figure 5.16 shows three stations with five-digit codes 11101 (terminal A), 11010 (terminal B), and 10011 (terminal C). All three terminals transmit their carrier during the first slot, because all codes have 1 in that slot. In the second slot, terminal C will listen and after sensing the other two withdraws from contention and waits for the next combing. In the third period, station B goes to a listening state and after sensing the carrier of terminal A defers its transmission until the next cycle. Terminal A continues its sequence of alternating transmissions and listening until the end of the comb period, when it transmits its data packet (as it heard no other terminal). After completion of the data transmission from station A, the other two terminals will wait for an interpacket spacing (IPS) for a new contention, after which station B transmits its packet. Station C will transmit after the second transmission cycle.

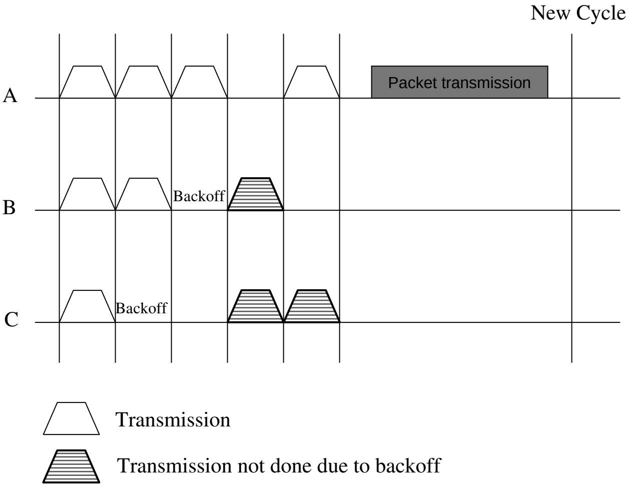


FIGURE 5.16 Illustration of combing.

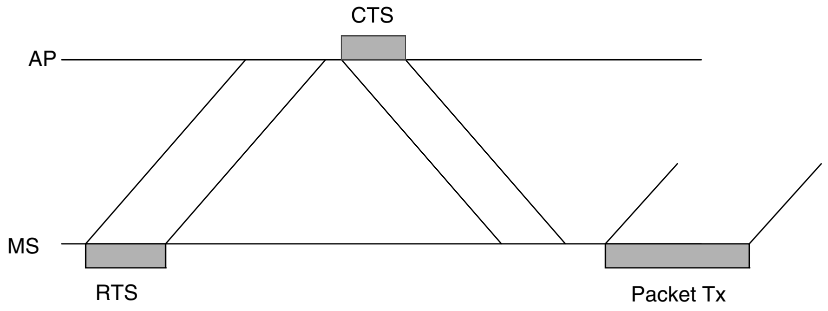


FIGURE 5.17 RTS/CTS in IEEE 802.11.

In CSMA/CA, as we will see later, priority is assigned by dividing the IFS into several differently sized intervals associated with different priority levels. In combing, priority can be arranged by assigning different classes of numbers to the codes. The lower priority packets will receive earlier zero codes, and higher priority packets will have a zero in their codes in the later intervals.

Another access method used in WLANs is the request-to-send (RTS)/clear-to-send (CTS) mechanism shown in Fig. 5.17. A terminal ready for transmission sends a short RTS packet identifying the source address, destination address, and the length of the data to be transmitted. The destination station will respond with a CTS packet. The source terminal will send its packet with no contention. After acknowledgement from the destination terminal, the channel will be available for other usage. IEEE 802.11 supports this feature, as well as CSMA/CA (see Chapter 9 for more details). This method provides a unique access right to a terminal to transmit without any contention.

Quality of Service in Wireless Local-Area Networks. With the emergence of VoIP as an important application over random access networks, it became important to provide support for QoS for voice and other real-time traffic over such networks. The IEEE 802.11e standard is an attempt to provide quality of service at the MAC layer in WLANs. The basic idea in IEEE 802.11e is to provide different priorities for different classes of traffic. Priority here refers to the ability of frames from certain classes of traffic to access the channel earlier than other classes of traffic. This helps in reducing the delay for such classes of traffic, as well as in providing higher throughput. One way of providing different priorities to different classes of traffic in IEEE 802.11 is to enable different waiting times (interframe spaces) or different back-off intervals for different classes of traffic. This way, voice traffic, for example, could wait for a much shorter period of time than Web traffic, reducing latency and improving throughput for voice packets.

This, however, implies that access to the channel is not fair (flows from some classes of traffic may be starved of bandwidth). Fair access to the medium while assuring throughput and latency is possible by using techniques that are distributed and yet provide fair access to the medium. Several techniques have been proposed to address this issue (see Pattara-atikom *et al.* [Pat03] for more details).

Medium Access Control Protocols for Wireless Sensor Networks. Wireless sensor networks (see Chapter 13) are a new class of networks that may include several thousand low-cost devices deployed in a sensor field for monitoring and sensing specific phenomena. In such networks, sensors may actually transmit and receive useful data (typically sensed quantities that have been processed locally) only very infrequently. Sensors may be deployed in areas that may be inaccessible to humans. Consequently, it is important to ensure that the limited batteries in such devices last for years. Medium access schemes designed for high-speed WLANs are quite unsuitable for sensor networks. As discussed in Chapter 13, coordinating sleep schedules of sensors to reduce the amount of idle listening is one method of enhancing the battery life in sensors. The reader is referred to Demirkol *et al.* [Dem06] for a survey of MAC protocols for sensor networks.

Enhancements for High-Speed Operation. One of the drawbacks of the CSMA/CA protocol employed in WLANs is that the waiting times employed severely limit the throughput of the network. Recall that no useful data is sent for the IFS time, during the back-off slots, or during the transmission of the acknowledgment frames. These times form an overhead for the transmission of every single MAC layer frame in IEEE 802.11. In particular, as the data rates increase, the waiting times become an increasingly large part of the transmission of a frame. Consider that the actual size of a frame reduces as the data rate increases. If the waiting times remain fixed, then eventually they dominate the transmission of frames; so, the throughput in a WLAN cannot exceed a certain threshold irrespective of how high the physical transmission rate is. To overcome this problem, in IEEE 802.11n (the high-speed standard for WLANs), additional MAC layer features have been introduced to reduce the overhead. These include aggregating several frames for transmission and block acknowledgments for several frames [Xia05, Sko08].

5.3.3 Performance of Random Access Methods

In voice-oriented circuit-switched networks, performance is measured by the probability of blockage (blockage rate) of initiating a call. If the call is not blocked, a fixed-rate full-duplex channel is allocated to the user for the entire communication session. In other words, interaction between the user and the network takes place in two steps. First, during the call establishment procedure, the user negotiates the availability of a line with the network and, if successful (not blocked), the network guarantees a connection with a certain QoS (data rate, delay, error rates) to the user. For real-time interactive applications, such as telephone conversations or video conferencing, if the user does not talk then the resource allocated to the user is wasted. If these facilities, originally designed for two-way voice application, are used for data application, then: (1) for bursty data, file transfers during the idle times between the transmission of two packet bursts' allocated resources are wasted; (2) large file transfers suffer a long delay or waiting time for the transfer because resources allocated to each user are more restricted.

Users of packet-switched networks are always connected and there is neither an initiation (negotiation) procedure to be blocked nor a fixed QoS to be allocated. In this situation, analysis of the performance for real-time interactive applications such as telephone conversations is complicated and will be addressed later. The performance of these networks for data applications is often measured by the average throughput S and average delay D versus the total offered traffic G. The *channel throughput* S is the average number of successful packet transmissions per time interval T_{p}. The offered traffic G is the number of packet transmission attempts per packet time slot T_{p}, which includes new arriving packets as well as retransmissions of old packets. The average delay D is the average waiting time before successful transmission, normalized to the packet duration T_{p}. The standard unit of

TABLE 5.1 Throughput of Various Random Access Protocols

Protocol	Throughput
Pure ALOHA	$S = Ge^{-2G}$
Slotted ALOHA	$S = Ge^{-G}$
Unslotted 1-persistent CSMA	$S = \dfrac{G[1+G+aG(1+G+aG/2)]\mathrm{e}^{-G(1+2a)}}{G(1+2a)-(1-\mathrm{e}^{-aG})+(1+aG)\mathrm{e}^{-G(1+a)}}$
Slotted 1-persistent CSMA	$S = \dfrac{G[1+a-\mathrm{e}^{-aG}]\mathrm{e}^{-G(1+a)}}{(1+a)(1-\mathrm{e}^{-aG})+a\mathrm{e}^{-aG(1+a)}}$
Unslotted non-persistent CSMA	$S = \dfrac{G\mathrm{e}^{-aG}}{G(1+2a)+\mathrm{e}^{-aG}}$
Slotted non-persistent CSMA	$S = \dfrac{aG\mathrm{e}^{-aG}}{1-\mathrm{e}^{-aG}+a}$

traffic flow is the Erlang, which can be thought of as the number of the packets per packet duration time T_{p}. The throughput is always between 0 and 1 Erlang while the offered traffic G may exceed 1 Erlang.

The analyses of the relationships between S, G, and D for a variety of medium access protocols have been a subject of research for a few decades. This analysis depends on the assumptions on the statistical behavior of the traffic, the number of terminals, the relative duration of the packets, and the details of the implementation. Assuming a large number of terminals generating fixed-length packets with a Poisson distribution,[4] Table 5.1 summarizes the throughput expressions for the ALOHA and 1-persistent and non-persistent CSMA protocols, including the slotted and unslotted versions of each. The expressions for p-persistent protocols are very involved and are not included here. The interested reader should refer to [Kle75b, Tob75, Tak85], where the derivations of the other CSMA expressions can also be found. The expressions in the table are also derived by Hammond and O'Reilly [Ham86] and [Kei89]. The parameter a in this table corresponds to the normalized propagation delay, defined as $a = \tau/T_{\mathrm{p}}$, where τ is the maximum propagation delay for the signal to go from one end of the network to the other end.

Example 5.30: Calculation of the normalized propagation delay Determine the parameter a in IEEE 802.3 (Ethernet) 10 Mb/s LANs and IEEE 802.11 2 Mb/s LANs.

Solution The IEEE 802.3 standard for star LANs allows a maximum length of 200 m between two terminals. The propagation speed in the cables is usually approximated by 200 000 km/s, resulting in $\tau = 1\ \mu\mathrm{s}$. The IEEE 802.11 allows a maximum distance of 100 m between the AP and the MS. The radio propagation is at the rate of 300 000 km/s, resulting in $\tau = 0.33\ \mu\mathrm{s}$. For a star LAN operating at 10 Mb/s with 1000 bit packets, the value is $a = 0.01$. For an IEEE 802.11 operating at 2 Mb/s with the same packet size, $a = 0.00066$

Figure 5.18 shows plots of throughput S versus offered traffic load G for the six protocols listed in Table 5.1, with normalized propagation delay of $a = 0.01$. All curves follow the same pattern. Initially, as the offered traffic G increases, the throughput S also increases up to a point where it reaches a maximum $S_{\max}$. After the throughput reaches its maximum value, an increase in the offered traffic actually reduces the throughput. The first region depicts the

[4] The Poisson distribution assumes packets are generated independent from one another and the interarrival time between the packets forms an exponentially distributed random variable.

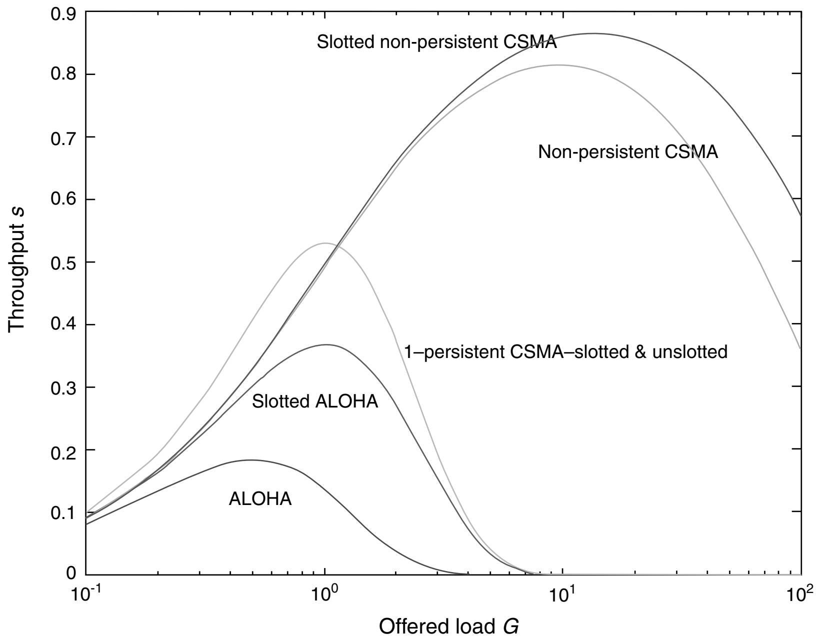


FIGURE 5.18 Throughput S versus offered traffic load G for various random access protocols.

stable operation of the network, in which an increase in aggregating traffic G, which includes arriving traffic as well as retransmissions due to collisions, increases the total successful transmissions and, thus, S. The second region represents unstable operation, where an increase in G actually reduces the throughput S because of congestion and eventually halts the operation. In practice, as we saw in the last section, retransmission techniques adopted for the real implementation include back-off mechanisms to prevent operation in unstable regions.

The throughput curves for the slotted and unslotted versions of 1-persistent CSMA are essentially indistinguishable. It can be seen from the figure that, for low levels of offered traffic, the 1-persistent protocols provide the best throughput, but the non-persistent protocols are by far the best at higher load levels. It can also be seen that the slotted non-persistent CSMA protocol has a peak throughput almost twice that of persistent CSMA schemes.

The equations in Table 5.1 can also be used to calculate capacity, which is defined as the peak value $S_{\max}$ of throughput over the entire range of offered traffic load G [Ham86]. An example is helpful to show how to relate the curve to a particular system.

Example 5.31: Relating Throughput and Offered Traffic to Data Rates To relate throughput and offered traffic to data rates, assume that we have a centralized network that supports a maximum data rate of 10 Mb/s and serves a large set of user terminals with the pure ALOHA protocol.

1. What is the maximum throughput of the network?
2. What is the offered traffic in the medium and how is it composed?

Solution

1. Since the peak value of the throughput is $S = 18.4\%$, the terminals contending for access to the central module can altogether succeed in getting at most 1.84 Mb/s of information through the network.
2. At that peak the total traffic from the terminals is 5 Mb/s (because the peak occurs at $G = 0.5$), which is composed of 1.84 Mb/s of successfully delivered packets (some mixture of new and old packets) and 3.16 Mb/s of packets doomed to collide with one another.

Plots of capacity versus normalized propagation delay are plotted in Fig. 5.19 for the same set of ALOHA and CSMA schemes. The curves show that for each type of protocol the capacity has a distinctive behavior as a function of normalized propagation delay *a*. For the ALOHA protocols, capacity is independent of *a* and is the largest of all the protocols (compared when *a* is large). As we discussed earlier, this is the case where the area of coverage is large and propagation delays are comparable to the length of packets. The plots in Fig. 5.19 also show that the capacity of 1-persistent CSMA is less sensitive to the normalized propagation delay for small *a* than is non-persistent CSMA. However, for small *a*, non-persistent CSMA yields a larger capacity than does 1-persistent CSMA, though the situation reverses as *a* approaches the range 0.3–0.5 [Ham86].

Another important performance measure for packet data communications is the delay characteristics of the transmitted packets. For real-time applications, such as voice conversations, if the delay is more than a certain value (several hundred milliseconds) the packet is not useful and it is dropped. Therefore, we need to analyze the delay characteristics of the channel to determine the capacity of the access method. In data transfer applications, the delay characteristics are usually related to the throughput of the medium and it usually

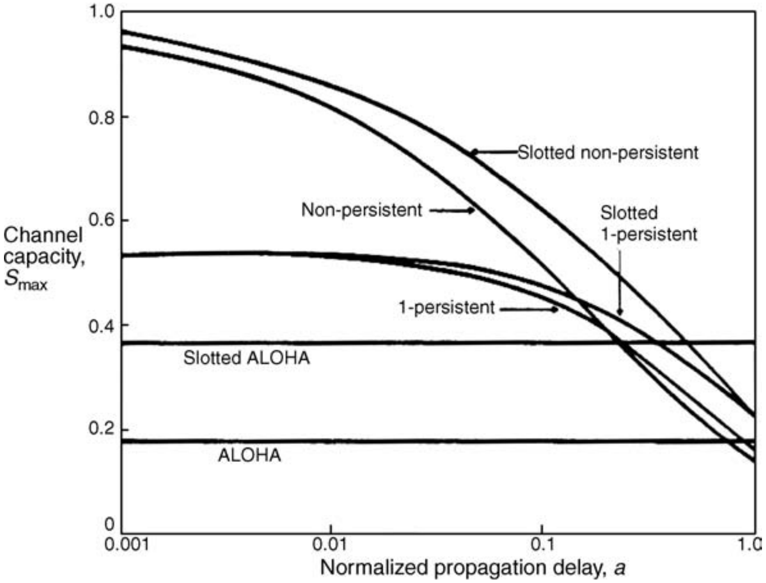


FIGURE 5.19 Capacity versus normalized propagation delay for various random access protocols.

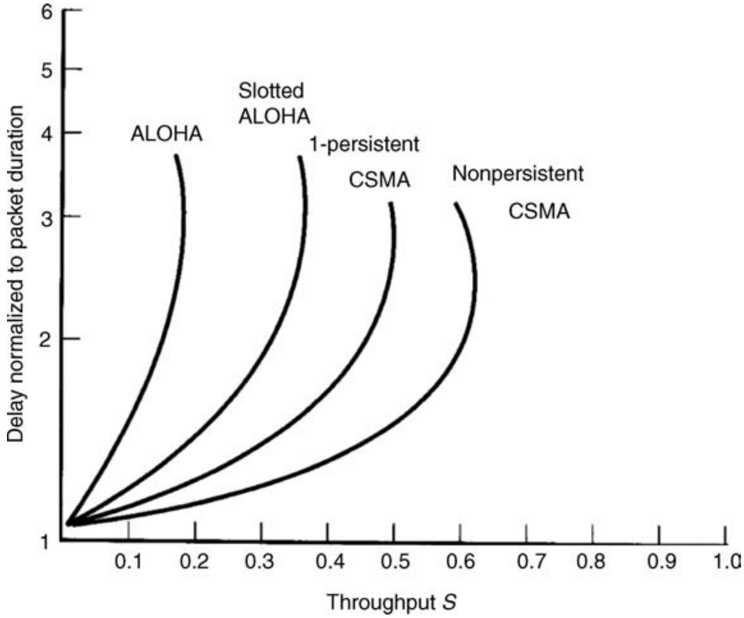


FIGURE 5.20 Delay versus throughput for various random access protocols.

follows a hockey-stick shape. At low traffic, when a small fraction of the maximum throughput is utilized, the delay often remains the same as the transmission delay. As the throughput increases, so the number of retransmitted packets increases, resulting in a higher average delay for the packets. Around the maximum throughput the delay retransmissions grow rapidly, pushing the network toward an unstable condition where the channel is dominated with retransmissions and the packet delays grow extremely large. Figure 5.20 shows the delay-throughput behavior of the ALOHA, S-ALOHA and CSMA protocols.

Practical Considerations. The analysis provided above is abstract and is used to provide an intuitive framework for the operation of different classes of access methods. In practice, implementations deviate considerably from the abstract and the performance is evaluated by analysis or simulation of case-by-case situations. Examples of this type of analysis for CSMA/CD with the exponential back-off algorithm used in the IEEE 802.3 Ethernet and the performance of the token-ring access method used in IEEE 802.5 are available in the last chapter of Stallings [Sta00].

Complications Caused by Wireless Channel. Three factors that are effective in throughput analysis in a wired environment are propagation delay, users' idle period (not transmitting), and packet collisions. In a wireless environment, analysis of the real throughput of a protocol is much more complicated because it involves hidden terminal and capture effects. To analyze these effects, let us assume we have a centralized AP with a number of terminals connected to it and communicating via a random access method.

Figure 5.21 demonstrates the basic concept behind the hidden terminal problem. The two terminals contending to communicate with the AP are both in the coverage area of the AP but they are out of the coverage area of each other. Limited antenna range and shadowing are two major causes for hidden terminal degradation. The hidden terminal problem does not

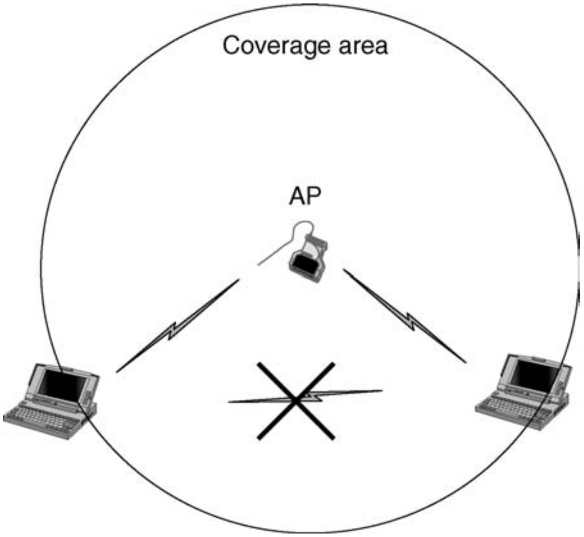


FIGURE 5.21 Hidden terminal problem.

affect the performance of the ALOHA-type protocols, but it degrades the performance of CSMA protocols. In a CSMA environment affected by the hidden terminal problem, some terminals cannot sense the carrier of the transmitting terminal and their transmitted packets have a higher probability of colliding, degrading the overall throughput.

In real installations, the coverage area of the AP is usually larger than that of the mobile terminals, because the AP is installed in a selected location to optimize the coverage (high on the walls or on the ceiling), which will increase the negative impacts of the hidden terminal problem. Assuming that the coverage areas of the AP and the mobile terminals are the same, there is still no guarantee that all the terminals in the coverage area of the AP can hear one another. This is because two terminals at the maximum distance L from the AP could be as far as $2L$ apart. Therefore, the hidden terminal problem is unavoidable and natural to the operation of the centralized access systems using the CSMA protocol that is common in WLAN operations.

Another phenomenon impacting the throughput of a radio network is capture. In radio channels, sometimes the collision of two packets may not destroy both packets. Because of signal fading or the near–far effect, packets from different transmitting stations can arrive with different power levels, and the strongest packet may survive a collision. Figure 5.22

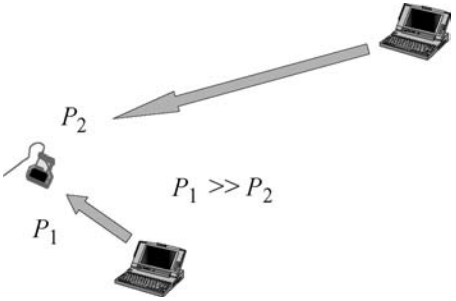


FIGURE 5.22 The capture effect.

shows the basic concept of the capture phenomenon. The received power from the terminal closer to the AP is much larger than the received power from the terminal located at a distance. If two packets collide in time, then the packet with the weaker signal will appear as a background noise and the AP captures (detects) the packet from the closer terminal successfully. The capture effect increases the throughput of the radio network because, in calculating the throughput, we always assume that the colliding packets are destroyed (not detected).

The hidden terminal effects were first analyzed for different types of CSMA protocol used in rapidly moving packet radio networks for military applications, and busy-tone signaling was suggested for eliminating the hidden terminal problem. More recently, there have been efforts to analyze the effects of capture and hidden terminals in WLAN environments using various assumptions [Zha92, Zah97].

In reality, the capture of a packet is a random process, which is a function of the modulation technique used for transmission, the received SNR, and the length of the packet.

Example 5.32: Capture Effect and Throughput Figure 5.23 [Zha92] shows the effects of capture on the throughput of the conventional slotted ALOHA and the CSMA systems for packet lengths of 16, 64, and 640 bits. Also shown for comparison are the curves for conventional non-persistent CSMA and slotted ALOHA without capture. With capture, the maximum throughput of CSMA with packet length 16 bits is 0.88 Erlang, which is 0.065 Erlang more than the case without capture. The maximum throughput for slotted ALOHA with the same packet length is 0.591 Erlang, which is 0.231 Erlang higher than the case without capture.

In slow fading channels, if the terminal generating the test packet is in a "good" location, then the interference from other packets is small and all the bits of the test packet survive the collision. In contrast, for a test packet originating from a terminal in a "bad" location, all the

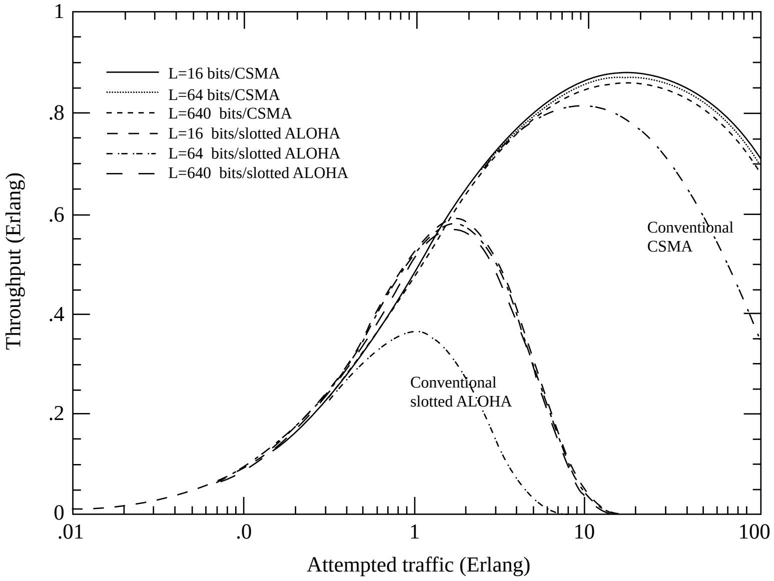


FIGURE 5.23 Effects of packet length on throughput for CSMA and slotted ALOHA with capture. The modulation is BPSK and the SNR is 20 dB.

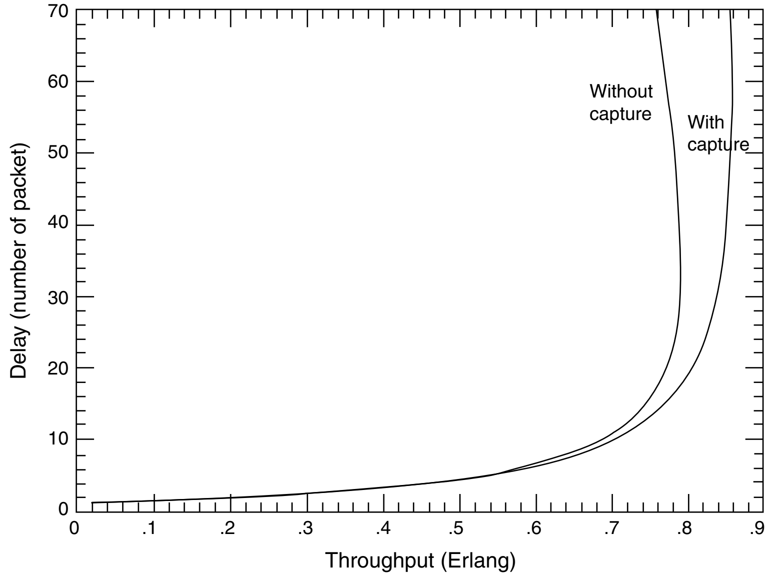


FIGURE 5.24 Delay versus throughput of CSMA for BPSK modulation and SNR of 20 dB, with and without capture.

bits are subject to high probability of error and the packet does not survive the collision. As a result, the system shows minimal sensitivity to the choice of packet length, which is consistent with our assumption of slow fading. Figure 5.24 [Zha92] shows the delay throughput for the CSMA protocol with and without capture for a 640-bit packet network. The packet delay is normalized to the length of the packet. For both cases, as the throughput approaches its maximum value the system becomes rapidly unstable, causing unacceptable delays for the delivery of the packets. When the capture effects are included, the maximum throughput is increased and the instability occurs at a slightly higher value of the throughput.

Example 5.33: Effect of Capture and Hidden Terminals in a WLAN Environment Figure 5.25 [Zah97] shows the throughput versus offered traffic curves for a WLAN AP using the CSMA protocol and surrounded by a large number of terminals uniformly distributed within the AP's coverage area. In this scenario, as shown in Fig. 5.26, each terminal senses a group of terminals within its coverage area (area I in the figure) and cannot sense those that are out of its coverage area but are still within the coverage area of the AP (area II in the figure). The throughput of the target terminal with respect to terminals in area I is the same as the throughput of a CSMA system. However, the throughput of the target terminal with respect to terminals in area II is the same as the throughput of ALOHA networks, because carrier sensing does not work and the terminals transmit their packets without knowledge of the transmission from terminals in area II. Using these facts, in [Zah97] the throughput at each point in the area of the coverage of the AP is calculated and then it is averaged over the entire coverage area of the AP for different coverage areas for the mobile terminals. Obviously, this average throughput will always remain between the throughput of CSMA and that of ALOHA. The lowest curve in Fig. 5.25 shows the throughput when the hidden terminal problem is considered and the coverage of each terminal is 70% of the coverage of the AP. Because a number of terminals cannot sense the

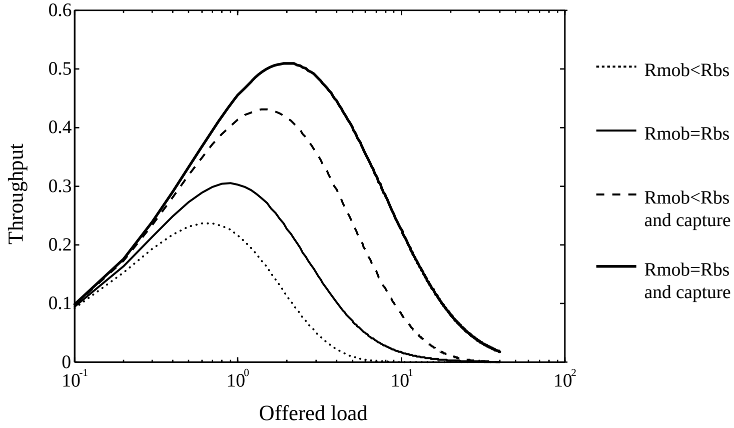


FIGURE 5.25 Throughput versus offered traffic for a WLAN with a large number of terminals.

transmission of others, the peak throughput has declined to less than 25%, which is slightly higher than the 18% maximum throughput of ALOHA and far below the maximum of CSMA. The second curve from below, with a peak value of around 30%, represents the same results where the coverage of the AP and the mobile terminals are the same. The third curve depicts the performance when the effects of both hidden terminal and capture are considered and the coverage of the mobile terminals is smaller than that of the AP. The capture effect increases the throughput to more than 40%. The top curve is the same as the third curve, where the coverages of the AP and the mobile terminals are the same. This situation has

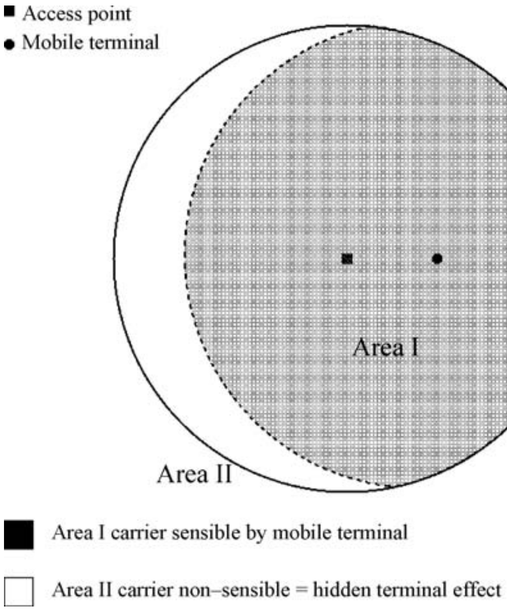

■ Access point
● Mobile terminal
Area I
Area II
Area I carrier sensible by mobile terminal
Area II carrier non–sensible = hidden terminal effect

FIGURE 5.26 Coverage areas of an AP and a tagged mobile terminal in a WLAN.

increased the throughput by another 10% to above 50%, which is getting closer to the performance of conventional CSMA.

5.4 INTEGRATION OF VOICE AND DATA TRAFFIC

As the wireless communications industry moves toward 3G and 4G networks, one of the important objectives is the use of a single wireless system for multimedia applications to support a variety of communications services, including voice and data and voice in various forms and combinations. A key technical problem to be dealt with in such integrated systems is that of multiuser access. As we saw earlier in this chapter, an access method that efficiently supports one category of service may be unsuitable for another category of service. In the 1G and 2G networks, as we saw in Chapter 1, the wireless industry evolved around two separate paths for voice-oriented and data-oriented applications. If a data service can be efficiently integrated with a voice service, then transmission resources that are otherwise wasted (because there is no voice transmission) can be used for data, which typically does not have stringent delay requirements. First, the voice-oriented networks evolved into supporting data. More recently, with the popularity of VoIP over the Internet and PSTN, supporting voice in WLANs has become attractive as well.

5.4.1 Access Methods for Integrated Services

As we saw earlier, in a packet communication environment, voice and data have different requirements. Voice packets can tolerate errors and even packet losses (a loss of 1–2% of voice packets has an insignificant effect on the perceived quality of reconstructed voice [Kum74]), while data packets are sensitive to loss and errors but can generally tolerate delays. Also, the rate at which information is transmitted is constant in the case of voice, thereby making circuit switching a viable and efficient approach, whereas information generated for data transmission is very bursty. As a result, voice- and data-oriented networks use different multiple access methods. In a wireless environment, the simplest approach is to assign different frequency bands to isochronous (voice) and asynchronous (data) packets. However, integration in one frequency band will result in a more efficient usage of the bandwidth, a simpler radio interface, and an environment that provides a better control for synchronizing voice and video (e.g. lip-sync).

5.4.2 Data Integration in Voice-Oriented Networks

Fixed access methods such as FDMA, TDMA, and CDMA were basically designed for access to the circuit-switched voice-oriented networks. Later on, as we saw in Chapter 1, several data services evolved around these systems. The economic incentive for using this medium for mobile data services is to take advantage, either partially or fully, of the existing infrastructure, the terminals, and the frequency bands designed for the voice-oriented networks. This way, the mobile data service provider saves in the major costs of deployment, which includes the cost of real estate and the installation of the antenna, and there is no longer a need to obtain new frequency bands for operation of the data service. If possible, using the same terminal for voice and data will reduce the cost and facilitate marketing of the service.

Example 5.34: Mobile Data over FDMA Analog Cellular The cellular digital packet data (CDPD) system described introduced in the early 1990s uses available frequency channels in the existing analog FDMA cellular telephone network (AMPS) to provide an overlaid packet data service supporting data rates similar to voice-band modems (up to 19.2 kb/s). In its present form this system does not exploit the pauses between talk spurts, but simply takes advantage of the frequency bands temporarily unused by mobile telephone users in each cell area. CDPD uses the unused AMPS channel to develop a communication link between a mobile data unit and a mobile data BS. Ideally, a CDPD terminal can use the RF and antenna of an AMPS terminal to communicate packet data bursts. However, the most important issue for the CDPD network is that it can use the same antenna site and the antenna towers and the frequency bands of an existing AMPS network. Since real estate, installation of the antenna post, and the frequency bands are perhaps the most expensive parts for implementation of a network, CDPD was perceived to provide a cost-efficient solution for a mobile data service with a comprehensive coverage. The air interface protocol and modulation technique used in the CDPD, however, are different from the AMPS system.

Example 5.35: Mobile Data over TDMA Systems The GPRS packet data network, introduced in the late 1990s, uses the air interface and the infrastructure of the GSM network to provide a mobile packet data service that can support data rates of up to a couple of hundred kilobits per second. GPRS uses the same physical packet format and modulation technique as GSM. The logical channels used in GPRS do not use the dialing procedure used in the GSM. In a manner similar to CDPD, through the wired infrastructure of the network, the packets of data in GPRS are routed to the packet-switched data networks rather than being switching through the PSTN. GPRS is designed to take advantage of the unused time slots of a TDMA voice-oriented GSM network.

Example 5.36: Mobile Data over CDMA Networks In TDMA systems and FDMA systems, data users may use free time slots and free channels respectively as they become available. In a CDMA system the situation is somewhat different. The structure of CDMA is such that all active users use the entire bandwidth time space simultaneously. The resource to be managed is signal power. With the application of efficient power control algorithms, the signal levels transmitted by MSs and the BS are continually adjusted in response to the changing locations of mobiles and the number of users on the system at any given time. In a CDMA network, the integration of data calls with voice calls is straightforward in principle, since various numbers of both categories of calls are readily mixed together, with each call accessing the channel with its unique user signal code. Therefore, in a CDMA system, no modification needs to be made to the channel access scheme to accommodate integration of voice and data "channels," and the information rate for voice or data traffic in any one channel can, in principle, be varied by a variable-rate scheme such as is used for voice service in the IS-95 standard. The integration of voice and data services in a single user channel is not necessarily straightforward.

From a technical point of view, there are two incentives for integration of data into voice-oriented fixed-assignment access methods:

1. The fixed-assignment access methods used in voice-oriented networks are designed to support a certain number of simultaneous users. When the number of active users falls below that number, some portion of the transmission resources is wasted.

2. A typical two-way conversation does not make full use of the call connection time, since only one party talks at a time. Furthermore, the flow of natural speech is actually composed of *talk spurts* with intervening short pauses. It is generally estimated that, in a two-way voice connection, the average *voice activity factor* for each party is in the vicinity of 40%; thus, about 60% of the available transmission time remains unused.

Assume that we have N voice channels available for a given area (e.g. the coverage area of a sectored antenna in a cellular deployment) to accommodate newly originated calls and calls handed off from other areas. Further assume that the overall calls in the area are generated according to a Poisson process with normalized rate of $\rho = \lambda/\mu$ calls/unit channel and the length of call holding is generally distributed with a unit mean. The number of idle channels N_{idle} according to renewal theory [Bud97] is given by

$$N_{\text{idle}} = N_{\text{u}} - \rho[1 - B(N_{\text{u}}, \rho)] \tag{5.10}$$

where $B(N_{\text{u}}, \rho)$ represents the blocking probability for the M/G/1 queue calculated from the Erlang B equation given by

$$B(N_{\text{u}}, \rho) = \frac{\rho^{N_{\text{u}}}/N_{\text{u}}!}{\sum\limits_{i=0}^{N_{\text{u}}} (\rho^{\text{i}}/i!)} \tag{5.11}$$

Figure 5.27 shows the average number of the idle voice channels per area N_{idle} versus the number of available channels N for variety of call blocking rates. At call blocking rates of around 2%, which is desirable for most cellular systems, a number of free channels are available for data communications. As the network accepts larger values of blocking rates, because most of the time all channels are in use, regardless of the number of channels, only a few channels will be available for data. If the integrated system uses the idle channels for data transmission, then, on average, the maximum throughput available to the data user is N_{idle} times the encoding rate of the voice channel. If the system can take advantage of the silence periods in the two-way telephone conversations, as we indicated earlier, then an additional throughput of up to 60% is available for data applications.

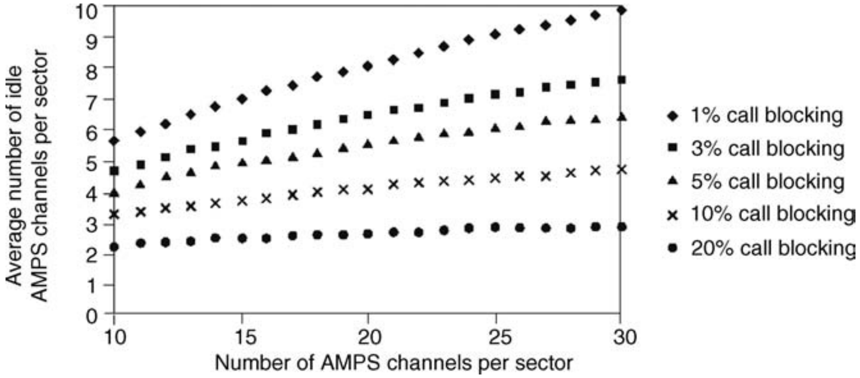


FIGURE 5.27 Average number of idle channels per area as a function of the number of available channels for a given call blocking rate.

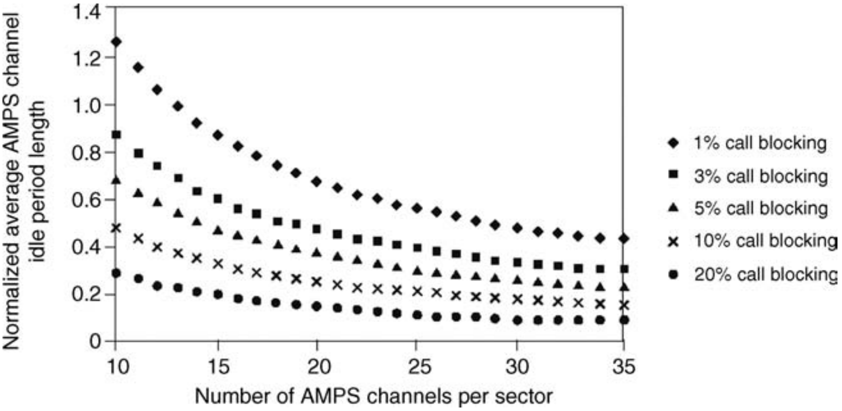


FIGURE 5.28 Normalized average idle period per channel as a function of the number of channels and call blocking probability.

Another important parameter is the idle time for a voice channel. The idle time is the period of time that a voice channel is not occupied by a voice user. In an N-channel voice-oriented network, assuming that each channel receives an equal fraction of the call load, one may calculate T_{I}, the average length of time a channel is idle, by [Bud97]

$$T_{\mathrm{I}} = \frac{N_{\mathrm{u}} - \rho[1 - B(N_{\mathrm{u}}, \rho)]}{\rho[1 - B(N_{\mathrm{u}}, \rho)]} \tag{5.12}$$

Figure 5.28 shows the normalized average idle period per channel T versus the number of available voice channels for different blocking rates. For a typical holding time of around a couple of minutes and a blocking rate of around 2%, the average idle periods are fairly long, implying that a data network has a reasonable time to detect the availability and send its bursts of data.

There are periods of time for which all the voice channels are occupied for the voice users and there is no channel available to the data service. If these periods are short and infrequent, then some data applications may accept the situation. Otherwise, specific channels should be allocated for data-only usage. In these situations the data users have their own channel as well as unused portions of the voice channels.

Assuming that we accept the block out periods and assign no dedicated channel resources to the data application, we will have periodic operation between the available and blocked out periods. Assuming that the holding time for the telephone conversations is exponentially distributed, it can be shown (See [Bud97]) that the average active period where a channel is available for data T_{a} is given by

$$T_{\mathrm{a}} = \frac{1 - B(N_{\mathrm{u}}, \rho)}{N_{\mathrm{u}} B(N_{\mathrm{u}}, \rho)} \tag{5.13}$$

The mean length of the blackout period T_{b} for this case is independent of the call load and, hence, blockage rate and is given by

$$T_{\mathrm{b}} = \frac{1}{N_{\mathrm{u}}} \tag{5.14}$$

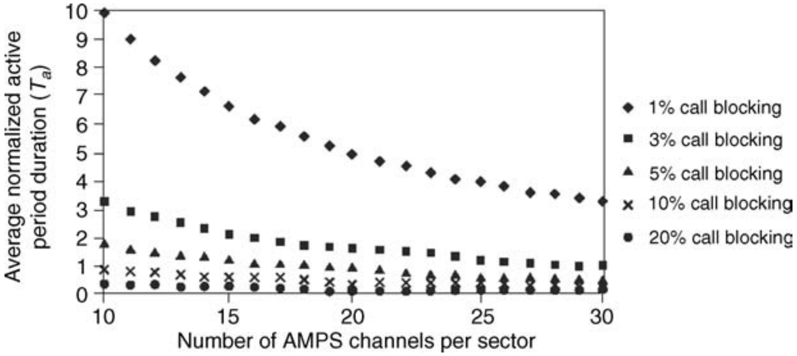


FIGURE 5.29 Mean length of available time for data as a function of the number of voice channels.

Figure 5.29 shows the normalized mean length of the active period T_a versus the number of channels. As the call blocking rate increases, the active period shortens, leaving the system in more blackout periods. For a blocking rate of 5% or less and with fewer than 25 channels, the system has the equivalent of one dedicated channel for data applications. At higher blocking rates and data applications that cannot tolerate blackout periods, the data service must be deployed with at least one dedicated channel. Some practical examples are helpful at this stage.

Example 5.37: Data Overlay in FDMA systems – CDPD The above analysis was actually developed for CDPD, and all of the above analyses and discussions are directly applicable to it. As we saw in Chapter 1, CDPD operates over analog FDMA cellular networks at a rate of 19.2 kb/s per channel, which corresponds to digital transmission using Gaussian minimum shift keying (GMSK) modulation over 30 kHz AMPS channels. CDPD supports a channel hopping feature that allows a mobile data terminal to move to another channel during a communication session, releasing the current channel for voice telephone conversation. This feature helps maintain the blockage probability at its nominal value and allows continual operation during the handoffs. The weakness of CDPD data overlay is that it does not assign several voice channels for one data user to support higher data rates. In general, in an FDMA system, the assignment of multiple voice channels to a single data user involves simultaneous operation of several RF channels by one terminal, which is not practically attractive. For the same practical reason, data overlay in FDMA systems encounters difficulties in taking advantage of the silence periods during telephone conversations.

Example 5.38: Data Overlay in TDMA systems – GPRS An example of TDMA data overlay is GPRS. All of the above analysis is applicable to GPRS applications as well. However, the format flexibility of TDMA allows multislot assignment to support higher data rates. GPRS also does not take advantage of silence periods in two-way telephone conversation.

An efficient method of integrating voice and data packets is a *movable boundary TDMA* scheme with *silence detection*. This method has been applied in the time-assignment speech interpolation (TASI) system used in T1-carrier telephone networks [Fis80] to maximize the number of voice users carried and to integrate data transmission into the channel. Using this

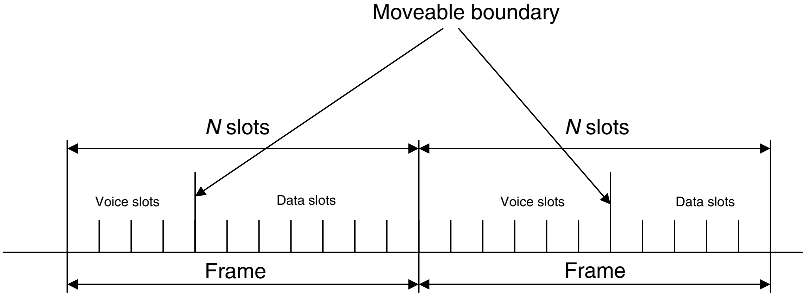


FIGURE 5.30 Frame structure in a moveable boundary frame-polling system.

basic idea, it is possible to design a *TDMA/framed-polling* protocol to integrate voice and data packets in a WLAN. This system consists of a number of voice and data terminals and a central station, which coordinates all the transmissions [Zha90]. The protocol for integration of voice and data packets is a movable boundary TDMA scheme, shown in Fig. 5.30. A frame is divided into two regions with a boundary between them. The first region is used for both voice and data traffic, where the voice traffic has priority. If not, voice packets occupy all the slots in this region and the remaining slots are used for data traffic. The second region is reserved exclusively for data traffic. The boundary between the voice and data regions moves in accordance with the number of active voice packets in each frame. The maximum number of voice packets per frames is N_1, which is assigned an appropriate value to ensure some minimum data traffic capacity and to keep the blockage of voice packets below a selected value (2% in Zhang and Pahlavan [Zha90]).

The result of extensive analysis and simulation by Zhang and Pahlavan [Zha90] provides a simple experimental relation between the capacity that can be allocated for voice and data applications: $D = R_T - 0.032N_v - 0.29$, where D is the data rate in megabits per second available for data applications, N_v is the number of active 64 kb/s PCM-encoded telephone conversations, and R_T is the transmission rate available on the medium. We can apply this equation to the TASI system that uses this protocol to accommodate 30 voice users with some additional data in a traditional 24-voice channels system (T1 carrier). The transmission rate is $R_T = 0.064 \times 24 = 1.536$ Mb/s to support 24 voice users at 64 kb/s. Using the equation for $N_v = 30$ active users, the data throughput with maximum delay of 10 ms (used in the simulation) is 286 kb/s. The new protocol supports six more voice users plus 280 kb/s data. With the same 24-voice users, 487 kb/s would be available for data applications, which is around 30% of the overall transmission rate.

Example 5.39: Data Overlay on CDMA As we saw earlier in this chapter, integration of bursty data with voice in the CDMA system is very simple, and CDMA systems already take advantage of the voice activity factor. Therefore, with the same infrastructure and terminals, data services can be overlaid on CDMA. If higher data rates are needed, then one can either reduce the processing gain of the data channel or assign several parallel channels for one data link. Indeed, the natural flexibility of CDMA to accommodate a variety of data services is one of the major reasons behind selection of the CDMA for 3G systems. Qualcomm has suggested its *high data rate* (HDR) technology, where asymmetric uplink and downlink data rates can be supported by simply using multiple carriers on the downlink for higher data rates. More discussion of data overlays is provided in Chapter 7.

5.4.3 Voice Integration into Data-Oriented Networks

Integration of voice and data has been discussed extensively in the literature. Most of these studies are concerned with protocols with explicit synchronization between the receiver and the transmitter. These approaches use assignment-based protocols for integration of voice and data that allocate a fixed reference time, such as a slot for transmission of packets. Synchronous systems provide more control on delay for voice traffic, but less flexibility for bursty data traffic. Another approach that does not need explicit synchronization between the receiver and the transmitter is the asynchronous approach. The asynchronous packet approach mostly uses protocols extended from packet data networks, which are more suited for bursty data traffic. The voice traffic in this approach requires relatively complicated handling to limit the delay.

Contention-based packet communications protocols such as ALOHA and CSMA are used for data-oriented wireless networks. They are especially well suited to networks comprising of many user stations each with low average data rate and potentially high peak rates. These protocols can operate with little or no centralized control and they can generally accommodate variable numbers of users in the network. However, contention-based schemes can become very inefficient in sharing the communications resources when the traffic load is heavy, as the system throughput degrades and the transmission delays increase. The unpredictability of throughput and the time delays make these access methods unattractive for voice-dominated communication services, where a minimum throughput and delay are essential for user acceptance of the QoS. Up to recent times (the Internet and wireless age), wired telephone services and the PSTN were producing the dominant source of income for the telecommunications industry. In the past century, telephone users have accepted the quality of the PSTN wired voice services as a normal standard for telephone conversations.

Quality of Service in Voice Services. In the language of digital packet communications, the QoS of the PSTN voice user is specified by a guaranteed 64 kb/s PCM (or 32 kb/s ADPCM) coded data rate and a maximum delay of around 100 ms. We refer to this QoS as *wireline-voice quality*. With the introduction of cellular telephone services in the late twentieth century, users accepted a lower QoS that suffered from the effects of fading due to the radio channel and dropped calls due to handoff, lack of coverage, or other reasons. As we saw in Chapter 1, cordless telephony and PCS services were aimed at bringing their QoS close to that of wireline quality. If the quality of voice in a cordless telephone were far below that of wireline quality, then users would have rejected the service. This is because they have the choice to receive a better QoS (no drops or fading effects) with their wired telephone that is also available at home or in the office. However, users had to accept the lower quality QoS with cellular telephony because there was no other alternative service provision for vehicles and other mobile applications.

Another recent event deviating from the wireline QoS is the emergence of voice over Internet or, as most people refer to it, the VoIP phenomenon. The popularity of the Internet, its penetration into the home market, its capability to support multimedia, and (the most important advantage) its uniform cost for local and long-haul communications have encouraged development of Internet telephony. Operating on a packet-switched environment with contention access, the QoS of these services is not guaranteed at all, and in its present stage of technology, the quality of voice calls is well below that of wireline quality. However, free international calls through an Internet connection have been an incentive for some users to try this option as well.

In wireless networks, VoIP does not make sense for mobile data applications because these services provide low data rates and, after all, they are evolving as an auxiliary network over already existing voice-oriented networks. Nevertheless, in 3G networks, recently, the use of VoIP over their high-speed packet access protocols is being investigated. However, VoIP can be considered for WLAN environments. Imagine a WLAN installation in a stock market hall supporting wireless terminals for the users working in the hall. It would be useful and beneficial if they had a VoIP service at the same terminal to use it for their telephone conversations as well. This incentive has initiated preliminary work on VoIP in a WLAN environment using contention-based access methods. The work in [Zah00, Fei00] determines the number of supported voice terminals in a WLAN environment under a variety of conditions.

Capacity of a Wireless Local-Area Network with Voice and Data. WLANs are becoming very popular in indoor applications, such as in stock exchange halls, where mobile users demand a high-speed wireless data access to the network and voice capabilities for telephone conversations. To deploy such a network, a mathematical framework is very helpful to compare the capacity performance of WLANs with voice and data services in different scenarios. Therefore, for an asynchronous WLAN using the TCP/IP suite, we need to find answers to two questions:

1. What is the number of network telephone calls that can be carried with a given amount of data traffic?
2. What is the maximum data traffic per user for a given number of voice users?

A mathematical framework to answer these two questions is provided by Zahedi and Pahlavan [Zah00], where the integration of voice and data with TCP/IP, which operates in an asynchronous CSMA access environment, is analyzed.

To integrate voice packets in a TCP/IP environment, the first step is to select a speech coder. A variety of speech coding algorithms exist with different rates, as discussed in Chapter 3. We first present some theoretical results and then discuss some experimental results reported in the literature. To reduce the load on the network generated by voice traffic, Zahedi and Pahlavan [Zah00] adopted IMBE, which is a popular low data-rate vocoder (4.8 kb/s) with an acceptable QoS. This vocoder has been used in INMARSAT-M and AUSSAT mobile satellite communication systems and is proposed for APCO-25 standard for narrow-band digital land mobile radio.

Using TCP/IP will provide two options for sending packets in the network: the TCP and the user datagram protocol (UDP). As a streaming protocol, UDP has no support for error correction, acknowledgment, sequencing, and flow control. Under high traffic conditions, the lack of flow control in UDP may cause bandwidth saturation in the Internet that should be prevented by the application program. In contrast, TCP has an error-correcting mechanism, uses acknowledgment, and guarantees in-order packet delivery. These requirements demand additional overhead that increases the average delay and reduces the overall throughput of the network. In general, data packets can tolerate delay but cannot tolerate packet loss, while the voice packets can accept packet loss of the order of 1% but cannot afford delays of more than 200 ms between consecutive packets. In the system described by Zahedi and Pahlavan [Zah00], TCP is used for the data packets to guarantee accuracy of information and UDP for voice packets to handle the delay requirement. This approach is adopted in several products available in the market, while other products use TCP for both voice and data. Although in our analysis TCP is selected for data transmission, there are

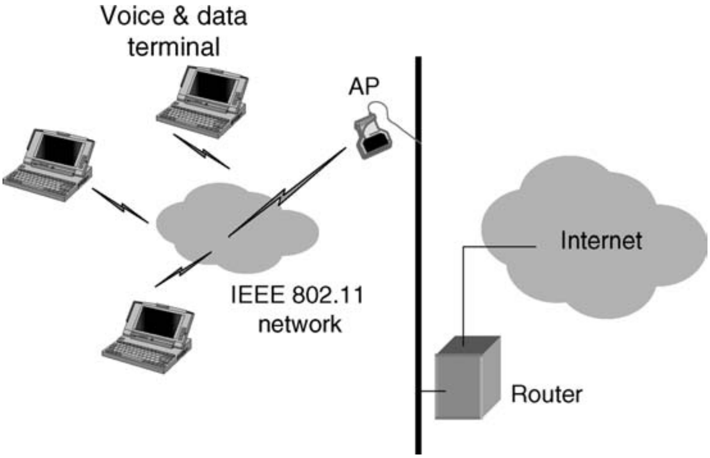


FIGURE 5.31 Schematic of a system employing voice and data terminals in a WLAN.

some products which use UDP for data transmission. In this case the upper layers are responsible for delivery accuracy.

Figure 5.31 presents a general overview of the system where several voice and data terminals communicate with an AP of a WLAN. Since the human ear is sensitive to time delays larger than 200 ms (T_{th}) in a voice conversation, wireless terminals should provide some facilities to minimize voice time delay. Allocating higher priority for transmitting voice over the data traffic is one way to decrease the voice packet time delay. Therefore, voice and data packets should be stored in a queue and wait for transmission, as shown in Fig. 5.32. The total delay for each packet transmission consists of a *queuing delay* at the

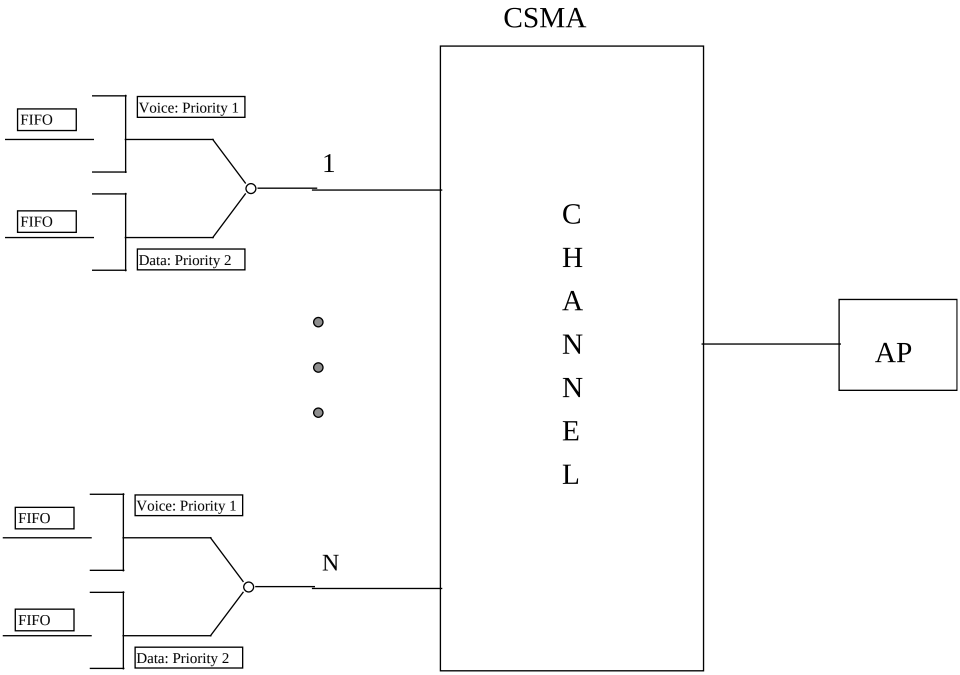
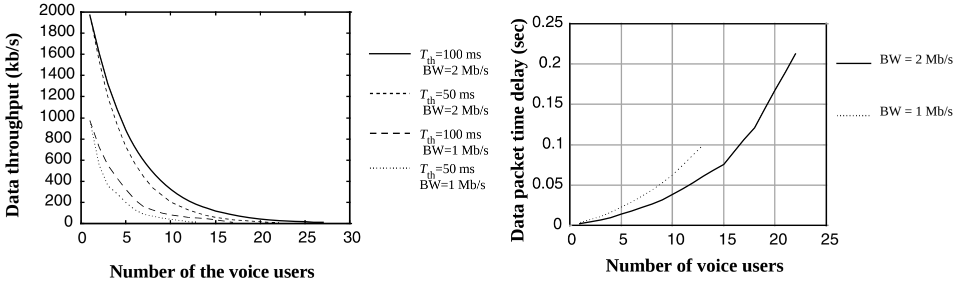


FIGURE 5.32 Queuing model for prioritizing voice traffic.



(a) (b)

FIGURE 5.33 (*a*) Throughput of data versus number of voice users for variety of thresholds for acceptable delay in voice packets. (*b*) Data packet time delay versus number of voice users.

terminal and *channel transfer delay*. The work by Zahedi and Pahlavan [Zah00] uses an M/G/1 queue[5] with two priorities and a single server for node modeling. The arrival is modeled by a Poisson random variable. Figure 5.33 shows the system capacity and delay for $T_{th} = 100$ and 50 ms with 1 and 2 Mb/s channel bandwidths. The maximum number of the voice users declines with reducing T_{th}.

Example 5.40: Capacity of a WLAN with Voice and Data Users Using Fig. 5.33, find the number of users for $T_{th} = 100$ ms and $T_{th} = 50$ ms when the channel bandwidth is 1 Mb/s.

Solution From the figure, a maximum of 18 voice users are supported with $T_{th} = 100$ ms and decreasing T_{th} to 50 ms will reduce the maximum voice users to 14. In this example the data traffic is less than 10 kb/s.

In practice, there are a number of VoIP software packages and services, such as Skype, Speakfreely, Net2Phone, DialPad, etc., that can be used to implement a real-time test bed to analyze the behavior of voice in an IEEE 802.11 WLAN environment. The purpose of setting such experimental test beds or simulations is to determine the number of voice users that can be supported and relate this to the design parameters.

So how do real WLANs do in terms of carrying VoIP calls? A simple analysis that was performed by Garg and Kappes [Gar03] indicates that a single 802.11b cell can support only 3–12 VoIP calls with a G711 codec and 20 ms audio payloads, depending on the average transmission rates of the MS. The analysis by Garg and Kappes [Gar03] was backed by limited experimentation, where calls were gradually added to an 802.11b cell. When a seventh call was added, it caused all calls to result in unacceptable quality. Experiments revealed that the downlink (AP to MSs) was the affected part. The waiting times (IFS and back-off) and the packet overhead (headers and ACKs) in IEEE 802.11 drastically limit the capacity of a WLAN to carry VoIP calls. Similar conclusions under different scenarios were reached by Elaoud and co-workers [Ela04a, Ela04b] in an experimental test bed. The work by Medepalli *et al.* [Med04] confirms the poor capacity of 802.11 WLANs for carrying VoIP calls and extends this to 802.11g and 802.11a systems. Packet losses were also ignored as VoIP quality considerations in this work. If packets are lost, the capacity drops further.

[5] An M/G/1 queue implies that packet arrivals are Poisson and the service rate of the queue has a general distribution with known mean and variance (see [Ber87]).

Work by Wang *et al.* [Wan05] addresses the degradation in capacity due to the downlink by *aggregating VoIP payloads* at the AP into a single multicast packet intended for many MSs. With this scheme, up to 22 VoIP calls could be supported using a GSM 6.10 codec instead of 12 calls without this scheme. This work also demonstrates the degradation of VoIP calls in the presence of even a single background TCP traffic flow (FTP in this case). This is confirmed by analysis using Markov renewal processes and simulations by Harsha *et al.* [Har06] even with the 802.11e standard (which supports priority-based QoS). 802.11e at 11 Mb/s is shown to have a higher capacity than plain 802.11 distributed coordination function (DCF; see Chapter 9 for details) for VoIP calls by Sai Shankar *et al.* [Sha04].

Internet Protocol Telephony Using a Wireless Local-Area Network. The principles of operation of voice over contention-based packet-switched networks and assigned-access circuit-switched networks are very different. In a circuit-switched network, a fixed connection between the two terminals is established during the call establishment procedure. This connection supports end-to-end communications with a fixed data rate and a controlled delay dominated by propagation delay and a negligible delay jitter. In packet voice communications with contention access and packet-switched networks, delay and jitter are the dominant sources of the performance degradation. To regulate jitter, the receiver has a buffer to store the received packets at different delays but pump them to the user at constant intervals to reconstruct real-time voice. The performance of the system is then related to the size of the buffer at the receiver. Next, we provide a summary of the experimental work presented by Feigin *et al.* [Fei00] to relate the throughput to the buffer size.

The first step is to describe the overall scenario in a practical situation for the implementation of VoIP in a WLAN environment. Figure 5.34 describes the arrival of packets. Because of random access and packet-switched networking, the packets sent at fixed intervals arrive at a variety of delays.

The overall delay is the minimum network delay plus the individual jitter per packet that will be regulated with the jitter compensation buffer at the receiver. The packets arriving at different delays are stored in a buffer and the application at the receiver reads this buffer periodically. When the packet with the right sequence number is available, the receiver reads the packet and plays it through the speaker. When the packet is not available, the application software at the receiver skips that packet. A simple example further clarifies the operation.

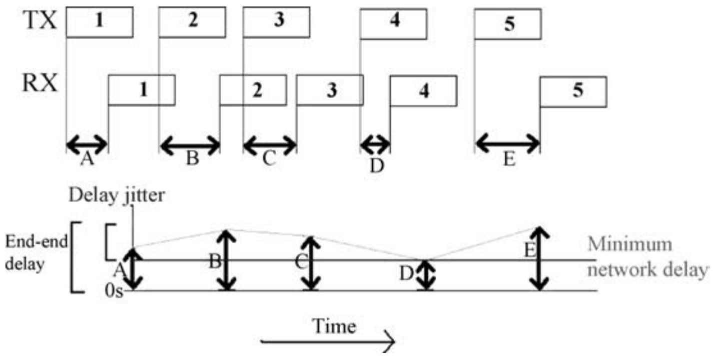


FIGURE 5.34 Illustration of the arrival of voice packets transmitted at a constant rate.

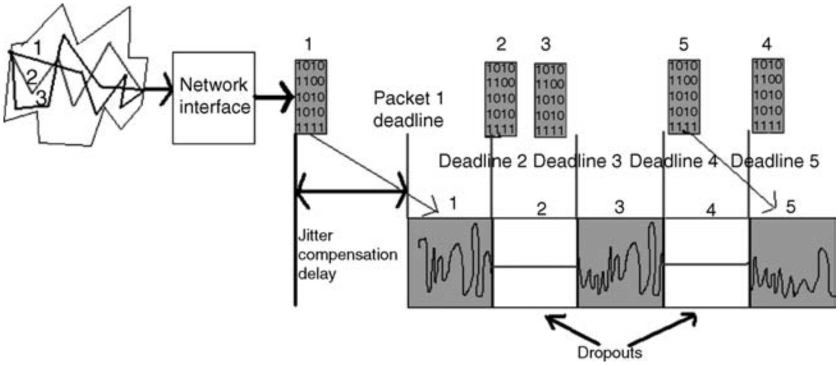


FIGURE 5.35 Reception of voice packets and buffering to maintain appearance of no jitter.

Example 5.41: Jitter in VoIP on WLANs Figure 5.35 illustrates the details of the operation of the receiver and the relationship between the jitter compensation buffer, arrival, and playing time of the packets. When the first packet arrives it is delayed in the receiver's jitter compensation buffer; after the maximum allowed delay, it is delivered to the user application to be played in the first slot. The second packet is discarded because it has arrived after the deadline for playing. The third packet arrives normally before the deadline and it is delivered at its appropriate time to the speaker. The fourth and the fifth packets arrive off-sequence. The fourth one is late (arriving after its deadline) and it is discarded. The fifth packet has arrived before the deadline of the fourth packet and it is shifted to its own time slot.

As we have discussed before, the user can accept a packet drop rate of around 1%. To reach the goal of 1% packet drop, the receiver has the choice of increasing the length of the jitter compensation buffer at the expense of additional overall delay at the receiver. Therefore, the length of the jitter compensation buffer is an important parameter in VoIP applications. This parameter is adjusted by changing the length of the buffer at the receiver. The delay observed by the user is the minimum network delay plus the jitter compensation buffering delay. The above example also illustrates that in VoIP applications, in addition to

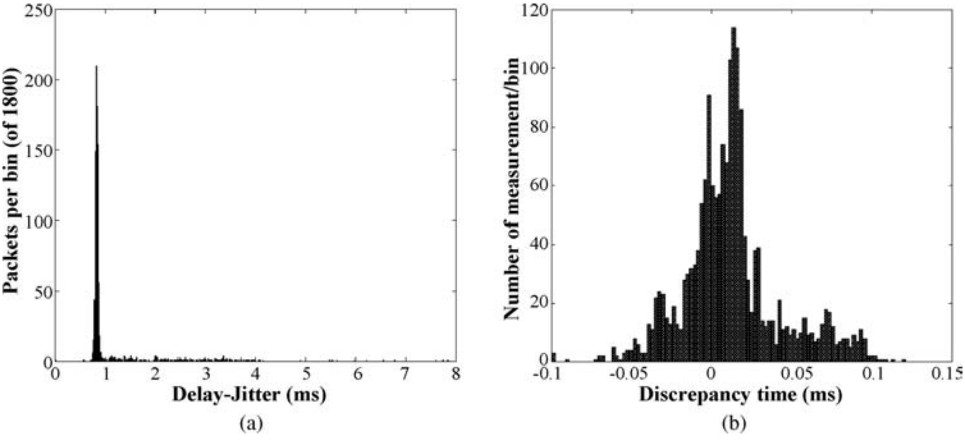


FIGURE 5.36 (*a*) Measured delay jitter in the WLAN test bed; (*b*) accuracy of the measurements.

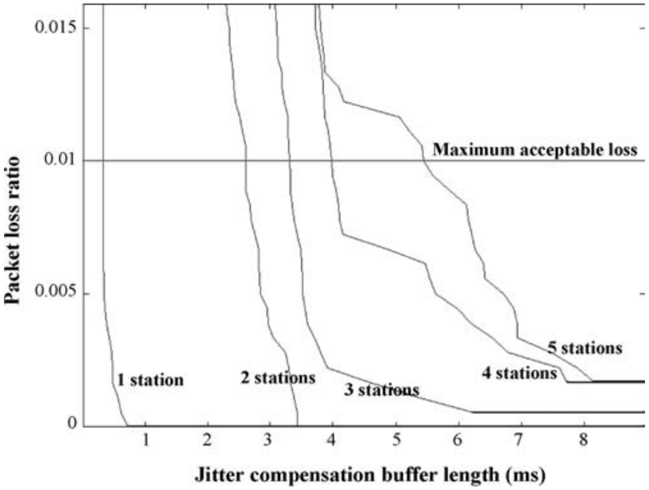


FIGURE 5.37 Packet loss versus jitter compensation buffer length.

transmission packet losses occurring in the network, we have packet losses due to late arrival at the receiver (which is a function of the length of the jitter compensation buffer). Therefore, the length of the jitter compensation buffer and packet loss are interrelated.

To determine the relationship between the jitter compensation delay and the packet loss rate (PLR), a test bed was developed by Feigin *et al.* [Fei00] to implement the scenario shown in Fig. 5.31. In this test bed, an infrastructure for WLAN operation using an AP and a number of laptops is used for measurement of the statistics of the delay jitter in a VoIP application. Figure 5.36*a* shows the statistics of the delay jitter for 1800 packets. The measurement system transmits, time stamps, and stores the packet at the transmitting and the receiving laptops. The stored files are then post-processed to eliminate the effects of differences between the clocks of the transmitter and the receiver laptops and we extract the refined delay jitter measurements. Figure 5.36*b* shows the accuracy of the system (whicj is measured by comparing the results obtained from two separate laptops connected to the same point and receiving the same message form a receiver). The measurement error (difference of measurements in two identical laptops) has a mean of around 0.01 ms and the mean of the measurements is around 1 ms, restricting the measurement error to around 1%. Using a delay jitter distribution, one can simply find the relationship between the packet loss and the jitter compensation buffer length. For any given jitter compensation buffer length, the probability of packet loss is the same as the probability of having a delay jitter larger than the jitter compensation buffer length. Figure 5.37 shows the experimental results for up to five stations operating in the WLAN test bed. If we fix the acceptable PLR to 1%, then the minimum buffer length increases from 0.5 ms to 7 ms when we increase the number of users from one to seven. The details of algorithms for the implementation of the test bed and the results of OPNET simulation for a large number of voice users are available in [Fei99].

QUESTIONS

1. Name two duplexing methods and one example standard that uses each of these technologies.
2. What are the popular access schemes for data networks? Classify them.

3. Name a cellular telephony standard that employs FDMA.
4. Why are guard bands necessary in FDMA?
5. What is binary exponential back-off algorithm, which standard uses that and what is the purpose of using it? What is its weakness?
6. Why are medium access schemes for wireless and wired networks different?
7. What is the difference between the access techniques of the IEEE 802.3 and IEEE 802.11?
8. Why do most PCS standards use TDD and most cellular standards use FDD?
9. Why did FDM lose its popularity to TDM in the PSTN backbone hierarchy?
10. Why did the 2G cellular systems shift from analog FDMA to digital TDMA and CDMA?
11. Name three standards using TDMA/TDD as their access method.
12. What are the advantages of CDMA as an access technique?
13. What is the difference between performance evaluation of voice-oriented assigned-access and data-oriented random access methods?
14. Explain the difference between the effects of power-control on the capacity of TDMA and CDMA systems.
15. In a radio ALOHA network, how does a terminal learn that its packet has collided?
16. What is the difference between the maximum throughputs of ALOHA and Slotted-ALOHA networks? What causes this difference?
17. What is difficulty with implementing CSMA/CD in a wireless environment?
18. Explain the difference between carrier sensing mechanisms in the wired and wireless channels.
19. Explain what is the hidden terminal problem and how it impacts the performance of CSMA based access method.
20. Explain what is the capture effect and how it impacts the performance of random access methods.
21. Explain the differences between integration of data into a voice-oriented network and integration of voice into a data-oriented network.
22. How is priority provided for different classes of traffic in IEEE 802.11e?
23. What makes medium access in sensor networks different?
24. Why is the capacity of an IEEE 802.11 WLAN to carry VoIP calls limited?
25. Explain the relation between the receiver buffer size and packet error rate in voice over IP applications.

PROBLEMS

Problem 1:

To provide public telephone access to commercial ferries a telephone company installs a multi-channel wireless telephone system in a ferry. This wireless radio system connects to a base station on the shore through the air. The base station is connected to the PSTN using wires.

(a) If the telephone company installs a four-channel system what is the probability of having a person come to the telephone and none of the lines are available? Assume that the average length of a telephone call is 3 minutes and 150 passengers of the ferry, on the average make one call per hour.

(b) What is the average delay for accessing the telephones?

(c) How many channels are needed to keep the blockage probability below 2%?

Problem 2:

(a) Neglecting the frequency spectrum used for control channels, what is the maximum number of two-way voice channels that can fit inside the frequencies allocated to the AMPS system with 30 KHz per channel and 25 MHz of total spectrum?

(b) What is the number of channels in each cell where frequency is reused in clusters of $K = 7$ cells? Note that $K = 7$ was originally used in the AMPS.

(c) Repeat (b) for IS-136/IS-54 in which $K = 4$ and number of slots per TDMA channel is three.

(d) Repeat (b) for IS-95 CDMA assuming the minimum required E_b/N_0 is 6 dB. Include the effects of antenna sectorization, voice activity, and extra CDMA interference.

(e) Repeat (d) for wideband CDMA where 5 MHz bands are used in each direction.

Problem 3:

(a) Sketch the throughput versus offered traffic G for a mobile data network using slotted non-persistent CSMA protocol. The packets are 20 ms long and the radius of coverage of each BS is 10 km. Assume the radio propagation speed is 300,000 km/second and use the worse delay for calculation of the "a" parameter.

(b) Repeat (a) for slotted ALOHA protocol.

(c) Repeat (a) for 1-persistent CSMA protocol.

(d) Repeat (a) for a wireless LAN with access point coverage of 100 m.

(e) Repeat (a) for a satellite link with a distance of 20,000 km from the earth.

Problem 4:

A cellular carrier has established 100 cell sites using AMPS with 395 channels and K = 7.

(a) Use the Erlang graphs to calculate the total number of subscribers for a blocking probability of 0.02, average of 2 calls per hours, and average telephone conversation of 5 min.

(b) Use software tools (MatLAB, MathCAD, etc.) to calculate the same values using the Erlang B equation directly.

(c) Determine (either from plot or calculation) the average delay for a call.

(d) Repeat (a) for a blocking probability of 0.01.

(e) Repeat (a) if IS-54 (IS-136) was used with K = 7.

(f) Repeat (a) if IS-54 (IS-136) was used with K = 4.

Problem 5:

We want to use a GSM system with sectored antennas (K = 4) to replace the exiting AMPS system (K = 7) with the same cell sites. In the existing AMPS system the service provider owns 395 duplex voice channels.

(a) Determine the number of voice channels per cell for the AMPS system.

(b) Determine the number of voice channels per cell for the GSM system.

(c) Repeat (b) if we were using a W-CDMA system with the bandwidth of 12.5 MHz for each direction. Assume a signal to noise ratio requirement of 4 (6 dB) and include the effects of antenna sectorization (2.75), voice activity (2), and extra CDMA interference (1.67).

Problem 6:

We provide a wireless public phone with four lines to a ferry crossing between Helsinki and Stockholm carrying 100 passengers where on the average each passengers makes a 3 min telephone call every 2 hours.

(a) What is the probability of a passenger approaches the telephones and none of the four lines are available?

(b) What is the average delay for a passenger to get access to the telephone?

(c) What is the probability of having a passenger waiting more than 3 minutes for access to the telephone?

(d) What would be the average delay if the ferry had 200 passengers?

Problem 7:

A WLAN hop accommodates 50 terminals running the same application. The transmission rate is 2 Mbps and the terminals are using the slotted ALOHA protocol. The commutative traffic produced by the terminals are assumed to form a Poisson process.

(a) Give the throughput versus offered traffic equation for the system and determine the maximum throughput in Erlangs.

(b) What is the maximum throughput in bits per second?

(c) What is the maximum throughput in bits per second for each terminal?

Problem 8:

A local 3 hour tour boat with 50 passengers has one AMPS radio phone to connect to the shore. On average, each user places one call per each tour and the average holding time for the calls is 3 min.

(a) What is the probability that a person attempts to use the phone and he/she finds it occupied?

(b) Repeat (a) if the AMPS phone is replaced by three IS-54 phones using the three slots of the existing IS-54 TDMA system over the same band.

(c) Repeat (a) if this phone is replaced by six upgraded IS-54 phones using 6-slot upgraded IS-54 TDMA over the same band.

Problem 9:

In a datagram packet switched network with

L_p: packet size in bits

N_h: number of hops between two given systems

R_b: data rate in bps on all links

L_o: overhead (header in bits per packet)

T_p: end-to-end propagation delay
N_p: number of packets
L_M: message length in bits
T_{ph}: propagation delay per hop

(a) Give N_p in terms of L_M, L_P, and L_o
(b) Give T_p in terms of L_M, L_P, L_o, N_h, R_b, and T_{ph}.
(c) What value of L_P, as a function of N_h, R_b, and L_o, results in minimum end-to-end delay T_p? Assume that the message length is much larger than the packet size and propagation delay is negligible ($T_{ph} = 0$).

Problem 10:

An ad-hoc 2 Mbps wireless LAN using ALOHA protocol connects two station with a distance of 100 meters from one another each, on the average, generating 10 packets per second. If one of the terminals transmits a 100 bit packet what is the probability of successful transmission of this packet. Assume that the propagation velocity is 300,000 Km/sec and the packets are produced according to the Poisson distribution.

Problem 11:

Let us suppose that the frame transmission in an IEEE 802.11 WLAN follows the process where a station waits for a period called the distributed inter-frame spacing (DIFS), which if clear, results in a backoff of CW time slots, followed by transmission of a frame, a waiting time called the short inter-frame spacing (SIFS) and the transmission of an ACK frame. The DIFS, SIFS, and the slot duration are respectively 50, 10, and 20 μs, and the minimum value of CW is 15. Assume that somehow the transmissions are collision free, there is no idle time outside of the waiting times, each data frame containing packetized voice is 242 μs number of voice octets × 0.73 μs long and the ack packet is 212 μs long. The number of voice octets produced by a station is 20 octets per packet every 10 ms (this means a packet has to be sent by a station every 10 ms).

(a) Compute how many simultaneous voice calls can be supported by one access point. Note that each call corresponds to two transmissions over the same link.
(b) Repeat the calculation if the number of voice octets produced by a station is 40 per packet every 20 ms.

Note that the delay for a call in the latter case increases because packets are transmitted only every 20 ms in the best case. This has an impact on the overall voice quality.

Problem 12:

In a slotted ALOHA wireless network, the received signal amplitude from each terminal forms a Rayleigh distributed random variable with average received power of P. In this problem we examine the throughput of this network using a simple power-based capture model. In the power-based capture model, if there is a collision of $n+1$ packets, the target packet will be captured if its instantaneous power, p_s, is at least z times more than the total power of the n interfering packets, p_n. In words, the target packet is destroyed if

$$\gamma = \frac{p_s}{p_n} < z$$

where z is the capture threshold and γ is the signal-to-interference ratio at the collision.

(a) Show that the probability density functions of the received power from the target user and the interferece power from other terminals are given by

$$f_{P_s}(p_s) = \frac{1}{P} e^{-p_s/P}$$

and

$$f_{P_n}(p_n) = \frac{1}{P} \frac{(p_n/P)^{n-1}}{(n-1)!} e^{-p_n/P}$$

(b) Show that the probability density function and probability distribution function of the signal-to-interference ratio are given by

$$f_\Gamma(\gamma) = n(\gamma+1)^{-n-1}$$

and

$$F_\Gamma(\gamma) = 1-\left(\frac{1}{\gamma+1}\right)^n$$

(c) Show that if the arrivals of the packets obey a Poisson distribution, the throughput of the system is given by

$$S = G\left[1-\sum_{n=1}^{\infty}\frac{G^n}{n!}e^{-G}F_\Gamma(z)\right] = Ge^{-Gz/(z+1)}$$

where G is the offered traffic.

(d) Use MathCAD or MatLAB to sketch S versus G for the capture parameter values of $z = 0$, 3, 6, 20, and ∞ (no-capture) dB.

PROJECTS

Project 1:

Use the equations given in the paper by Budka et al. to reproduce Figures 6 and 8 in the paper. See reference [Bud97].

Project 2:

Wireshark is a free sniffer/network protocol analyzer for different operating systems. It is available as a free download at http://www.wireshark.org/. Read the documentation to understand its basic operation. Make sure that you are allowed to sniff a network you are connected to or use it within your home network. Capture packets using the promiscuous mode on the default interface on the computer on which you have installed wireshark. Make a list of MAC addresses that you see on the network. Can you identify the devices having these MAC addresses?

PART TWO

WIDE-AREA NETWORKS

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

6

THE INTERNET

6.1 INTRODUCTION: INTERNET INFRASTRUCTURE

The Internet is the public network connecting numerous smaller computer networks together to create a platform for worldwide packet switching using the IP. The Internet allows implementation of applications running over millions of smaller networks owned by government, private sectors, academic institutions, and industrial organizations.

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

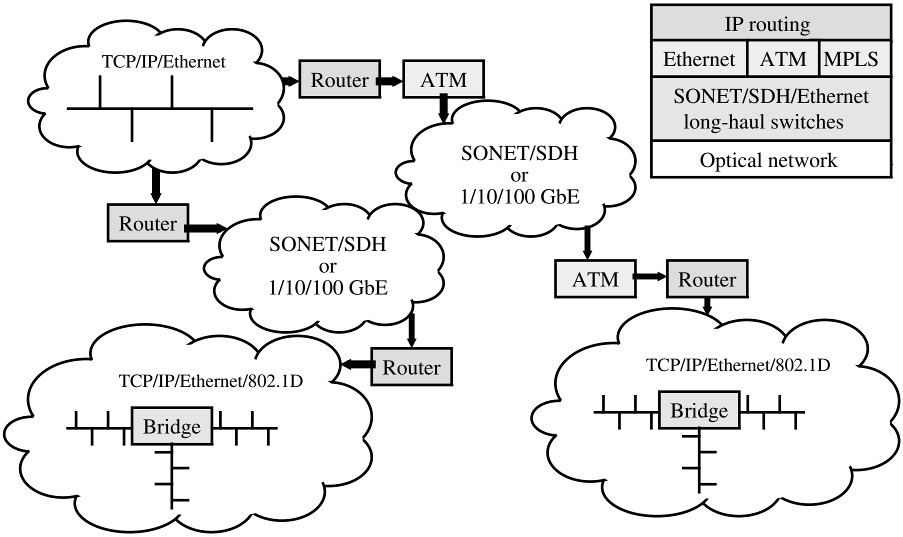
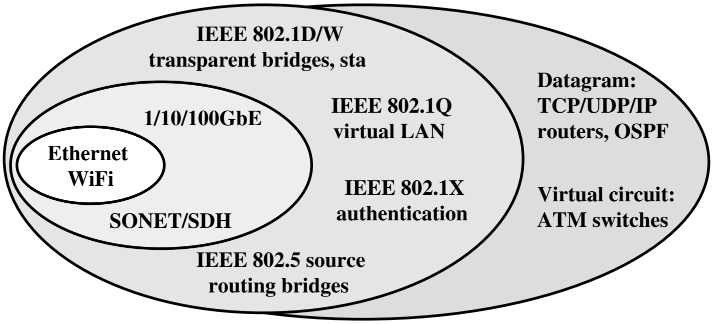


FIGURE 6.1 Overview of the Internet infrastructure.

The physical infrastructure of the Internet uses bridges, switches, and routers to interconnect networks and a number of protocols describe how data can be transferred over the network. The Internet is defined by its interconnections and routing policies. Figure 6.1 shows a simplified description of the Internet infrastructure and its relation to interconnecting devices. One can view the entire network as a number of LANs connecting in local areas through bridges and connecting in wide areas through switches and routers.

Considering Fig. 6.1, the first natural question which comes to mind is "why so many options to interconnect networks?" The simple answer to this question is that "no one was able to solve all the issues" or maybe "the networking industry evolved that way." Long-haul telcommunication TDM switches for voice traffic in PSTN came first. Then ISDN and ATM switches came to integrate the data with the voice. These switches are fast hardware equipment designed for connection-based services with support of QoS, but service per user is expensive. ATM was designed to answer our question but it failed to integrate everything and that was the last attempt of telecommunication industry. In the computer communication industry, bridges, which are also called LAN switches, interconnect LANs using hardware MAC addresses which do not scale well, but they are fast and inexpensive. Routers scale well using IP software addresses and can handle data traffic at reasonable price for long-haul communications, but still they are more expensive than LAN bridges and they cannot guarantee the quality of the received messages. Therefore, in this chapter we address all three interconnect technologies, because each addresses a separate technical aspect of interconnects for communication networking. Before we discuss bridges, switches, and routers, we have two sections devoted to addressing techniques and the Quality of Service (QoS) to prepare the reader for a better understanding of the details of these interconnecting devices.

Figure 6.2 illustrates an overview of the important technologies related to the Internet. The core of the network is wired Ethernet and wireless WiFi LANs. This core is connected or



OSPF: Open shortest path first
STA: Spanning tree algorithm

FIGURE 6.2 Overview of the technical aspects of the Internet infrastructure.

extended to wide areas using SONET/SDH or gigabit Ethernet technologies. Local networks are interconnected with popular transparent bridging and diminishing source routing bridging technologies with some additional features for security and virtual LAN (VLAN) operations. All these local connection technologies are specified by the IEEE 802 community, which originally started regulating LANs around 1980. Long-haul connections are made either by routers using a datagram or by ATM switches using virtual circuit switching. There are two algorithms frequently used for interconnection: the spanning tree algorithm (STA) for bridges and open shortest path first (OSPF) for routers. In the rest of this chapter we go over the details of these technologies. We first pay special attention to the two important issues of addressing and QoS to differentiate these aspects from each other. Then we provide the details of bridging, switching and routing technologies. The interested reader is referred to textbooks, e.g. [Pet07, Kur01, Tan03], for more details of the elements and protocols that make up the Internet.

6.1.1 Fundamentals of Packet Forwarding

Figure 6.3 shows the basic principles of packet transmission over a network. Packets from terminal A are destined to terminal B through a set of interconnecting elements to network the transmission media. Each interconnecting network has a few options to

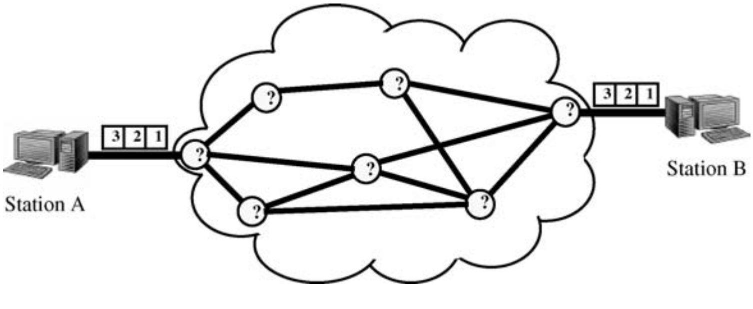


FIGURE 6.3 Packet forwarding problem.

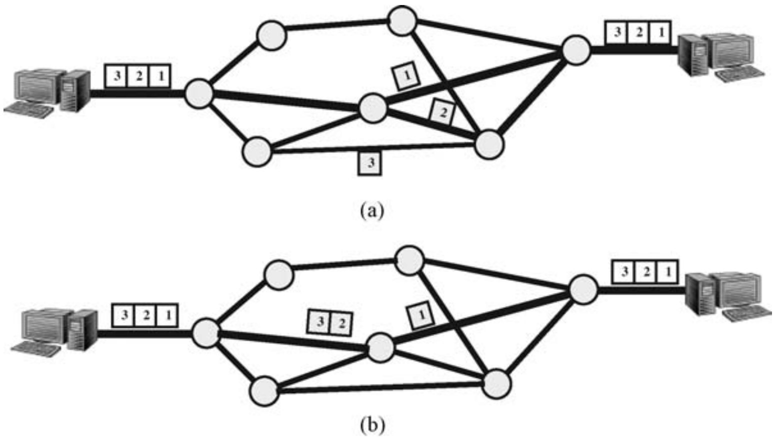


FIGURE 6.4 Packet switching approaches: (*a*) connectionless datagram; (*b*) connection-based virtual-circuit.

forward the packets, raising two fundamental questions for the design of the interconnecting elements:

1. Who should decide what the appropriate route for the packet is: the terminal or the interconnecting elements?
2. Should all packets from one user go through the same route or each may take a different route?

In bridges, switches, and routers used for traditional and popular LANs and WANs, such as Ethernet, PSTN, and Internet, decision making for packet forwarding is transparent to the terminal. These techniques are generally divided into datagram and virtual circuit switching techniques. Figure 6.4 illustrates the basic concept of datagram and virtual circuit packet switching. In datagram transmission, used for connectionless transmissions in Ethernet bridges and Internet routers, each packet finds its own route and they are reassembled in the last stage of transmission in the proper order. In virtual circuit switching, used in connection-based networks such as ATM, the network first defines a route for the packet delivery during the connection phase and then all packets follow the same route while the connection is alive. In the IEEE 802.5 and some of the adhoc and sensor networks the terminal is involved in the routing of the packet. The particular technique used in IEEE 802.5 is referred to as source routing, and bridges operating based on this principle are referred to as source routing bridges.

In the rest of this chapter we address popular technologies for bridges, switches, and routers. We start with two important issues concerned with networking, namely addressing and QoS handling in connection-based and connectionless networks and then go over the details of bridges, switches, and routers.

6.2 ADDRESSING

Modern networks have evolved around connection-based voice-oriented PSTN and the data-oriented connectionless Ethernet/Internet networks. Addressing in these two networks is quite different: PSTN is a circuit-switched network and Ethernet/Internet is a packet-switched network. In circuit-switched environments, the network assigns a virtual path for transmission during the call establishment and all the traffic packets are transmitted through that path. Therefore, the address identifies a destination which is used during call establishment. In a packet-switched network we have MAC addresses for LANs and local networking and IP addresses for WAN operation. Wireless access to these networks adds a new dimension of complexity to this environment, because connection to the network is mobile and it connects to the network from different APs or BSs. Another dimension of complexity is the handling of multicast and broadcast. LANs have embedded multicast and broadcast features which are desirable for a number of modern applications. To extend these features to the Internet of PSTN we need to find new approaches for scaling these features. In the rest of this subsection we provide an overview of the issues involved in addressing with more details of how these issues are handled in a modern network.

6.2.1 ISDN Addressing in Connection-Based PSTN

The legacy connection-based circuit-switched network is the PSTN, which was basically used for voice application and later on in ISDN to provide an integrated voice and data service. ISDN services never gained the popularity that was expected; however, digital cellular networks used a modified version of this technology for implementation of cellular networks with an integrated service. ISDN addressing is the core for this type of communication. Virtual packet switching networks, such as X.25 and ATM networks, also have their own addressing schemes. In the rest of this subsection we provide a fundamental overview of the addressing techniques used in these networks.

In connection-based networks a terminal needs to forward its message to a destination terminal; for that reason, it passes the destination address to the first node in the network so that the network can find the best route and establish the connection, as shown in Fig. 6.4*b*. This process involves five steps:

- *setup time*, in which the terminal indicates to the network that it needs to establish a connection with a given address;
- *call processing*, in which the route is determined and the connection to the destination terminal is established;
- *alerting*, in which the network informs the calling party that the destination terminal is ringing;
- *connecting*, in which network sends back a message to the calling party indicating that the intended destination has answered the call;
- *release*, in which either the source or the destination indicates that the call is to be terminated.

Cellular telephones such as GSM use connection-based networking and ISDN addressing. Figure 6.5 shows the general format of the ISDN addresses. The international ISDN number carries a country code assigned by the ITU, telecomm authorities in different countries distribute the number groups to cities and services with a national destination, and service providers assign the subscriber number to the end users. The international ISDN number has a maximum length of 15 digits.

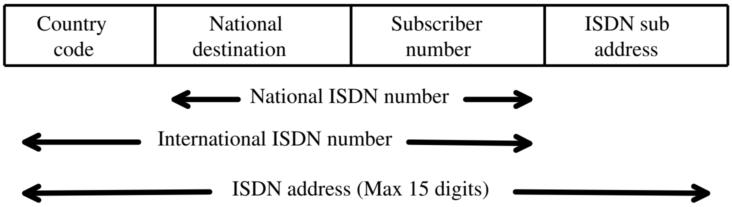
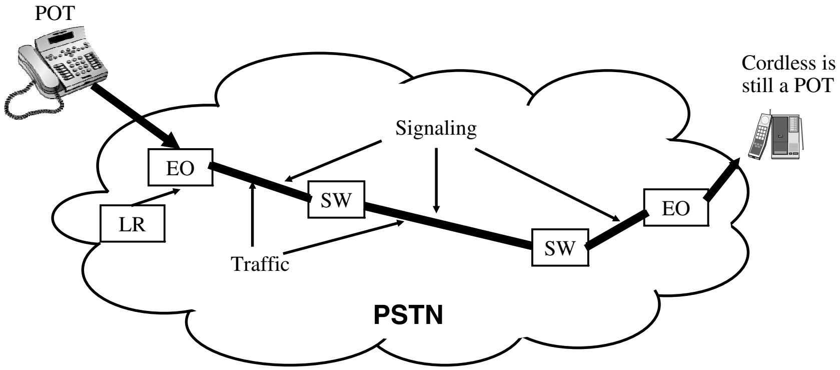


FIGURE 6.5 General format of an ISDN address.

Example 6.1: ISDN Numbering The ISDN number 358-40-525-5436 is a mobile number for which 258 specifies Finland, 40 specifies Telia, the mobile operator in Oulu Finland, and 525-5436 is the subscriber number. An ISDN address consists of an ISDN number plus a sub-address with a maximum length of four digits. The sub-address is transparent to the public network and can be used for implementation of additional private features, such as an extension number.

In PSTN, one can track an ISDN address geographically to the home or office location and the network uses the number to route the call using signaling messages among the interconnecting elements of the network. Figure 6.6 shows the details of a telephone call in the PSTN. When the telephone terminal user picks up the phone and dials the destination number, the end office, which is the last switch in the network connected to the phone, uses the signaling network and the ISDN destination address to establish the route. Then traffic is transferred through the traffic channel. The number is stored in a location registration database at the network for billing purposes. In cellular telephones this number is stored in the mobile terminal and registered in an interconnecting device of the network. Figure 6.7 shows an overview of mobile ISDN addressing used in cellular networks. The interconnecting device, which is referred to as an MSC, has two databases: one to keep the address of home users and one for the visitor users. A mobile user is registered in a home location database by its permanent address. When the mobile moves



POT: Plain old telephone
LR: Location registration data base
EO: End office switch
SW: Telephone switch

FIGURE 6.6 ISDN addressing connection-based PSTN.

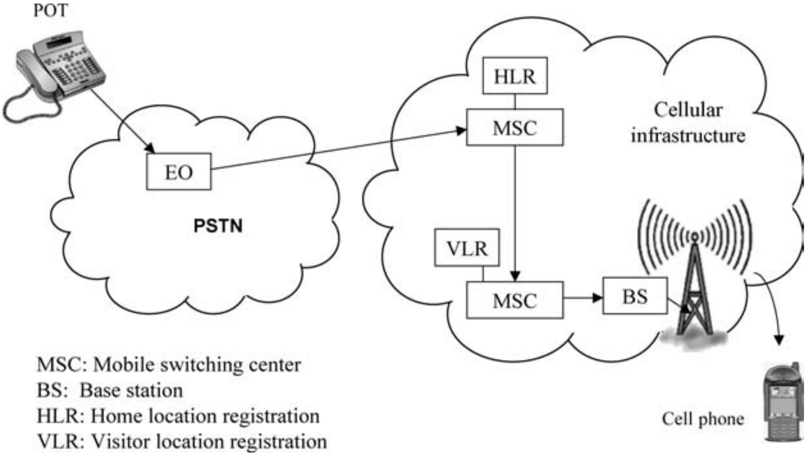

POT
PSTN
EO
HLR
MSC
Cellular infrastructure
VLR
MSC
BS
Cell phone
MSC: Mobile switching center
BS: Base station
HLR: Home location registration
VLR: Visitor location registration

FIGURE 6.7 Mobile ISDN addressing in PSTN and cellular networks.

from the home to connect as a visitor to another part of the PSTN it receives another address from the visitor database. The visitor and home databases communicate with one another to allow call forwarding to the new location. In cellular phones, the actual address is not broadcast on air. Each time that a user connects to the network it receives a temporary address for communication through the air. This is to protect the user from fraudulent connections.

Virtual packet switching data networks such as X.25 or ATM, discussed later on in this chapter, have their own addressing format for packets to travel between the switches. In X.25, each packet has a 12-bit virtual circuit identifier divided into 8 bits of logical channel number and 4 bits of logical channel group number. ATM packets use 12 bits for virtual path identifier and 16 bits for channel identifier. Figure 6.8 shows the relation among virtual channels, virtual paths, and the physical medium in an ATM network. Each of the channels and paths is addressed with a unique number.

6.2.2 MAC Addressing in Connectionless Local-Area Networks

Connectionless networks carry the information using datagram packets. In a datagram we expect each packet to carry the source and destination address so that an

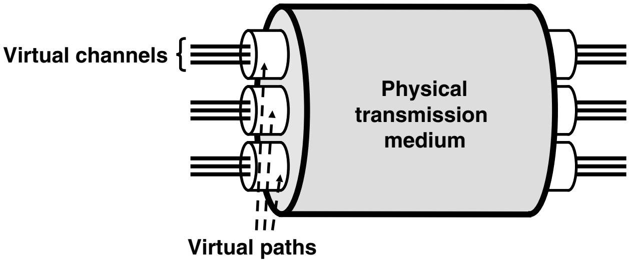


FIGURE 6.8 Virtual channels and virtual paths in an ATM virtual packet-switching network. Each channel or a path is identified by an address.

interconnecting element, such any bridge or a router, can forward the packet towards its destination. Two addresses have evolved for datagrams: the MAC address introduced for LANs and the IP address introduced by the Internet community. The MAC or layer 2 or Ethernet hardware address is a unique address identifier for network interface cards (NICs) to connect to LANs.[1] This addressing technique was originally introduced in the Ethernet and it uses 48 bits[2] assigned to manufacturers of the NICs by the IEEE organization.

Example 6.2: The MAC Address The MAC address 00:00:5e:67:35:60 is represented in 12 hex letters each carrying 4 bits. The first three pairs of octet numbers represent the manufacturer and the next three the serial number. This 48-bit address space contains potentially 2^{48} or 281 474 976 710 656 possible MAC addresses. An NIC is uniquely identified by a MAC during the manufacturing process, and products of a manufacturer are randomly distributed all over the world. As a result, the MAC address uniquely identifies the hardware wherever it is, but it does not carry information on geographical location of the terminal.

Compared with ISDN addresses, the maximum 15 digit ISDN number is more than three times the potential 2^{48} possible MAC addresses. The ISDN number, assigned by the service provider, carries geographical information, but MAC addresses, assigned by the manufacturer, do not. MAC addresses are enabled for multicast and broadcast in addition to traditional *unicast* addressing. In *multicast*, a single packet transmitted from a source can address a number of destinations in the network, and in *broadcast* mode that packet arrives at all destinations in the network. Broadcast and multicast features utilize network infrastructure efficiently by requiring the source to send a packet only once when it is delivered to a large number of destinations. The nodes in the network take care of replicating the packet to reach multiple destinations only when it is necessary. A MAC multicast or broadcast address is used by the source and the destination terminal to exchange information. These addresses are useful for implementation of user groups in a LAN.

The MAC address for unicast operation has a zero as the least significant bit of the most significant byte of the address. If the least significant bit of the most significant byte is set to a 1 it is a multicast address which reaches several destinations enabled for multicasting with that number. Packets sent to a multicast address are received by all stations on a LAN that have been configured to receive packets sent to that address. Packets sent to the broadcast address carry an all-1 address. All terminals in a LAN receive these packets. In hexadecimal the broadcast address would be "FF:FF:FF:FF:FF:FF." Ethernet and other IEEE LANs, IEEE 802.11, Bluetooth, FDDI, and several other networks use MAC addresses.

Figure 6.9 shows the overall frame format for the IEEE 802.3 Ethernet and most other LANs. Before adding the MAC header and trailer we have the LLC header which also carries source and destination addresses each 8 bits long. The addresses identify the so-called logical service AP, allowing alternative services for logical end-to-end transmission links over the MAC of a LAN. LLC defined by the IEEE 802.2 allows connection-based and

[1] Using a MAC spoofing technique it is possible to change the MAC address on most hardware designed recently. In this approach, a locally administered address is assigned to a device by a network administrator to override the burned MAC address.

[2] An unpopular 16-bit option for MAC address also exists.

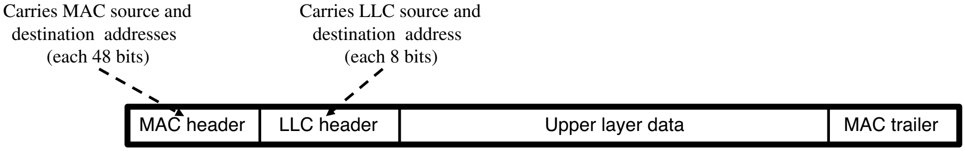

Carries MAC source and destination addresses (each 48 bits)
Carries LLC source and destination address (each 8 bits)
MAC header
LLC header
Upper layer data
MAC trailer

FIGURE 6.9 LLC and MAC overhead.

connectionless services with and without acknowledgement. The connectionless services allow another layer of multicast and broadcast to the terminal. In practice, all products use unacknowledged connectionless services. The two layers of connectionless LLC and MAC addressing, each with three options for unicast, multicast, and broadcast, allow nine different addressing options in a LAN.

In a wireless local network, we need two MAC addresses when we need to address a WLAN: one to identify the AP and the other to identify the wireless terminal. As we will see in Chapter 12 on 802.11 WLAN, the MAC frame of this standard has four address fields that are used to identify the AP and the mobile terminal. When a mobile node with a WLAN connection moves from coverage of one AP to another it keeps its own MAC address, but the MAC address of the AP will change. This mobile MAC addressing mechanism allows implementation of local roaming among different points of connection for a terminal using a WLAN.

6.2.3 IP Addressing in the Connectionless Internet

The IP or layer 3 or Internet software address is an address identifier for an electronic device allowing it to network with other devices through the Internet. All network devices, such as routers, switches, computers, printers, and IP telephones, have their own address that is unique within the scope of the specific network. IP addresses are created and distributed by the Internet Assigned Numbers Authority (IANA) which allocates super-blocks to different regions to be divided into smaller blocks for different ISPs and enterprises. The small blocks of IP addresses are then assigned dynamically by the network administrator of each organization to the individual terminals. As a result, part of the IP address is used to locate an IP device and from that device one can find the target device and interact with it. Therefore, similar to ISDN addresses, IP addresses have a hierarchical geographically selective structure which helps in routing the message toward the destination address through a number of independent interconnecting elements. However, unlike ISDN used in fixed or mobile telephone, they do not identify a fixed connection to a location or a specific mobile terminal. Compared with MAC addresses, IP addresses provide geographical selectivity, whereas the geographical distribution of MAC addresses is random; and an IP address is a software address that is changed, whereas a MAC address is burned into the NIC hardware. The popular IP version 4 (IP-v4) uses 32-bit addresses, which are usually represented as four decimal number equivalents of the four eight-digit blocks.

Example 6.3: IP Addressing The IP address 130.215.10.14 represents 10000010, 11010111, 00001010, 00001110. A 32-bit address results in close to 4 billion ($2^{32} = 4\,294\,967\,295$) different possibilities. Considering that, at the time of this writing, we have around 3 billion cell phones and more than a billion fixed Internet connections, this number

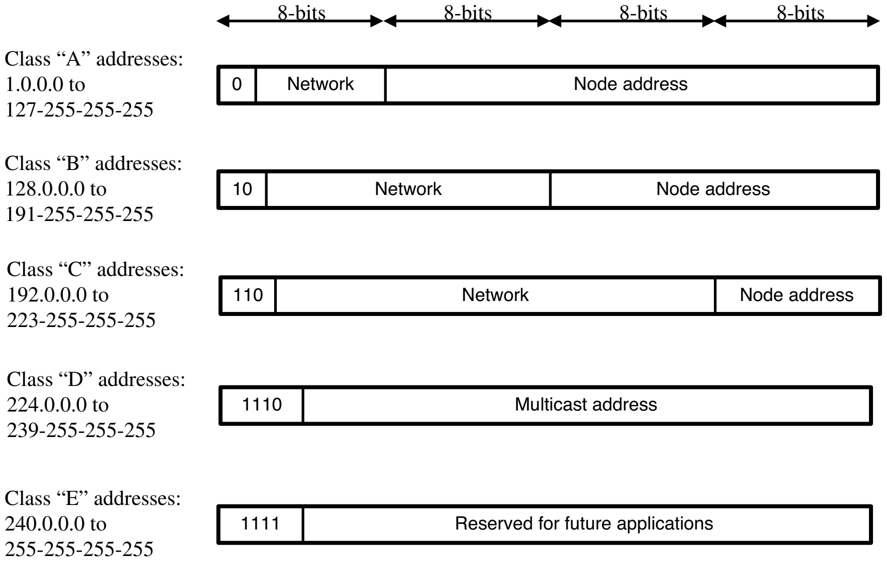


FIGURE 6.10 Overall structure of an IP address and different classes.

is small and we need a longer address. IP version 6 (IP-v6), with 128-bit addresses, was introduced to solve this problem.

IP addresses have two parts, which identify the network and the node in the network. There are five coded classes of addresses and a subnet mask to differentiate the network address from the node address. The class of the address and the subnet mask determine which part belongs to the network address and which part belongs to the node address. The first 4 bits can be used to determine each class. Figure 6.10 shows the five classes of address for IP-v4 and different fields for each class. Using Fig. 6.10 one can determine the class of the address and from which network and node address for classes A, B, and C.

Example 6.4: Parts of IP Addresses and the Loopback IP Address In the IP address 130.215.10.14, the first three digits "130" indicate that this is a class B address and by default the first two sets 130.215 represent the network address and the second set 10.14 the address of the terminal. In other words the address for the main router in the network or the network address is 130.215.0.0. If we set the node address to all "1," which means 130.215.255.255, this signifies a broadcast for that network, which delivers the message to all nodes of the network. Addresses beginning with 127 (01111111) are reserved for loopback and for internal testing on a local machine. For example, if you ping 127.0.0.1 it always works because it points at yourself.

Figure 6.11 shows an overview of a typical network and its relation to IP addressing. A network such as a university campus consists of a number of sub-networks, each representing a department in a university for example. Each sub-network has an Ethernet LAN which connects a number of nodes to a router. The subnet routers are connected to the main router. In each Ethernet segment, packets from each node are accessible by all other nodes. Subnets in an IP network preserve the address space of each Ethernet and their associated securities. In addition, the subnet provides for network traffic control. To provide for implementation of this hierarchy in a network, the node address part of the IP address is

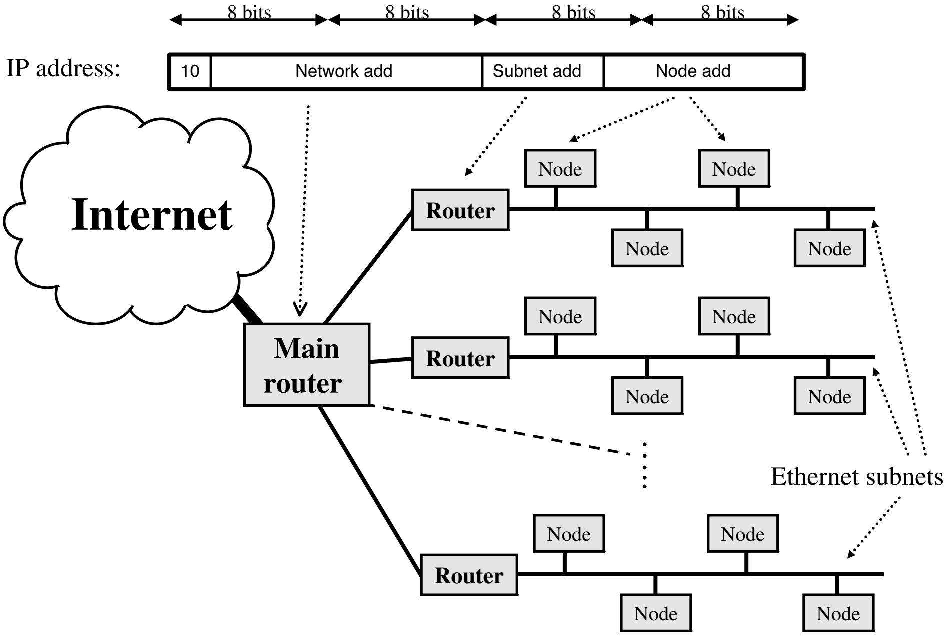


FIGURE 6.11 Overall structure of an IP address and different classes.

further divided into subnet address and node address within the subnet. The main router masks the received IP addresses by performing an AND operation between the received address and a mask number which has all 1s for the network address bits and all 0s for the node address bits. If the result shows the network address, then it accepts the packet. Subnets do a similar operation, but for a mask differentiating the subnet address from the node address in the subnet.

Example 6.5: Subnet Masks for Different Classes of IP Addresses The default subnet masks for class A, B, and C addresses are

Class A	255.0.0.0	(11111111.00000000.00000000.00000000)
Class B	255.255.0	(11111111.11111111.00000000.00000000)
Class C	255.255.255.0	(11111111.11111111.11111111.00000000)

A typical subnet mask for class B addresses using 3 bits for the subnet to address up to seven subnets is 255.255.224.0 (11111111.11111111.11100000.00000000). To represent this submask the address is represented by 130.215.10.14/19, in which the number after the slash represents the length of the mask in bits.

IP multicast provides for one-to-many messaging; this is similar to MAC multicast, but it takes place in an IP infrastructure. This approach utilizes the network infrastructure efficiently by allowing the source to send a packet only once while it is delivered to a large number of destinations. The nodes in the network manage the packet switching so that it arrives in multiple destinations without any prior knowledge of the specific destinations. IP multicast is implemented based on the IP multicast group address, a multicast distribution tree which is created by the destination terminals. The IP multicast group address is used by sources and the destination terminals in different ways. A source terminal uses the group address as the IP destination address. The destination terminal uses the group address to inform the network nodes that it is interested in accepting the broadcast. IP multicast is ideal for applications such as distance learning, where a number of widely distributed users demand a wideband service, such as the class video. Implementation of IP multicast requires much more complex operation at the interconnecting element, and its complexity increases as the problem scales to a huge number of users.

Consider one of the Ethernet LANs in Fig. 6.11; each node is identified by a MAC address in the LAN and with an IP address in the network. IP addresses are not permanent, but MAC addresses are permanent. How can we keep track of the mapping between the two addresses? The simplest approach that comes to mind is that we keep a table in the router which maps the MAC addresses to the IP addresses; but updating such a table is difficult and because when we need the mapping we are not certain that the table has been updated. The more common approach is using the so-called address resolution protocol (ARP), which broadcasts a question in the LAN asking who owns the IP of the arriving packet. The terminal that has the address releases the MAC address to the network. The ARP is used for implementation of the mobile IP operation.

Unlike mobile addressing for cellular networks, mobile IP is not designed only for wireless connection to a network. Mobile IP allows connection to the network from any place in the network. By definition, mobile IP is an IETF standard communications protocol that is designed to allow mobile device users to move from one network to another while maintaining a permanent IP address. Figure 6.12 shows the basic elements for implementation of the mobile IP protocol. We have a home network in which a mobile host is

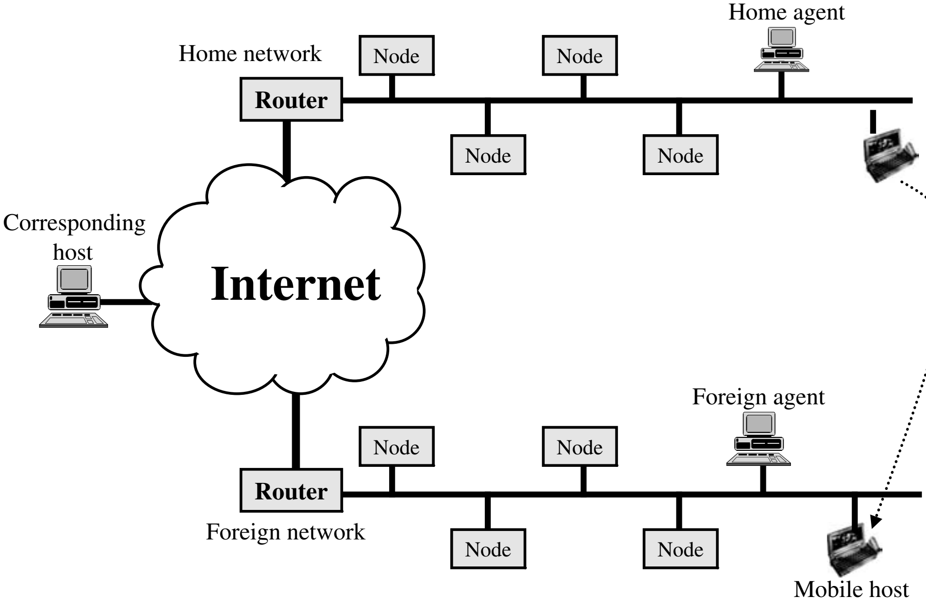


FIGURE 6.12 Basic elements for implementation of mobile IP.

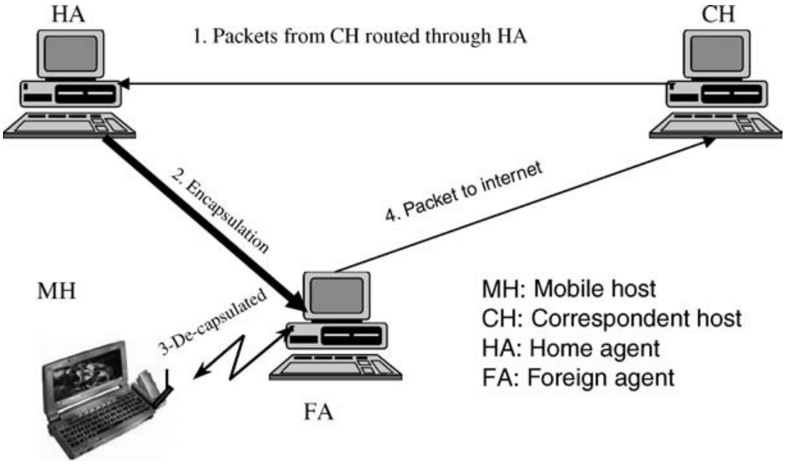


FIGURE 6.13 Basic principle of mobile IP operation for a wireless laptop.

registered and its IP is taken, there is a visiting or foreign network to which the mobile host desires to connect, and a corresponding host which desires to send a message to the mobile host with the address taken from the home network not knowing that the mobile host is currently connected to a foreign network. To enable the network to provide this type of mobility to the host terminal, the IETF recommends addition of two agents in the home and the foreign networks. The mobile host has two addresses: a *permanent address* for the home network and a *care-of address* associated with the visiting foreign network. The home agent stores information about the mobile host and its permanent address in the home network. The foreign agent stores information about the mobile host visiting its network and advertises the care-of address.

Figure 6.13 shows how a terminal sends a packet to a wireless mobile terminal registered at home network but visiting the foreign network. The corresponding host sends its packet for the mobile host using the home address. This packet is intercepted by the home agent, which uses a table to tunnels the packets to the mobile host's care-of address. The tunneling operation adds a new IP header with the care-of address while keeping the original IP header in the packet. Upon arrival of the packet at the foreign agent the packet is decapsulated at the end of the tunnel to remove the added IP header and then delivered to the mobile host. When a foreign mobile host wants to send a packet it simply sends the packet directly to the corresponding host through the foreign agent.

6.3 QUALITY OF SERVICE

As a packet is delivered to the network, until it arrives at the destination, depending on the type of network, a number of impairments occur which affect the quality of the received information. The quality of the received information is usually measured by a number of characteristics: the throughput or maximal data transfer rate or bandwidth that can be sustained between two end points, e.g. average rate, peak rate, minimum rate; t he delay time of the packet to transport from the source to the destination caused by transmission facilities and interconnecting elements; variation in end-to-end transit delay, which is

TABLE 6.1 Quality Requirements for Popular Applications

Application	Bandwidth	Delay	Jitter	Reliability
E-mail	Low	Low	Low	High
File transfer	Medium	Low	Low	High
Web browsing	Medium	Medium	Low	High
Audio streaming	Medium	Low	High	Low
Video streaming	High	Low	High	Low
Telephony	Low	High	High	Low
Video conferencing	High	High	High	Low

referred to as delay jitter; the reliability of the network in delivering the packets or packet loss, which is the ratio of the number of undelivered packets to the total number of sent packets. Different applications have different sensitivities to these performance measures. Table 6.1 shows the relation among these performance measures and a number of popular applications. The access methods and interconnecting elements of different networks have evolved around specific applications which were the main reason to generate income in an industry.

6.3.1 Quality of Service in Connection-Based Networks

In the connection-based circuit-switched telecommunication industry, QoS was sometimes defined as the ability of a network to have some level of assurance that its traffic and service requirements can be satisfied. To support QoS, all layers and every network element should cooperate to guarantee a number of features, such as time to provide service, voice quality, echo level, and connection loss rate. A subset of telephony QoS is the grade of service requirements related to blockage and outage probability. The term QoS is sometimes used as a quality measure rather than referring to the ability of the network to reserve resources. The QoS is not the same as the ability of the network to support high bit rate, low error rate, and low latency. It is the ability to support a certain level of assurance for traffic requirements. To define the level of QoS, the telephone and video industry uses subjective measures based on user-perceived performance and mean opinion scores, which reflect the cumulative effects of all system imperfections affecting the service. This approach uses collective human opinion in the assessment process to determine objective measures such as delay, throughput, and error rate for designing the network. As an example, the wired telephone network industry required measures such as 1% packet loss and 150 ms end-to-end delay for telephone services to satisfy mean opinion of scores. The switches and transmission facilities in the PSTN are designed to support this level of QoS for wired telephony application using 64 Kbps PCM-encoding techniques. When it came to cellular phone applications, the limitation of bandwidth forced the industry to resort to different encoding techniques with data rates around 10 Kbps. The lower data rate and the fading characteristics of the radio channel reduced the expected QoS from the wireline quality traditionally used in PSTN over the past century. This experience showed that the religious adherence to QoS specifications can be compromised. To integrate data application into connection-based networks, ATM technology evolved, which has an elaborate framework to plug in QoS mechanisms and will be discussed later on in this chapter. Recently, ATM has been increased used for carrying live professional uncompressed video and audio in environments demanding low latency and very high QoS, such as in the professional media production industry. However, popular commercial applications tend to delve further into the connectionless Ethernet/Internet environment for most applications.

6.3.2 Quality of Service in Connectionless Networks

In connectionless packet-switched networks for computer networking, QoS refers to resource reservation control mechanisms rather than the achieved service quality. This is due to the fact that the packet-switched networks were originally designed for connectionless packet-based data applications, such as file transfer, Telnet, or e-mail, for which user satisfaction is much less sensitive to the continual QoS needed for two-way telephone conversations. Data applications recover the packet loss at higher layers, and in most legacy applications they are not that sensitive to the delay. As a result of this situation, as we discuss in Chapter 11, the legacy LANs were designed without any provisioning for QoS. The Ethernet MAC packets do not have any field for priority assignment and the MAC mechanism is contention based, having no control on the delay. Other legacy LAN technologies, such as FDDI or token ring, which had priority bits in their MAC packets, were designed to differentiate users (e.g. giving higher priority to the network manager), but not to ensure a certain level of quality to the users and they were never used in practice. Indeed, early bridges connecting LANs were eliminating priority assignments when they were connected to long-haul networks. When the Internet was first deployed it was not designed to support QoS guarantees, due to the limitation in router computing power, and it operated at "best effort." Although the IP addressing header has allocated four "type-of-service" bits in each message, which are similar to priority bits in some of the legacy LANs, these bits were also ignored in practice. As real-time streaming multimedia applications such as VoIP and IP-TV complemented traditional Ethernet/Internet applications, type-of-service bits in IP and other available resources in LANs were exploited for traffic provisioning. The need for QoS in Ethernet/Internet networks became more important when streaming became popular in resource-limited cellular networks. In summary, in packet-switched networks, QoS is the ability to provide a different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow.

In the packet-switched environment of the Ethernet/Internet, the QoS can be provided by overprovisioning a network so that interior links are considerably faster than access links. With the emergence of broadband services, this relatively simple approach is practical for many applications, such as audio and video streaming for which high jitters can be compensated with large buffers or VoIP under low load. Under high traffic load conditions, the delivered QoS for applications such as VoIP degrades significantly, and the use of a QoS mechanism in the network would allow significantly higher user satisfaction and a more balanced traffic among subscribers. Application of QoS mechanisms becomes more important as we encounter lower bandwidth wireless applications, such as satellite IP [Kot04]. To support QoS in packet-switched networks, in the early days the integrated service "IntServ" philosophy of reserving network resources was used. In this model, applications used the resource reservation protocol (RSVP) to request and reserve resources through routers in the network. Using these methods, core routers for a major service provider needed to accept, maintain, and tear down numerous reservations (and maintain this information for many flows). This did not scale well with the rapid growth of the Internet and the notion of designing core routers at the highest possible packet switching rates. As a result, a second approach called differentiated service "DiffServ" emerged as an alternative. In the DiffServ approach, packets are marked according to the type of service they need. Routers supporting DiffServ use these marks to form multiple queues for different priorities.

In the Internet part of the WAN it is the traffic class bits and in the LAN it is the priority bits that are used to mark the packets. Router manufacturers provide different capabilities for configuring this behavior, to include the number of queues supported, the relative priorities of queues, and bandwidth reserved for each queue. Packets carrying VoIP, for example, are assigned to the highest priority queues; control packets will receive a default portion of the bandwidth and the leftover bandwidth is allocated to best-effort traffic.

6.4 BRIDGES

A bridge is a device that connects two networks using the same or different data link protocols. Since in the Open Systems Interconnection (OSI) model the DLL is the second layer of the protocol stack, they are sometimes referred to as layer 2 devices. One of the early applications of bridges was to connect two separate segments of a network, such as two LANs in two different buildings of a building complex, with a point-to-point link. Another traditional application of bridges was to connect two different LANs, such as an Ethernet and a token ring, as shown in Fig. 6.14. These traditional bridges were hardware devices operating at LAN speed using the MAC address (rather than IP address) to direct the packet to the appropriate segment of the network. Although the function of a bridge was to connect separate LANs together, it was also used to connect two LANs using the same protocol. Bridges use the MAC addresses of the nodes residing on each LAN

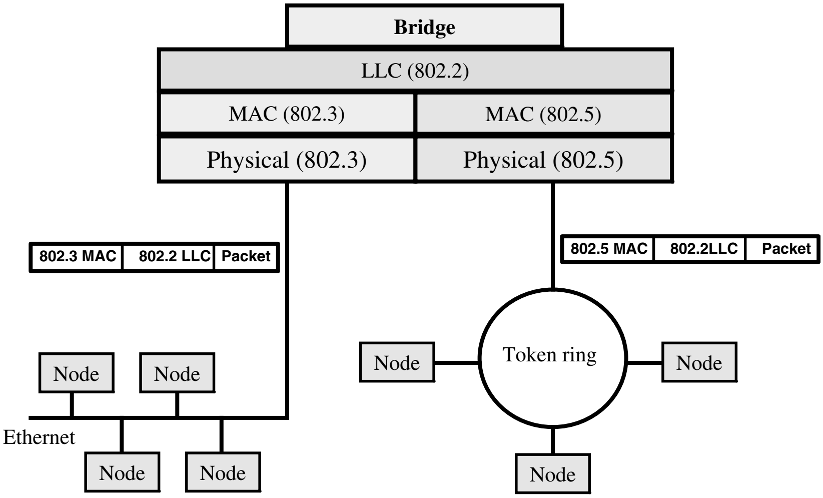


FIGURE 6.14 A legacy bridge connecting an Ethernet LAN to a token ring LAN.

segment of the network and allow only the necessary traffic to pass through them. A bridge can also filter out certain traffic and prevent it from passing through. When a packet is received by the bridge, the bridge determines the destination and source segments. If the segments are the same, then the packet is filtered (dropped); if the segments are different, then the packet is forwarded to the appropriate segment. In addition, bridges prevent unnecessary and corrupted packets from spreading throughout the network by dropping them. For example, a token-ring packet carries priority and has token packets for operation of the LAN which are not used in the Ethernet. A bridge connecting a token ring to an Ethernet LAN eliminates the priority bits and prevents token packet entering the Ethernet segment of the network. Bridges improve the overall throughput by distributing the shared traffic into different shared domains, increase security and resistance to failure by partitioning the traffic, and increase overall throughput and geographical coverage by using several LANs. A LAN switch is a *bridge* connecting more than two LAN segments together preventing unnecessary traffic entering each segment. The terms LAN bridge and LAN switch are sometimes used interchangeably addressing the same entity.

A bridge or a LAN switch receives packets on one port and retransmits them on another port and it does not start retransmission until it receives a complete packet. As a result, stations on either side of a bridge can transmit packets simultaneously without causing collisions. A fixed bridge directs every packet toward all ports. A learning bridge examines the destination address of every packet to determine to which port the packet should be directed using a table which has been build by analyzing previous arriving packets in different ports. This approach to bridging increases efficiency because the bridge avoids unnecessary transmission of the packets. Learning bridges are very common in industry. Another approach to switching used in the IEEE 802.5 token ring LANs is source route bridging, in which the source terminal identifies the path of the packet.

WLAN APs are sometime referred to as wireless bridges. An AP connects multiple users on a WLAN to each other and to a backbone Ethernet. As shown in Fig. 6.15, a

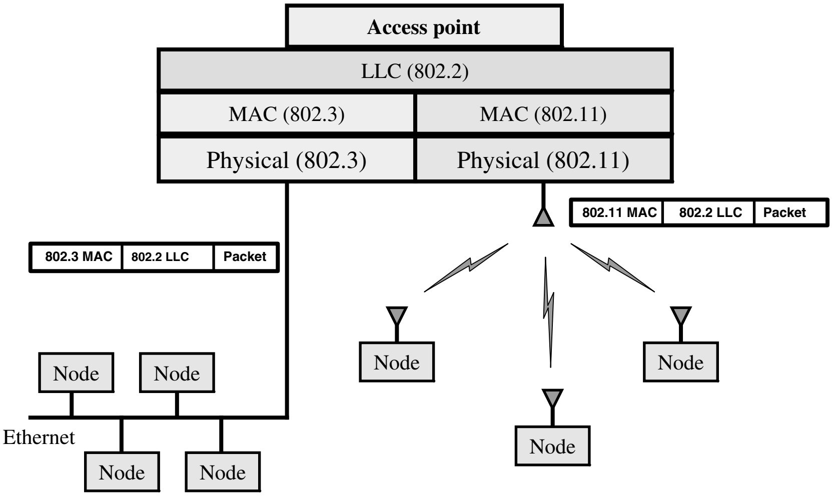


FIGURE 6.15 An AP connecting a WLAN to the Ethernet backbone.

single 802.11 NIC in the AP is connected to the backbone Ethernet network and each WLAN user has access to the Ethernet network and to each other through the AP. The AP here is also a bridge, but it interfaces the backbone network to multiple users rather than to another user. Wireless APs can also be used to connect two cluster of users (for example, located in two different buildings) to one another. In such applications, WLAN technology is used to bridge two LANs, and this is sometime referred to as a wireless bridge.

Bridges, though, connect networks and are often less expensive than APs. For example, a WLAN bridge can interface an Ethernet network directly to a particular AP. This may be necessary if you have a few devices, possibly in a far-reaching part of the facility, that are interconnected via Ethernet. A WLAN bridge plugs into this Ethernet network and uses the 802.11 protocol to communicate with an AP that is within range. In this manner, a bridge enables you to connect wirelessly a cluster of users (actually a network) to an AP.

6.4.1 Standardization and Bridges

Bridges are devices which connect different LANs together. Since the IEEE 802 community has developed all popular LAN standards, popular bridge standards are also developed by the IEEE 802 community. Since LANs carry variable-length packets, as opposed to telecommunication networks carrying fixed-length packets, bridges or LAN switches are designed to handle variable-length packets. The number 802 came about because it was the next available number for an IEEE committee, but sometimes it is associated with data of the first meeting of this group (February 1980). Figure 1.10 shows the overall perspective of the IEEE working groups of the IEEE 802 project. IEEE 802.1 is a general group working on issues common to all other 802 LAN standards and it is concerned with a number of issues related to all LANs, including internetworking among 802 LANs, MANs and other WANs. Other concerns of the IEEE 802.1 include defining the architecture, link security, network management, and higher layer protocols. IEEE 802.2 defines the LLC protocol, which is also used for all other standards. Starting from IEEE 802.3 Ethernet, traditionally each number was associated with a specific MAC technique with a number of physical layer options. For example, as describe in Chapter 11, the IEEE 802.3 standard define a single CSMA/CD MAC and a number of physical layers operating at different data rates over a variety of media. Most important IEEE recommendations for bridges are defined under the 802.1 group. The most popular IEEE 802.1 recommendation related to bridges is IEEE 802.1D, which defines the operation of the transparent bridges and STA for interconnecting bridges. Another popular bridge related to the IEEE 802 community is the source routing bridge, which was defined for IEEE 802.5 token ring LANs. Other interesting 802.1 recommendations are IEEE 802.1Q, which defines the operational environment for the VLAN, and IEEE 802.1x, which defines authentication algorithms for connecting to a LAN.

6.4.2 IEEE 802.1D Transparent Bridges

The word “transparent” is used for the IEEE 802.1D bridges because bridging does not involve any transaction with the stations connected to the bridge, so they are transparent to the user terminals. These bridges have an automated address-finding mechanism which builds up a table based on the address of the arriving packets to each port. The IEEE 802.1D

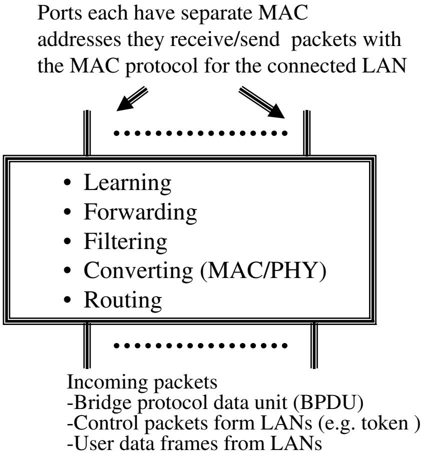

Ports each have separate MAC
addresses they receive/send packets with
the MAC protocol for the connected LAN
• Learning
• Forwarding
• Filtering
• Converting (MAC/PHY)
• Routing
Incoming packets
-Bridge protocol data unit (BPDU)
-Control packets form LANs (e.g. token)
-User data frames from LANs

FIGURE 6.16 Overall structure of the IEEE 802.1D transparent bridges.

transparent bridges have packets exchanged within themselves to establish the routing protocol known as the STA. The traditional IEEE 802.1D bridges were designed to support protocol conversion between legacy LAN technologies such as Ethernet, token ring, and FDDI. These bridges were filtering control packets associated with different LAN technologies and adjusted for differences in MAC priorities.

Figure 6.16 shows the general structure of the traditional IEEE 802.1D transparent bridges. A bridge receives and sends three different types of packet in each port: user terminal's data frames from the connected LANs; control packets such as tokens used for the operation of certain LAN technologies, such as token ring; and bridge protocol data unit (BPDU), which is the protocol for communication among different bridges. Each port of the bridge has a separate MAC address to connect to different LANs. The functionality of the bridge is to *learn* about the terminals connected to each port, *forward* frames to appropriate ports, *filter* unwanted frames, execute *routing* algorithm, and *convert* the MAC and PHY layer of the packets to fit the associated LAN technology connected to each port.

During the *learning* process, a transparent bridge reads the source address of the arriving packets and associates them to the port number they are arriving from. The arriving information is time stamped so that it can be refreshed (e.g. around every 300 s) to manage for changes in network nodes. During the *forwarding* process the bridge looks at the destination address of the arriving packet; if it matches the addresses associated with an existing port, then the bridge forwards the packet to that port. If the destination address is not known, then the bridge flood all ports except the source port and blocked ports. If the source and destination address is the same, then the packet is removed. The *filtering* process eliminates control packets, such as token or ring maintenance packets used in token ring LANs, to prevent them from entering the other segments of the network. If ports are connected to LANs using different technologies or media, then the physical and MAC layers are *converted* at the bridge. For example, if one segment of the network carries 10 Mbps Ethernet and the other segment uses 100 Mbps, then the physical layer conversion is performed at the bridge. Packets of IEEE 802.5 and FDDI carry seven different priorities, IEEE 802.3 and IEEE 802.11 packets have no priority, and IEEE 802.12 has two levels of priority. IEEE 802.1D provides a standard for mapping these priorities to one another which is used during the conversion of MAC headers. IEEE 802.11 uses the STA for routing, which prevents circulation of packets inside a network. The following section provides the details of this algorithm.

6.4.3 The Spanning-Tree Algorithm

The spanning-tree protocol for bridges was first introduced in 1985 at the Digital Equipment Corporation by Radia Perlman [Rad85]. In 1990, the IEEE 802.1D published the first standard for the protocol based on the algorithm designed by Perlman. Subsequent extensions of the algorithm that were published by the IEEE 802.1 community include a spanning tree for VLANs in 1998 and to increase the speed of operation of the algorithm in 2004. Local networks support broadcast and multicast addressing; therefore, bridges should have the capability of flooding all their ports for broadcast or multicast messages. Figure 6.17 shows a typical local network with a number of bridges and LANs. These networks evolve over time and network managers do not have complete information about all connections all the time. This situation is alarming because loops may occur in the network, causing infinite circulation of the packets. In addition to broadcast messages, as we discussed in the previous section, learning bridges broadcast arriving messages in all ports in initial training periods. No matter how the message is broadcast, in each loop a broadcast message from one of the LANs is carried to another LAN and then comes back to the original LAN through a different bridge, and this circulation of the packets destroys the proper operation of the network. To secure proper operation of a local network with a number of LANs and bridges connecting them we need to have at least one connection to every LAN in

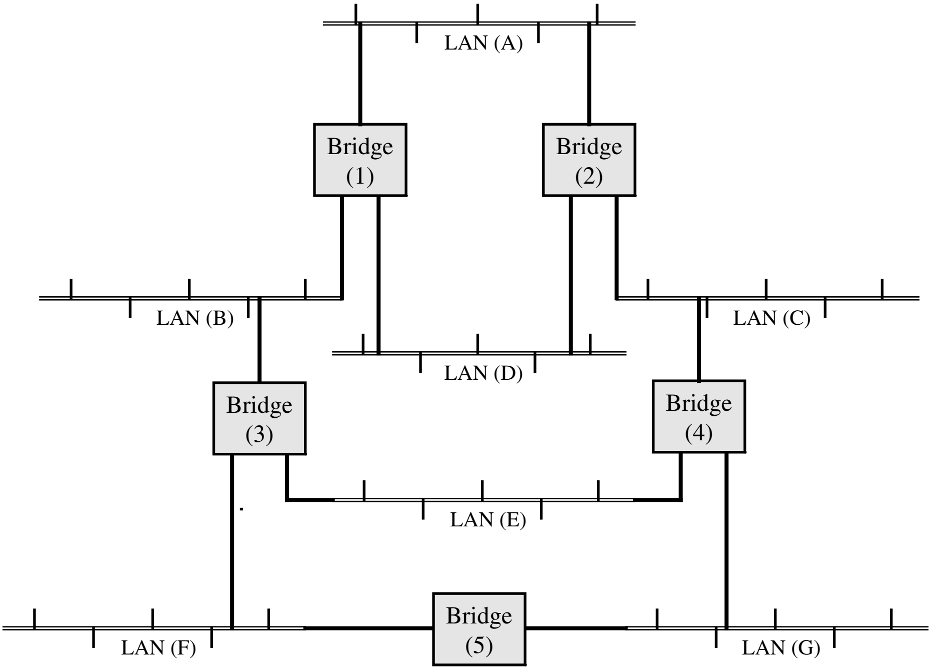


FIGURE 6.17 A local network with seven LANs and five bridges.

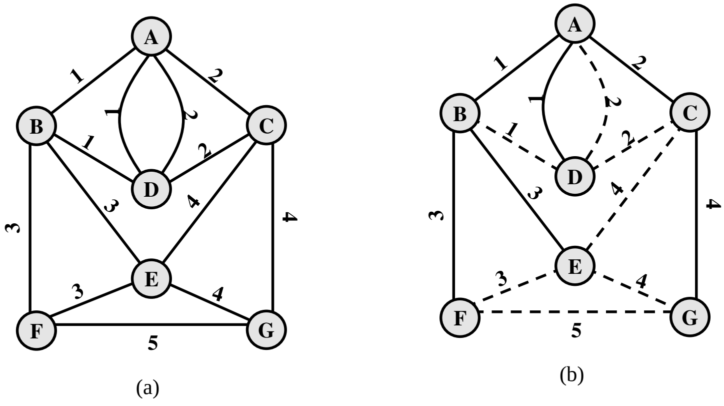


FIGURE 6.18 (a) Topology of the network shown in Fig. 6.17; (b) redundant bridge connections in the network.

the network without having any loop. Figure 6.18*a* shows the topology of the network shown in Fig. 6.17, where each LAN is shown as a node and each bridge as a connection between two nodes. If a bridge has two ports then it connects a pair of LANs one time, and if it has three ports it connects a pair of LANs three times. Figure 6.18*b* shows the minimum paths needed to connect all the LANs and the redundant paths. If the redundant dashed connections are blocked then there is no loop in the network while all LANs are connected to each other.

IEEE 802.1D recommends an STA which automatically breaks loops in the network and maintains access to all LAN segments. This algorithm is implemented in several stages using a set of messages exchanged among the bridges using BPDU messages. Each BPDU is issued from a bridge regularly (every 2 s) and it is broadcast to all other bridges. A BPDU carries the ID of the original bridge and its port, as well as its root path cost. The algorithm operates in a few steps to form the tree that connects all LANs and it blocks ports which cause loops. To start the algorithm the STA finds the *route bridge* and assigns *root ports* and *route path cost* for all other bridges. The *root bridge* of the spanning tree is the bridge with the smallest *bridge identification* number. The bridge identification number has a unique identifier and a configurable priority number. To compare two bridge identities, the priority is compared first; and if two bridges have equal priority, then the MAC addresses are compared. This allows the network administrator to assign a specific bridge as the route bridge by configuring the highest priority to a specific bridge.

The process of selection starts by each node assuming it is the root bridge and broadcasting a packet declaring that claim. When a packet from other bridges arrives in a node, its priority and address are checked against that of the existing bridge. If the arriving packet from a bridge has higher priority than the bridge receiving the packet, then the receiving bridge assumes that the transmitting bridge is the root bridge and sends a message to all ports other than the packet-arriving port to identify the route bridge and the cost of the path to the root bridge. The cost is calculated by adding the exiting cost on the message to the cost of the arriving port. This way, after a starting period for circulation of the messages in the network, the only bridge which sustains its announcement is the actual root bridge and all other nodes have noticed that fact. When the selection of root

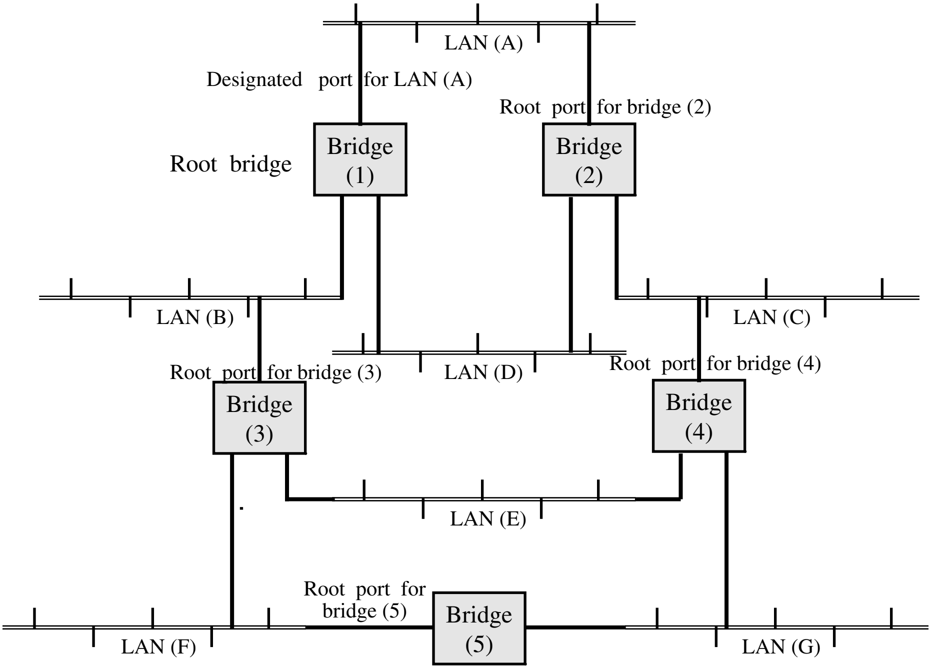


FIGURE 6.19 Root bridge, root port, and designated ports for the network shown in Fig. 6.17.

bridge is settled, each bridge examines the cost of arriving packets from the root bridge to each of it ports and selects the port with the least-cost path as the *root port* of the bridge. The cost of traversing a path is the sum of the costs of the segments on the path, and the cost per segment depends on the bandwidth of the channel: the wider the bandwidth is, the lower the cost is. Alternatively, the network administrator can configure the cost of different segments of the network. Figure 6.19 shows the *root bridge* and *root path* for each bridge in the network shown in Fig. 6.17 assuming that all path segments have the same cost. In such a case, the number of bridges in between the target bridge and the root bridge is the actual cost. To select a root port when we have two candidates with the same cost, the one connecting to the neighboring bridge with lowest identification number is selected.

In the next step of implementation of the STA the algorithm decides on the lowest cost bridge to connect that segment of the network to the root bridge. The port connecting this bridge to the network segment is called *designated port* for the segment. Figure 6.19 also shows a sample designated port for LAN (A). Any active port of all the bridges that is not a *root port* or a *designated port* is then blocked. This blockage is for the flow of data from the terminals, but it does not block the packets exchanged among the bridges using BPDU; therefore, communication among bridges is alive, allowing them to adjust to changes in the network topology in time. Figure 6.20 shows the complete picture for all the root ports, designated ports, and the blocked ports and the resulting implementation of the STA for the network shown in Fig. 6.17. Figure 6.21*a* shows the topology when each node is a bridge and each LAN is a connector. Figure 6.21*b* shows the blocked paths in the network and the resulting tree topology. To select a root port when we have two candidates with the same cost, one of them is selected randomly. If two bridges have

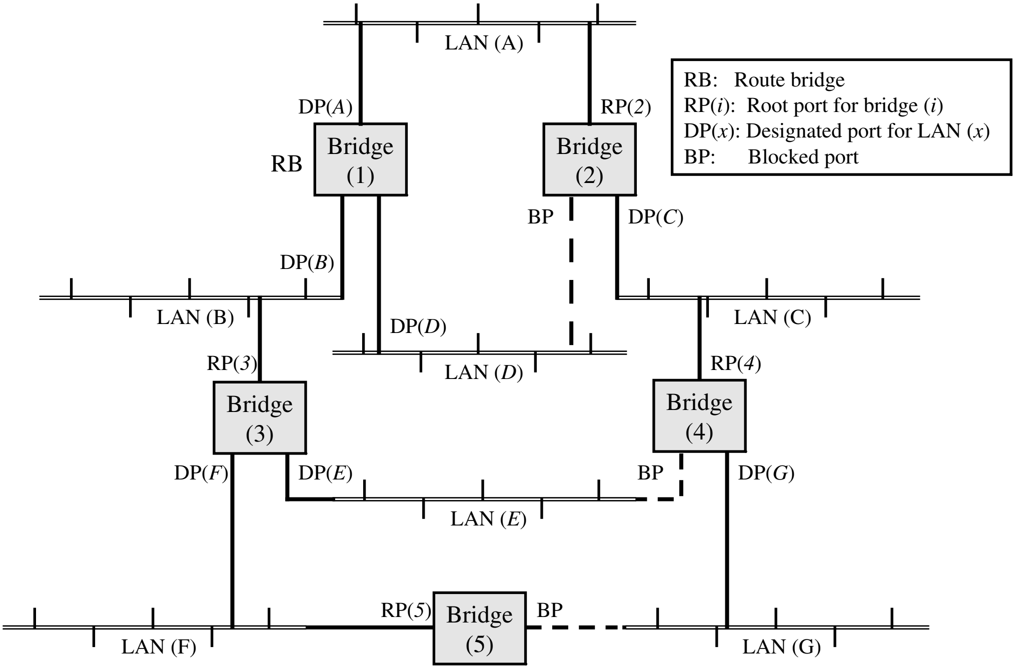


FIGURE 6.20 Implementation of STA for the network shown in Fig. 6.18.

the same cost to connect the LAN to the root port, the bridge with lower identification number is selected.

The spanning tree protocol was replaced by the rapid STA introduced by IEEE 802.1w and adopted by IEEE 802.1D in 2004. The STA was a passive algorithm waiting for time to pass for information collection. The rapid STA actively sends an inquiry packet seeking information from neighboring switches, which results in a faster convergence. The algorithm also has a faster root detection procedure, digs for back-up ports in case forwarding ports failed, and uses other practical tricks to speed up the process. Other modifications to the STA have been made to make it applicable to VLAN environments, as described in Chapter 11.

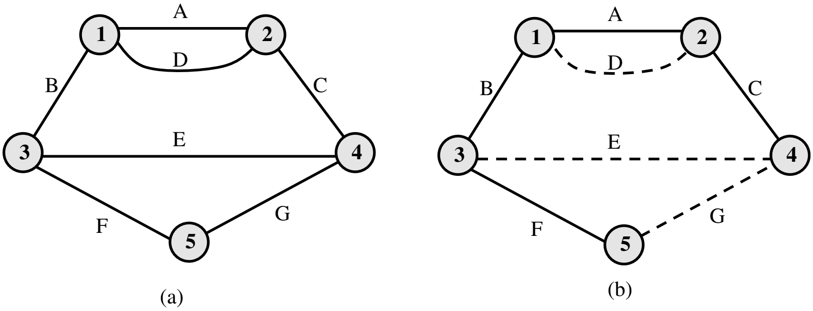


FIGURE 6.21 (a) Topology of the bridges and (b) blocked paths for the network shown in Fig. 6.17.

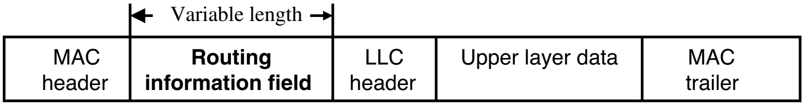


FIGURE 6.22 Packet format for source routing.

6.4.4 IEEE 802.5 Source Routing Bridging

In source routing, the main responsibility of routing a packet is assigned to the source terminal, resulting in a simpler design for the bridges. This approach was used primarily on token ring networks and it was standardized in a section of the IEEE 802.2 standard. As shown in Fig. 6.22, a routing information field is inserted between the MAC and LLC fields of the packets to assist source routing bridging. This field is attached to the packet by the host terminal indicating the sequence of bridges and network segments to be used for delivering of the packet to its destination. The bridges follow the list specified in the routing information field. If the address of the bridge is the next address in the list then bridge forwards the packet, otherwise it ignores it. To find the sequence of bridge and network addresses in the routing information field for a new destination used for the first time by the host, the host broadcasts a special packet. This special broadcast packet instructs the network bridges to append their bridge number and network segment number to each packet as it is forwarded. At the destination, broadcast packets are modified to be standard unicast packets and returned to the source using the reverse path listed in the routing information field. To avoid loops in the path, each bridge ignores packets which already contain its bridge number in the routing information field. Using this approach, the source terminal receives back one packet per each possible path through the network from the source to the destination. At this stage, the source terminal can decide on the path with lowest cost as the communication path of choice for further communications with the destination. The IEEE 802.2 standard allows the coexistence of source and transparent bridging in the same bridged network by using source routing with hosts that support it and transparent bridging otherwise. In transparent bridging, the burden of routing is on the bridge, whereas the main burden in source routing is on the terminal.

6.4.5 IEEE 802.1Q Virtual Local-Area Network

A VLAN allows a group of terminals to form a subnet regardless of their physical location. Figure 6.23 illustrates an example of formation of a VLAN among a terminal connected to a LAN and a LAN located in another place in the network. Implementation of a VLAN is based on attachment of a tag to the MAC address which carries a unique address to identify a particular VLAN group. The format of the tag is defined by IEEE 802.1Q, and bridges complying with this standard have the means to read the tag and identify the VLAN address. The VLAN tag is then used by the bridge to direct the packet to specific segments of the network which are a part of the VLAN. Figure 6.24 shows the general format of a MAC packet and insertion of a VLAN tag which carries the VLAN address. A VLAN operates in the same way as a physical LAN to group end stations together even when they are not located on the same LAN segment. Using the VLAN feature a network manager can configure the network through software control rather than physical relocation of terminals. A VLAN allows the creation of a broadcast domain using hardware MAC addressing

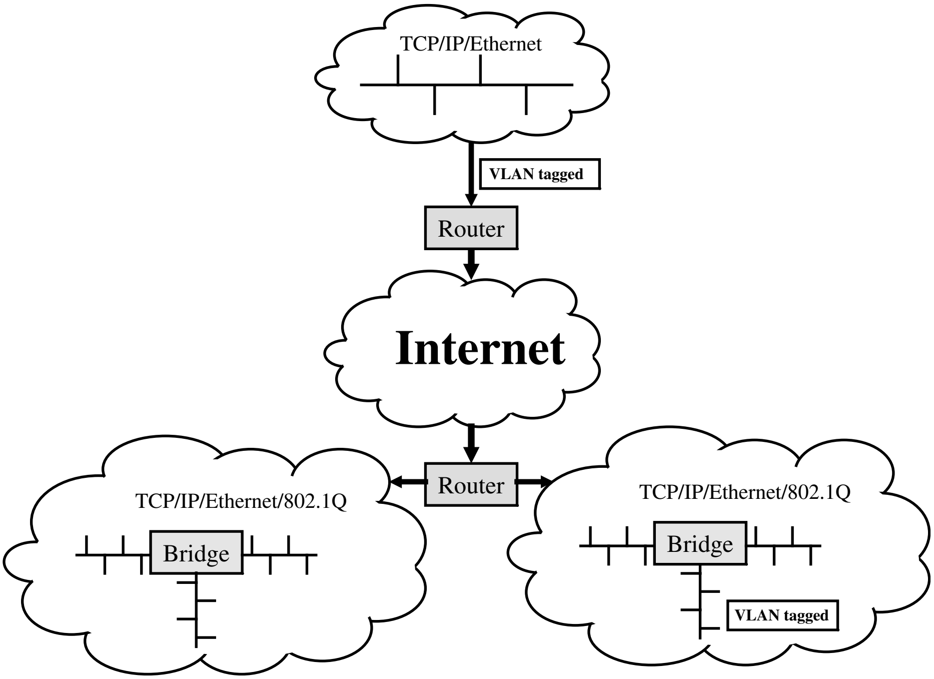


FIGURE 6.23 VLAN formation in geographically dispersed networks.

MAC header	**VLAN identifier**	LLC header	Upper layer data	MAC trailer

FIGURE 6.24 Packet format for VLAN implementation.

mechanisms at the so-called layer 2. In a larger network, switch ports can be assigned to a VLAN to allow packets to be forwarded and flooded only to stations in the same VLAN. VLANs allow the creation of segmentation services traditionally provided by routers in LAN configurations. AVLAN forms a logical network, and packets destined for stations that do not belong to the same VLAN must be forwarded through a routing device using an IP address. An STA can be implemented to a VLAN. More details of the implementation of a VLAN are described in Chapter 11.

6.5 SWITCHES

In the last section we discussed LAN bridges, which are also referred to as LAN switches. These network interconnecting elements evolved for datagram packet switching in a LAN environment for today's Ethernet/Internet datagram environment. Another class of switches evolved out of circuit-switched connection-based PSTN which are fundamentally different from the packet-switched connectionless Internet. In this section we discuss the evolution of these switches.

6.5.1 Circuit Switching in Public Switched Telephone Network

Originally, PSTN was designed using analog voice connections through manual switchboards in the late nineteenth century. The manual switchboards were subsequently replaced by automated telephone exchanges in the early twentieth century; later, digital switches took over core connection for the PSTN in the mid twentieth century. Today, analog two-wire circuits are still used to connect to most telephones, but switches are digital circuits. The basic digital circuit in the PSTN has a data rate of 64 Kbps carrying a typical phone call from a source telephone terminal to a destination terminal. The analog audio sound is digitized at 8 KSps using 8-bit PCM, resulting in a 64 Kbps channel. The digitized voice is then transmitted from one end of the PSTN to another through a series of digital circuit switches. The call is switched using a signaling protocol known as signaling system number 7 (SS7) between the telephone exchanges under an overall routing strategy. SS7 is a set of signaling protocols to set up and tear down telephone calls. The SS7 protocol is also used for number translation, prepaid billing mechanisms, short message services, and a variety of other services. The IETF has also defined level 2, 3, and 4 protocols that are compatible with SS7 but use an IP transport mechanism. This suite of protocols is called SIGTRAN. In principle, PSTN can be thought of as an overlay of a circuit-switched network used for traffic and a packet data network for signaling and control. After call establishment, each circuit cannot be used by other callers until the circuit is released and a new connection is set up. As a result, during the connection period, the QoS is maintained; but even when there is no traffic, the channel still remains unavailable for other users. Technically speaking, these switches are designed for constant bit delay during the connection, whereas an interconnecting element in packet switching uses queues, which may cause varying delay for arriving packets. Many observers believe that, in the long term, future PSTN and circuit switching will be integrated in the Internet, but the current differentiation is basically around the QoS guarantees.

6.5.2 Integrated Service Data Network Switching

In the 1970s, the telecommunications industry started to pay attention to digital services and the concept of ISDN emerged in industry. In ISDN, voice and data are both treated as digital circuit-switched services. In ISDN, the access of the POTS, which was an analog wiring similar to the original access in the early days of the industry, changes to digital access. In the POTS, the analog signal arriving from the customer would be digitized at the network, whereas in ISDN the analog voice is digitized at the terminal to be integrated with the data. Data in the POTS was modulated by modems to an analog signal and then it was sampled to digitized data at the front end of the network to be carried over longer distances. This was an ironically funny situation: the digital data at lower rates, like 9600 bps, was transmitted over the analog lines connected to the user using very expensive voice-band modems. Then, at the network, the same analog modulated data would be turned to a 64 Kbps digital stream to be transferred to the destination where the digitized data would turn to an analog signal to be sent to the modem at the destination terminal. This process was an irrational, expensive solution that had evolved because the main income of the industry was the analog voice and data services were auxiliary applications. With the growth of importance of data services, ISDN emerged to solve this problem. ISDN is able to deliver at minimum two simultaneous connections, in any combination of data, voice, video, and fax, over a single line. Multiple devices can be attached to the line and used as needed.

At a top level, ISDN is a set of protocols for establishing and breaking circuit-switched connections with advanced call features for the user, such as caller ID, call forwarding, and three-way conferencing. The basic ISDN provides two traffic channels at 64 Kbps and one control channel at 16 Kbps, all using digital transmission. Compared with POTS, for delay-sensitive applications, ISDN provides a better quality voice with two simultaneous channels when both ends are ISDN. As a result, it found some niche applications, such as transmission of radio signals in radio stations, video conferencing, and many PBX systems to connect the end office of the PSTN through a T1 carrier. For call setup, control, and other administrative purposes, ISDN services use a separate dedicated signaling channel from the end node to the network, whereas POTS does not.

Compared with the voice-band modems used for data application, ISDN avoids analog transmission and multiple digital–analog–digital–analog conversions by providing a complete end-to-end solution with defined switching infrastructure. Compared with DSL modems, ISDN cards provide lower data rates at a considerably lower complexity and a direct connection to the end office in the PSTN. Although wired ISDN never met the expected market and never achieved a large penetration, it was the favorite backbone of the 2G GSM cellular networks. In fact, 2G digital cellular networks were the fist popular application of an integrated voice and data service over a circuit-switched network in which digitization of the voice takes place at the terminal not at the network.

6.5.3 Packet Switching over Public Switched Telephone Network

The early computer networks were designed to connect remote data terminals and computers together. These networks were using voice-band modems over circuit-switched leased telephone lines. These networks emerged after the Second World War, first for military applications to connect computers and terminals in remote airbases and shortly thereafter spread to airline reservation, banking, and many other commercial applications. The main technical challenge in that industry was actually the design of higher speed modems to save on the expensive cost of leased lines. These networks were using HDLC or similar algorithms for DLLs protocols along with extensive coding to support reliable data transmission over the unreliable telephone network. The HDLC protocol provides connectionless and connection-based services with multipoint capabilities, but it is commonly used for point-to-point connections. The first packet-switched network, X.25, emerged in that environment as the first international packet-switching network in 1978. It had large global coverage during the 1980s and into the 1990s and it is still in use, mainly in transaction systems.

The X.25 model was the traditional telephony application which establishes reliable circuits through a shared network, but X.25 uses software to create "virtual calls" over the connection. These calls connect the data terminal and looks like point-to-point connections. As a result, each endpoint can establish many separate virtual calls to different endpoints. The X.25 protocol was a CCITT (now ITU-T) standard for packet switching over virtual circuits provided by PSTN and provided a reliable data transmission with a data rate of up to 64 kbps. These virtual circuits carry variable-length packets. Today, the dominant Ethernet/Internet packet-switching technology is connectionless and it does not use a dedicated physical or virtual circuit, whereas X.25 packet switching operates based on connection-oriented virtual circuits. In X.25, a connection is established, the data packets are transferred, and then the

Flag	Frame relay header	Information field	CRC	Flag
8 bits	16 bits	128 bytes (max)	16 bits	8 bits

FIGURE 6.25 Frame relay frame format.

connection is terminated. This protocol allows virtual switching for multiple streams over a single line. HDLC is adopted by the ITU as the layer 2 protocol for X.25. This protocol, used commonly on legacy computer networks, provides connectionless and connection-based services with multipoint capabilities. Today, this protocol is commonly used for point-to-point Internet connections, such as transaction processing for credit card authorization and for automatic teller machines, where distant branch offices could be connected to central hosts for a cost that was considerably lower than a permanent long-distance telephone call. X.25 was typically billed as a flat monthly service fee depending on link speed, and then a price-per-packet on top of that. Link speeds varied from 2400 bps up to 64 Kbps.

With the widespread use of digital and optical links in the PSTN, transmission became very reliable and almost error free; consequently, there was no need for the large overhead of X.25 to ensure reliability and it was replaced by frame relay. Frame relay is a fast packet-switching technology that provides no error checking, leaving error control to the end user. X.25 checks point-to-point connections with error correction over all intermediate switches, which causes a large delay and reduces the throughput. Frame relay does not have error and flow control and it provides an end-to-end connection with a much higher throughput which is suitable for LAN connections. Frame relay is generally purchased as an alternative to a dedicated leased line. X.25 prepares and sends packets which are checked throughout the network when they are traveling from one node to another, whereas frame relay prepares and sends frames which are traveling across the network. Therefore, X.25 has flow and error control, whereas frame relay does not. X.25 specifies processing at layers 1, 2 and 3, whereas frame relay operates at layers 1 and 2 only, which has significantly less processing at each node. In addition, X.25 has a fixed bandwidth which wastes the throughput on low load, whereas frame relay dynamically allocates bandwidth during call setup negotiation to avoid throughput waste. Figure 6.25 shows the frame format for frame relay. Flags are patterns of 01111110 and the data is bit-stuffed so that this pattern is not repeated. The overall format is very similar to the HDLC packet format, but control bits are defined differently. In HDLC, control bits and CRC codes are used for implementation of complex flow and error control mechanisms, whereas most of these bits in frame relay are used for addressing and congestion notifications.

6.5.4 Asynchronous Transfer Mode

Figure 6.26 illustrates the overview of evolution of switching techniques in PSTN. The legacy circuit-switched networks supporting POTS using analog connection to an end-user

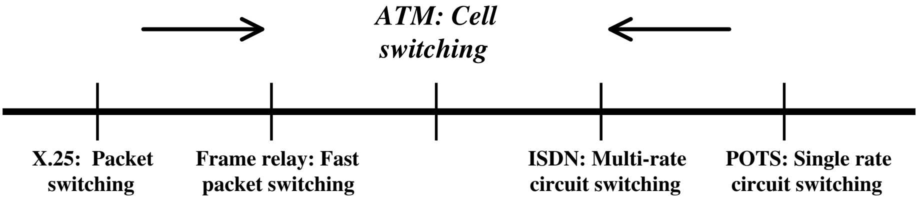

ATM: Cell switching
X.25: Packet switching
Frame relay: Fast packet switching
ISDN: Multi-rate circuit switching
POTS: Single rate circuit switching

FIGURE 6.26 Evolution of ATM switches to integrate circuit-switched and packet-switched traffic.

terminal evolved into ISDN using multi-rate circuit-switching techniques using a connection-based virtual circuit. The packet-switched networks using datagrams over virtual circuit-switched networks started with the lower speed X.25 and emerged into fast frame relay. The next step in the evolution of this industry was the ATM or cell-switching technology which became a standard first in the mid 1980s. The objective of this technology was to design a single networking strategy that could support both delay-sensitive applications (such as real-time video and audio) usually transported over circuit-switched media and data applications (such as transport of image files, text and e-mail) usually carried over packet-switched networks. The standard was designed by a special group called the ATM Forum and became an ITU standard. ATM also is a virtual circuit technology, which uses fixed-length cell relay connection-oriented packet switching.

Figure 6.27 shows the basic concept of the ATM networks. The encoded information traffic is broken into small 53-byte packets with 48 bytes of information and 5 bytes of header to be transferred over the network. This approach is different from other packet-switched networks such as frame relay, Internet, and Ethernet, in which variable-sized *packets* or frames are used. Similar to frame relay and X.25, ATM is a connection-oriented technology, in which a logical connection is established between the two endpoints before the actual data exchange begins.

ATM was designed for efficient integration of voice streams, which was the main income of the service providers at the time, and the emerging data applications, which were expected to form the future. But this entire idea was before the commercial success of the Internet and its associated industry, which generated incomes comparable to voice telephony. At that time, AT&T alone commanded a budget equal to the fifth economy of the world, and that income was dominated by voice telephony application. Therefore, the design of the system pays more attention to the needs of the telecom industry and its popular telephony application. As a result, the motivation for the use of short packets or *cells* was the control of the delay jitter in the integration of voice and data applications in a high-speed link. At the time ATM was designed, 155 Mbps SONET/SDH with a payload of 135 Mbps was considered a fast optical network link. At this rate, a typical full-length

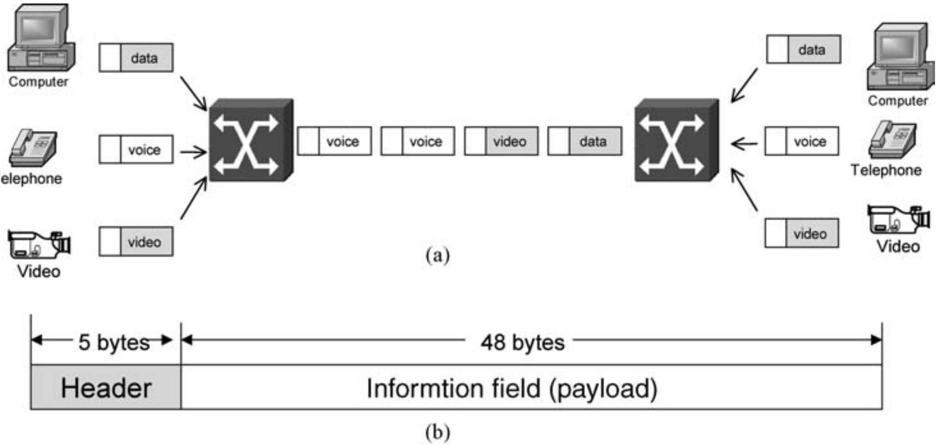
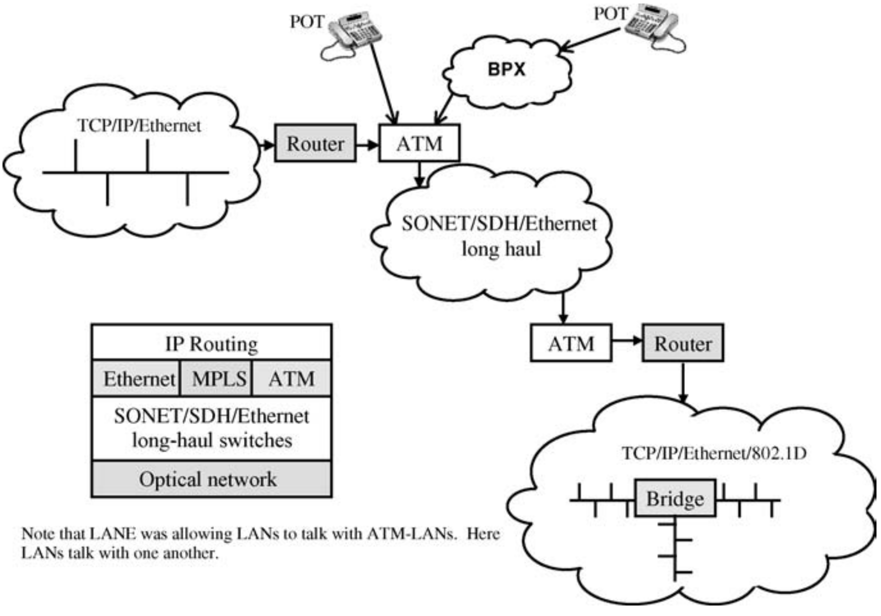


FIGURE 6.27 Principle of operation of an ATM network: (a) integration of media into fixed packet length; (b) format of the packet.

1500 byte (12 000-bit) data packet from a LAN would take 77.4 μs to transmit. In a lower speed T1 carrier link, at the rate of 1.544 Mbps that packet would take up to 7.8 ms. A queuing delay caused by a few of these randomly generated data burst packets between very short and regulated streaming speech packets causes excessive jitter, bringing QoS of the voice below the accepted level by service providers. Designers of the ATM standard were convinced that to be able to provide short queuing delays for voice and carry large datagrams they had to have fixed short packets or cells. As a result, ATM breaks both long data packets and voice streams into 48-byte pieces and adds a 5-byte switching header to each piece to form a cell. The header is used for transport over the switching fabric and reassembly at the destination. The choice of 48 bytes was a compromise between American 64-byte and European 32-byte proposals. The 5-byte headers provided around 10% of overhead for switching information. By multiplexing these 53-byte cells instead of long data packets, ATM technology reduced the worst-case queuing jitter for voice applications to more than an order of magnitude, resulting in a better QoS for the voice users.

In the core of modern optical networks, a 1500-byte Ethernet packet takes 1.2 μs to transmit on a 10 Gbit/s link, allowing reasonable chances to control the jitter for voice streaming packets and the consequent need for small cells. In addition, the cost segmentation and reassembly hardware at those high speeds makes ATM less competitive for IP traffic. On the edge of the networks using data rates below a few megabits per second, adoption of ATM technology makes sense. For example, many ADSL systems use ATM technology between the physical layer and a layer 2 protocol like PPP or Ethernet.

ATM's ability to carry multiple logical circuits on a single physical or virtual medium is useful. Figure 6.28 provides an overview of the integration of PSTN and Internet applications using ATM networks.



Note that LANE was allowing LANs to talk with ATM-LANs. Here LANs talk with one another.

FIGURE 6.28 Integration of plain old telephone and LANs over long-haul optical fiber fabric.

Traffic Classes. An ATM network operates based on the concept of the traffic contract during which a traffic class is negotiated between the terminal and the network on the QoS of the connection. When an ATM circuit is set up, each switch on the path is informed of the traffic class and QoS of the connection. The ATM Forum defines four basic types of traffic, each with a set of parameters describing the connection for different services. The basic specification of the traffic classes is based on the bit rate of the connection, but also on other parameters, such as cell delay variation tolerance defining the clumping of the cells in time. Services are divided into real- and non-real-time as follows:

- Real-time services
  - constant bit rate (CBR), for video/audio conferencing and streaming, which specify a constant peak cell rate;
  - variable bit rate (VBR), for variable rate compressed video/audio, which specify a minimum guaranteed rate and delay characteristics.
- Non-real-time services
  - VBR for critical response time requirement, such as airline reservations and banking transactions which specify a minimum guaranteed rate;
  - available bit rate (ABR), which specifies a minimum and peak rate requirement;
  - unspecified bit rate (UBR), for text/data/image messaging without any requirement, which receives all remaining transmission capacity.

Traffic contracts are usually maintained by traffic shaping and policing. Traffic shaping uses a combination of queuing strategies and marking the cells usually performed at the entry point to an ATM network. Traffic policing maintains network performance by watching the virtual circuits against their traffic contracts. If a circuit is exceeding its traffic contract, then the network can either drop the cells or mark it as a discardable cell so that the network can decide on dropping all cells related to one long packet and save the bandwidth.

Asynchronous Transfer Mode Protocol Reference Model. Figure 6.29 shows the ATM protocol reference model. Traffic and signaling and control packets are processed through different types of ATM adaptation layer (AAL) before they are delivered to the ATM layer for transmission over the physical layer. The physical layer specifies the transmission medium and the signal encoding technique. The ATM layer defines transmission in fixed-size cells and the use of connections. The AAL is a service-dependent layer that maps higher layer packets into ATM cells. The higher layer data includes the actual traffic and flow and error control, as well as call management for the traffic channel. Cross-layer management

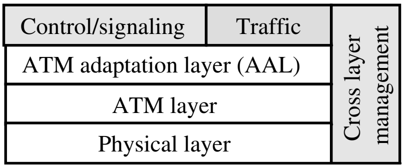


FIGURE 6.29 ATM protocol reference model.

TABLE 6.2 ATM Adaptation Layer Alternatives to Support Different Services

AAL Options	CBR	VBR (real-time)	VBR (non-real-time)	ABR	UBR
AAL-1	Circuit emulation for video/audio conferencing/ streaming				
AAL-2		Variable rate compressed audio/video			
AAL-3/4			Critical response time data		
AAL-5		Voice on demand	Frame relay over ATM LAN		Text/voice/image IP over ATM

operates at the system level to provide coordination among different levels and manage resources and parameters residing in the protocol entities.

There are five AALs defined by the standardization committee for fragmentation and reassembly of different packets: the more popular AAL1, AAL2, and AAL5, and the rarely used AAL3 and AAL4. Table 6.2 shows the relation between service classes and AAL options of the ATM. AAL1 is used for CBR services and circuit emulation, AAL2 through AAL4 for VBR services, and AAL5 for general data. AAL5 is similar to AAL3/4 with a simplified information header scheme, so it is the most widely used AAL for data applications and it is used for classic IP over ATM, Ethernet over ATM, and LANE. LANE technology was perceived to use a connection-based ATM backbone to emulate full-featured connectionless LANs with broadcast and multicast capabilities. In general, AAL5 was intended to provide a streamlined transport facility for higher layer protocols.

Structure of the Network and the Asynchronous Transfer Mode Cells. Figure 6.30 shows the vision of the ATM networks in the mid 1990s. ATM switches in the network connect all the multimedia terminals using the ATM protocol and they also connect to existing LANs. In ATM, a connection has to be established for two parties before they send cells to each other. Call establishment is initiated by the requesting party using a signaling protocol to indicate the address of the receiving party, the type of service it requested, and give traffic parameters if applicable to the selected service. Call admission is then done by the network to confirm that the requested resources are available and that a route exists for the connection. If one desires to connect two LANs using the ATM fabric or connect a LAN to an ATM local network or a terminal, then one needs to create an environment for integration of the connectionless LANs into the connection-based ATM fabric. This technology, although not used in practice today, has been developed by the ATM Forum and is referred to as LANE technology. To connect a connectionless LAN with MAC and IP addresses and broadcast and multicast features to a connection-based ATM which lacks a simple mechanism for broadcasting and uses ISDN addressing, the LANE technology adds three servers to the architecture of the network. These servers are responsible for configuration of the addresses, establishment of ATM virtual links for datagram packets from the LANs, and arranging virtual circuits for broadcast and multicast messages.

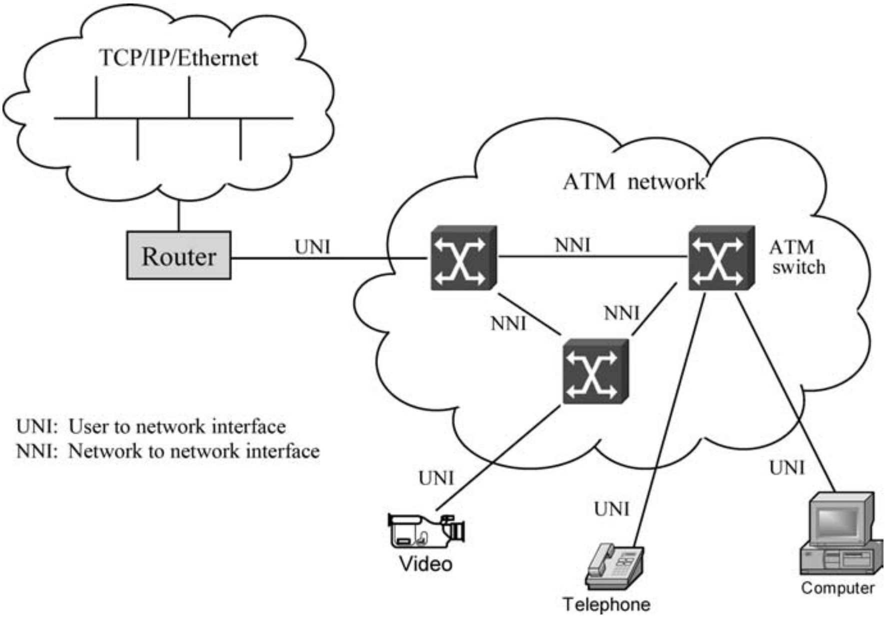
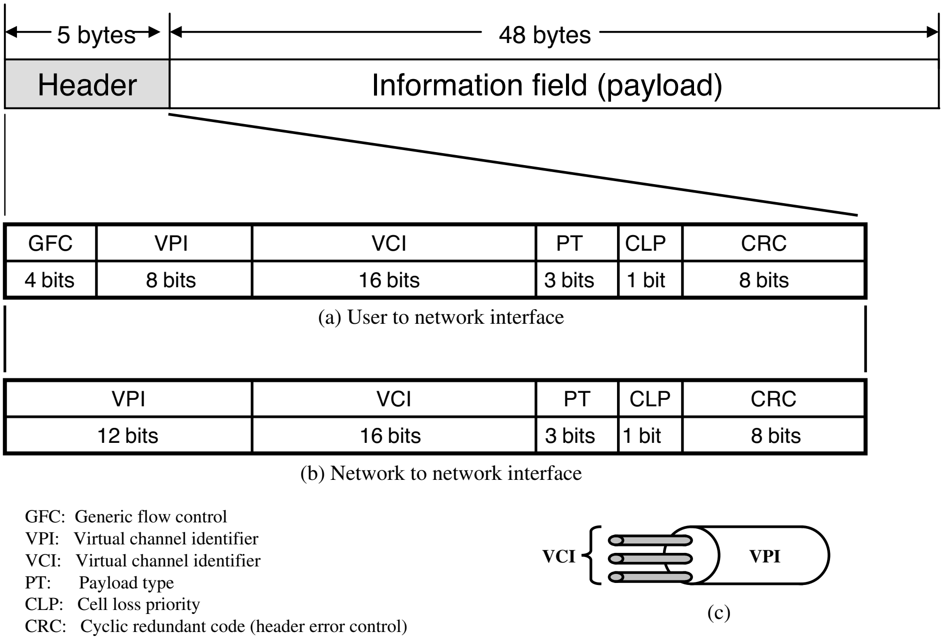


FIGURE 6.30 Vision of integration of voice and data into ATM networks.

There are two sets of interfaces in ATM networks for user-to-network and network-to-network connections. All of the cells exchanging information among these elements are 53 bytes with a 5-byte header and a 48-byte payload. These cells have two different cell formats for network–network interface (NNI) and user–network interface (UNI). Figure 6.31 shows the details of these two cell structures. Most ATM links use the UNI cell format, in which we have a 4-bit generic flow control (GFC) with a default value of four-zero bits and they have not been used widely. The GFC field only belongs to UNI cells and they are reserved for a local flow control system between users to allow several terminals to share a single network connection similar to ISDN services, which could support two phones and a data link over a single basic rate connection. These bits were expected to be overwritten when the cell is passed to the network; therefore, they are not included in the NNI.

As shown in Fig. 6.31, ATM cells have an 8-bit and 12-bit virtual path identifiers (VPIs) for UNI and NNI connections respectively. Each cell is also identified by a 16-bit virtual channel identifier (VCI). As shown in Fig. 6.31*c*, the VPI identifies the physical connection number on a switch port and the VCI specifies one of the channels going through that physical connection port. These two addresses together identify the virtual circuit used by the connection. As a cell traverses the ATM network, at each ATM switch the VPI/VCI addresses change to addresses for the next switch in the connection. As a result, although the VPI/VCI addresses are changing as the packet traverses through the network, the connection to the other end is always through the same set of switches. This concept is in contrast with IP routing, where any given packet could get to its destination by a different route, while the address remains the same. In other words, in ATM the connection is fixed and it is established when the call is made; the addresses are shorter and they only identify the next stage switch on the connection. In IP networks, addresses are longer and they are



GFC	VPI	VCI	PT	CLP	CRC
4 bits	8 bits	16 bits	3 bits	1 bit	8 bits

(a) User to network interface

VPI	VCI	PT	CLP	CRC
12 bits	16 bits	3 bits	1 bit	8 bits

(b) Network to network interface

GFC: Generic flow control
VPI: Virtual channel identifier
VCI: Virtual channel identifier
PT: Payload type
CLP: Cell loss priority
CRC: Cyclic redundant code (header error control)

FIGURE 6.31 Structure of an ATM cell and details of the 5-byte (40-bit) header: (a) for UNI; (b) for NNI; (c) relation between the two layer addresses.

maintained in the packet, while the route of the connection may change from one packet to another. Short ATM packets going over a known connection allow more control on delay and the QoS.

The payload type (PLT) field in Fig. 6.31 is used to designate various kinds of cells for operation and management of the network. The cell loss priority (CLP) bit is used to identify whether the cell can be dropped or not. The 8-bit CRC is used to correct single-bit errors and detect multibit errors in the header. When a multibit header error is detected, cells are dropped until a cell with no header errors is found. The 8-bit CRC polynomial used for the header is given by $G(x) = x^8 + x^2 + x + 1$.

The advantage of virtual circuits is to support delay-sensitive real-time applications. Datagram packet switching over connectionless networks in the original Ethernet/Internet networks were best-effort networks lacking such mechanisms. Technologies such as multiprotocol label switching (MPLS) and the RSVP created virtual circuits on top of datagram networks. MPLS provides services similar to ATM for variable-length packets and it is sometimes referred to as ATM without cells. Modern routers using multi-gigabit per second speeds, however, do not require these technologies to be able to forward variable-length packets because the transmission time of the packets is so short that the associated queuing delay causing jitters does not harm the QoS significantly.

6.6 ROUTERS

Routers work in a manner similar to switches and bridges, in that they are intermediate nodes directing the traffic from the source terminal to a destination terminal. Compared with connection-based switches, routers have evolved for connectionless datagram applications. Compared with LAN bridges, routers are designed to connect dissimilar networks for wide area internetworking using the third layer and below. In the 1980s and early 1990s a router was designed to internetwork different LAN technologies such as Ethernet, token ring, and FDDI with long-haul X.25 networks. This network was designed mainly for data applications and the PSTN was handling the voice and video applications through its connection-based network. Traditional routers designed for that environment had the capabilities to adjust addressing schemes and maximum frame size which are not the same in different networks. For example, IEEE 802 LANs normally use 48-bit MAC addresses which are binary numbers and the maximum packet length of these sub-networks is 1500 bytes; the X.25 frame-switching network uses 48-bit address, with each 4 bits representing a decimal number to form a 12-digit address and the maximum length of the packets is 1000 bytes. A router has the capability to cope with these situations and adjust the address and the lengths. These routers were designed on minicomputers on software which created another differentiation between them and the bridges and switches, which were hardware devices. Today, multiprotocol routers once connecting networks with different transport layers such as DECnet, AppleTalk, Xerox protocol, and IP are absolute and all routers use IP at the network layer. Modern high-speed routers are specialized computers with extra hardware added to accelerate both common routing functions, such as packet forwarding, and specialized functions, such as encryption.

Long-haul connection-based switches were the technology of choice for the telecommunication industry generating its income from reliable telephone connections, which was the dominant source of communication in the past century. Routers were developed by the computer communication industry to connect computer terminals and a variety of LAN technologies worldwide. Until the mid 1990s and the start of the mass popularity of the Internet, the income of the data networking applications and, consequently, the size of the data communication industry was not comparable to that of the telecomm industry generating its income mainly from telephony application. Therefore, the telecom industry dictated its view of long-haul networking using connection-based services led by the telephony application. The emergence of the ATM technology around the 1990s should be thought of in this light. Comparing an ATM network with the Internet, one should realize that ATM was designed mainly for the dumb analog telephone terminals with strict delay requirements for the core network, whereas the Internet was designed for smart digital computers with plenty of computational power and many applications without strict requirements on delay. As a result, ATM technology is very complex, but it supports the QoS needed for the telephony application, whereas the Internet is simple and unreliable using complex computations and protocols at the terminals and the routers to implement a variety of applications. As shown in Fig. 6.1, parts of the Internet traffic in the core of the network are still carried using ATM technology today.

During the 1990s, three phenomena happened which affected the design of long-haul interconnecting devices. The Internet became popular, generating an income comparable to PSTN, the cellular industry shifted the circuit-switched popular telephony application to mobile telephone, and Ethernet emerged as the dominant LAN technology. This change promoted a sharp and fast growth of router design technology and the idea of integration of connection-based cellular telephones into IP networks. In the past decade, the dominance of the fiber lines, the popularity of the voice and video over Internet, the growth of home networking, and the success of smartphones have resulted in the emergence of home router technology. Today, we have core routers, edge routers, and home routers all designed for IP networking. The small home office uses small routers to connect to an ISP and corporates use racks of routers to manage connection to the Internet.

An IP router partitions a network into subnets and directs the arriving traffic destined for particular IP addresses toward a segment of the network using a routing algorithm which optimizes the cost. The cost for intelligent forwarding and filtering is usually calculated using the speed of the network. This protocol filtering usually takes more time than the packet filtering used in LAN bridges. Routers examine data as it arrives, determine the destination address for the data, and use routing algorithms to determine the best way for the data to continue its journey. Unlike bridges and switches, which use the hardware to determine the destination of the data, routers use the software network address to make decisions. This approach makes routers more functional than bridges or switches, but it also makes them more complex and expensive.

IP addresses are soft addresses, and the address of a device may change in time; when a packet arrives in a LAN, it needs to know the MAC addresses which uniquely identify the hardware. To map an IP address to the MAC address of different devices, routers use the ARP. Assume terminal A wants to send an IP packet to terminal B in the same LAN. Terminal A must already have the IP address of terminal B. However, in order to physically send the packet to terminal B the hardware or MAC address of that terminal is needed. If terminal A does not already know that MAC address of terminal B, it sends an ARP request to ask for that MAC address. If terminal B is still connected to the same LAN then it sends a response with its own MAC address to the ARP request for the MAC address. If the terminal has moved to another location and uses a mobile-IP protocol, as shown in Fig. 6.13, then the home agent in the LAN responds to the request with the MAC address of terminal B. The home agent captures the packet addressed to terminal B and encapsulates it in a new IP address and sends it to a foreign agent IP address to decapsulate the message and delivers it to the mobile user. In addition to IP and MAC address resolution, the ARP is also used for IP over ATM links.

6.6.1 Types of Router

Figure 6.31 provides an overview of the role of the routers in the Internet. ISPs provide wide area connectivity among different enterprise and home networks. Each ISP organization manages a group of routers connected together exchanging information using a common protocol. This group of routers is referred to as an autonomous system (AS). The routers inside an AS domain use the interior router protocol (IRP) to communicate with each other and an exterior router protocol (ERP) to communicate with outside of the AS domain. The most popular IRP is the OSPF, which uses Dijkstra's shortest path algorithm. The most popular ERP is the border gateway protocol (BGP). The BGP is the core ERP routing protocol of the Internet operating on TCP/IP. It allows negotiation between bordering routers to learn if another AS is willing or able to carry the traffic of an AS. The BGP maintains a table of IP networks for network reachability between ASs. The BGP is fundamentally a distance vector protocol in which distances are defined based on political, security, or economics considerations. These policies are manually configured in the BGP routers and they are not part of the protocol. Internet service providers must use BGP to establish routing between one another. So, BGP is comparable to SS7, the inter-provider call setup protocol for PSTN. The Internet does not have a clearly identifiable backbone similar to the PSTN.

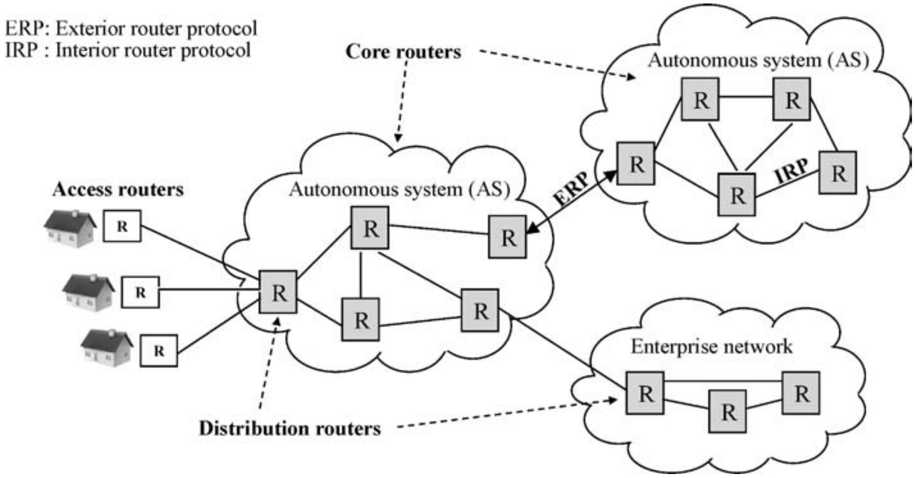


FIGURE 6.32 Access distribution, and core routers.

Different sizes of routers have emerged for variety of applications. Smaller routers are often found in small homes and small offices, while the large routers are found in ISP core networks, academic and research network facilities, as well as in networks for large businesses. One way to classify different routers, shown in Fig. 6.32, is to divide them according to their application into *access*, *distribution* and *core routers*. *Access routers* are small routers used by small subnetworks at customer sites, such as a branch office of an enterprise, a small office, or a home office, to connect to ISP services such as IP over cable, DSL, or fiber optic links. These small subnetworks do not need hierarchical routing of their own and are designed for minimal cost. These routers or gateways commonly use network address translation instead of routing algorithms to support the connection. Therefore, instead of connecting local computers to the remote network directly, these routers make all local computers appear to be a single computer connecting to the ISP. Today, most of these routers integrate WLAN access for the laptops and carry several other ports to connect to desktops. A new technology called *femtocell* technology is emerging in that area to integrate cellular telephone services into the home router unit. This way, the troubled coverage of cell phones inside residential areas will improve and the gateway supports integrated services for access to both Internet and the connection-based PSTN.

Distribution routers connect access routers to the rest of the Internet. These routers collect and police multiple streams of data from different access routers to assure the QoS for the user and to enforce that across the WANs. In enterprises, distribution routers connect multiple buildings of a campus or distributed buildings of large enterprise locations. As a result, distribution routers have large memories, complex algorithms and processing intelligence, and multiple connections to the WANs to support diversified resources for transmission. When an enterprise is primarily concentrated in a relatively large local geographical area, such as a university campus or an office building, the need for a distribution router is eliminated and the access router connects to the core router directly.

Core routers provide a backbone to interconnect the distribution routers from multiple buildings of a campus or large enterprise locations. These routers process huge amounts of information and they tend to optimize for high bandwidth. If the campuses of an academic institution or branches of a large enterprise are widely distributed over a wide geographical area, then each campus or branch office can be directly connected to the WAN service providers, eliminating the need for core routers, which makes the distribution routers the highest tier router in a networking operation.

6.6.2 Network Protocols for Routers

Routers use two classes of protocols to route a packet throughout a network. These protocols are referred to as *routable or routed protocols* and *routing protocols* and often they are confused with one another. These protocols are designed for implementation of different functions needed for communication between user applications in source and destination devices, they are routed over the internetwork, and they vary among different protocol suites. IP, DECnet, and AppleTalk are examples of routed protocols. Routing protocols are designed to implement routing algorithms and are used by routers to build tables for determining path selection of routed protocols. Examples of these protocols include the IRP, OSPF, BGP, and routing information protocol (RIP).

Routed or Routable Protocols. To route a packet in a large WAN consisting of a number of LANs we need a protocol identifying the address of different networks and a unique address which identifies individual terminals within a specific network. Protocols having both of these features are called routable protocols, and TCP/IP, internetwork packet exchange/sequenced packet exchange (IPX/SPX), and AppleTalk are three examples of such networks. TCP/IP, designed by the Department of Defense during 1970s, is today's dominant protocol. TCP/IP uses the RIP and OSPF algorithm for routing. RIP is a distance-vector routing protocol and OSPF is a link-state routing protocol. IPX/SPX was created by Novell for use on NetWare networks and today it is replaced by the TPC/IP even inside Novell. AppleTalk was designed for the Macintosh computer networks in the early 1980s and it is widely used in Apple networks. AppleTalk uses an algorithm called the routing table maintenance protocol (RTMP) for routing functionality, which is similar to the RIP used by IPX/SPX and TCP/IP.

The TCP/IP is based on a four-layer reference model by the IETF consisting of an application layer, a transport layer, a network layer and a link layer. A number of protocols at a higher level use different options for the lower layer protocol to accomplish their objectives. For example, an application layer protocol such as FTP, SSH, POP3, TELNET or SMTP has transport option protocols such as TCP or UDP which can be sent over an IP-v4 or IP-v6 network protocol using a link layer protocol such as Ethernet, WiFi, or ATM. The most popular of the transport/network protocols is the TCP/IP. Figure 6.33 shows the basic concept behind the operation of the TCP/IP. The TCP provides an end-to-end connection between the two applications and IP is used for routing the protocol. The TCP/IP is so important for the Internet that sometimes the entire suite is referred to as the TCP/IP suite.

IP is designed for datagram delivery over the Internet using packets of information also referred to as datagrams. Information packets are short sequences of binary data consisting of a header and a body. The body contains the application data, and the header describes the destination and some additional bits to be used by the routers on the Internet to pass the packet along a sequence of routers towards the destination. Since in datagrams different packets may take different paths, the packets may arrive out of sequence. Where

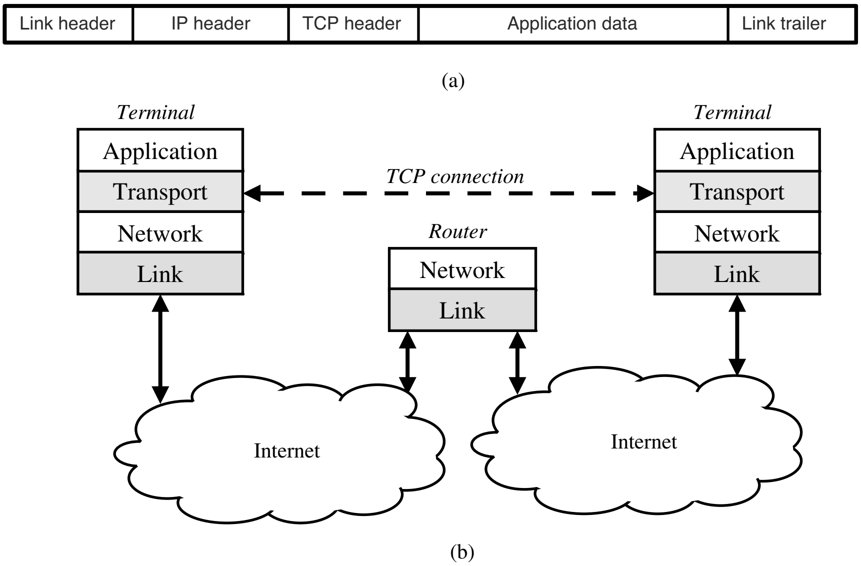


FIGURE 6.33 IP stack and TCP/IP: (a) packet headers; (b) operation for terminals and routers.

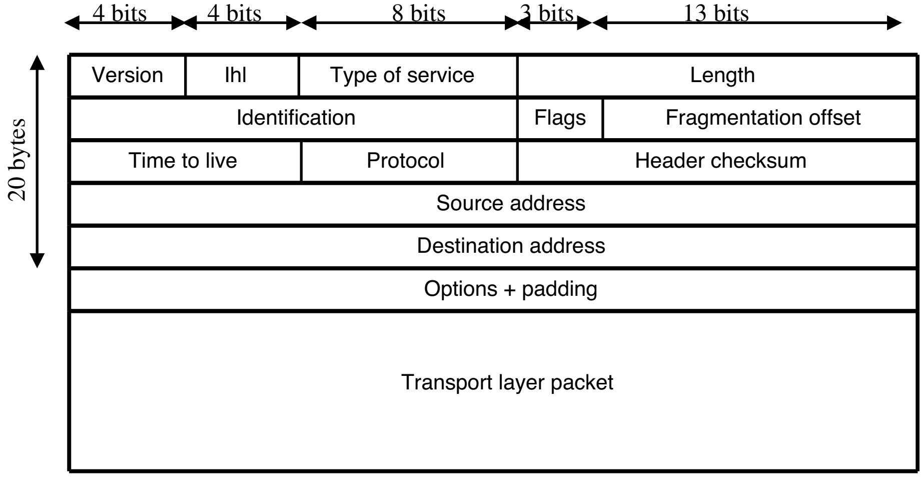
there is congestion in a specific router, the IP may discard a packet, causing a packet loss in a stream of packets. Since packets sometimes go through LAN bridges or other interconnecting devices which sometimes flood the ports, we may have packet duplications at the destination. Therefore, a stream of packets generated by an application may arrive at a destination in the wrong order and with missing or duplicate packets. TCP is an end-to-end connection-based protocol operating on the arriving packets to provide a reliable delivery service that guarantees to deliver a stream of packets sent from a host to a destination without packet duplication or packet loss. TCP also allows multiplexing several applications over an IP stream. In a sense, the functionality of the TCP is similar to DLLs, because they are both designed to provide reliable communications. DLL protocols are working over an unreliable physical link, while TCP operates over a randomly varying sequence of reliable links with the possibility of delivering packets in duplication or out of sequence. Therefore, similar to popular DLL protocols such as connection-based HDLC, the TCP is a connection-based protocol using congestion or flow control as well as error control mechanisms.

Since TCP is optimized for accurate delivery rather and is not focused on timely delivery, sometimes it results in relatively long delays on the order of seconds. This delay is caused while the TCP waits for out-of-order messages or retransmissions of lost messages. As a result, TCP is not particularly suitable for real-time applications such as VoIP, in which delays greater than 100 ms cause undesirable disturbances in the quality of the voice, while losing up to 1% of the packets does not have any significant impact on the quality of the voice for the user. For such applications, the connectionless UDP usually replaces the TCP. Application layer protocols such as the real-time transport protocol (RTP) can run over the UDP transport layer. The TCP provides reliable in-order delivery of a stream of bytes, making it suitable for traditional Internet applications such as file transfer, e-mail, or Web access. TCP is the transport protocol that manages the individual conversations between Web servers and Web clients. In these applications, TCP divides the long messages into smaller segments and it is responsible for controlling the size and rate at which messages are exchanged between the server and the client.

TCP has been optimized for wired networks in which lost packets are caused by congestion mostly in the core network. In wireless links, packets are also lost because of multipath and shadow fading, as well as handoff in the wireless access portion of the network. The congestion control mechanisms used to remedy the packet loss in long-haul networks with relatively long window sizes are not suitable for the packet losses in short distances due to wireless channel impairments. Modern cellular networks integrating voice and data use different approaches to address this problem.

The IP is a data-oriented network layer protocol used for communicating datagrams or packets across a heterogeneous packet-switched network. The packets with IP headers were originally designed to be encapsulated in data link protocols such as Ethernet, FDDI, token ring, or others. The service of an IP header to a lower layer protocol with its own MAC layer address is to provide for a communicable unique global addressing amongst computers. The current and most popular version of the IP is IP-v4 originally defined in 1981. IP-v6 is the proposed successor to IP-v4, whose most prominent change is the modification of 32-bit addresses (approximately 4 billion) to 128-bit addresses (approximately 3.4×10^{38}).

Figure 6.34 shows the general format of an IP-v4 packet and the details of its header. The length of the header is 20 bytes, in which 4 bytes are used to identify the address. This compares with the 26-byte length with 6-byte addresses for IEEE 802 MAC addresses. The main difference between the two addresses, however, comes from the fact that MAC addresses are randomly distributed, whereas IP addresses carry a certain order and a distribution technique that allows reasonable tracing of the address into geographical locations. In the header, the first 4 bits are associated with the Version field that identifies the version of the IP. The next 4 bits are used for the Internet header length (IHL), which specifies the length of the header in 4-byte (32-bit) words. Most of the time there is no option or pad, leaving the value of IHL at 5. The Type of Service field is 8 bits and is



IHL: Internet header length

FIGURE 6.34 Frame format for the IP-v4 packets with details of the header.

reserved for implementation of priority for a given packet. This type of priority is similar to the priority of the LANs which can be implemented if the interconnecting bridge or router has multiple queues with different priorities, which is not the case for commonly used bridges and routers. The next 16 bits identifies the length of the packet, including the header, in bytes. Therefore, the length of an IP packet is $2^{16} - 1 = 65\,535$ bytes. The link layer for carrying IP packets such as Ethernet often has a lower limit for packet length, which calls for fragmentation and reassembly mechanisms. The 16-bit Identification field is used to identify the length of the fragments of a long packet and it is the same for all fragmented packets. The 3-bit Flags field indicates whether the packet is fragmented so that at the destination the packet can be reassembled. The 13-bit Fragment offset shows the sequence number for a fragmented packet in 8-byte (64-bit) units. This way, the fragmented packets can be reassembled in the destination terminal and routers do not get involved in the reassembling process. The 8-bit Time to Live field is used in practice to count the number of hubs to allow implementation of a mechanism to discard packets which are trapped in routing loops. The value can be set to any number up to the limit of 255, and each router reduced counts this number one time down until this field represents a "0," for which the router drops the packet. The default value of the Time to Leave field commonly used in practice is 64. The 8-bit Protocol field identifies the transport layer protocol that the packet should be delivered to. For example, a Protocol field of 6 identifies TCP and a Protocol field of 17 signifies UDP transport.[3] The 16-bit Checksum field is calculated by dividing the entire header into 16-bit blocks and determining the remainder in a simple binary form to generate a checksum (details are given in Example 4.8 in Chapter 4). Since the generation of the checksum includes the destination address, if the checksum does not match then the packet can be discarded, leaving it to the TCP layer to recover that information. Compared with CRC codes used for IEEE 802 MAC packets, the coding technique used for IP headers is simpler but less powerful. The source and destination addresses follow the IP addressing format described in Section 6.2.3. The Options field is used for security, source routing, record routing, time stamping, and other features. The Padding field is designed to make sure that the length of the packet is a multiple of 32 bits.

Comparing the IP-v4 header with the header of the Ethernet LANs, the IP header addresses provide for destination traceability, the means for fragmentation and reassembly of the packets, a timer for discarding wandering packets, and a few bits for priority. These features are exploited by routing algorithms operating in the routers to direct a packet towards its destination. Using an IP header, data from an upper layer protocol is encapsulated inside one or more datagrams without any circuit setup and it is delivered to its destination host with which it may not have previously communicated; for this reason we call it a connectionless protocol. This approach is quite different from PSTN, in which we establish a connection before a phone call goes through and we refer to it as a connection-based protocol. The TCP also provides an end-to-end connection, but that connection is for reliable packet transmission and does not specify a physical or virtual connection on the network elements, as PSTN does.

The IP works based on encapsulation of lower layer packets, which enables networking over heterogeneous DLL environments such as Ethernet, WiFi, or ATM. When using IP to connect two computers, it makes no difference to the upper layer protocols which DLL network is connected to the individual computers. Each of these DLL networks can have its

[3] Details for other protocols are available at www.iana.org.

own method of addressing with a corresponding need to resolve IP addresses to data link addresses. This address resolution is handled by the ARP. The IP provides an unreliable best-effort delivery service which may result in corrupted, lost, or duplicated packets; the only reliability assurance of the IP is that the header is error free with very high probability through the use of a checksum. The primary reason for the lack of reliability is to reduce the complexity of the routers and allow them to discard packets when necessary. Perhaps the most complex aspects of IP are IP address assignments and routing. Address assignment to subnets and individual hosts should be so that the routers can somehow find the host with the destination address. The IP datagram is routed through routers, typically using interior gateway protocols (IGPs) or external gateway protocols (EGPs) to communicate between routers of an AS and among different ASs and forward the packet towards its destination address.

Figure 6.35 shows the format of the IP-v6 packets. The overall header is 40 bytes, compared with the 20-byte IP-v4 header. The address is 128 bits rather than 32 bits, allowing practically an unlimited number of addresses to accommodate the envisioned future networks in which every light bulb can have an IP address. Another feature of IP-v6 is the header compression option, which can reduce the overhead. The 4-bit Version field is the same as IP-v4. The 8-bit Traffic Class is an extended version of the Type of Service field in IP-v4. The 20-bit Flow Label is reserved for flow control and together with the Traffic Class field is used to implement QoS over the Internet. The 16-bit Payload Length field is the same as the Length field in IP-v4, with the difference that the length in bytes in IP-v6 excludes the length of the header. The 8-bit Next Header field replaces the IP-v4's Options and Protocol fields. Normally, this field shows the transport layer code; but if there is a need for an option; that is also indicated in this field. In case there is an option, that field follows the header. The 8-bit Hop Limit is the same as the Time to Live field of IP-v4, with a more appropriate name used in practice.

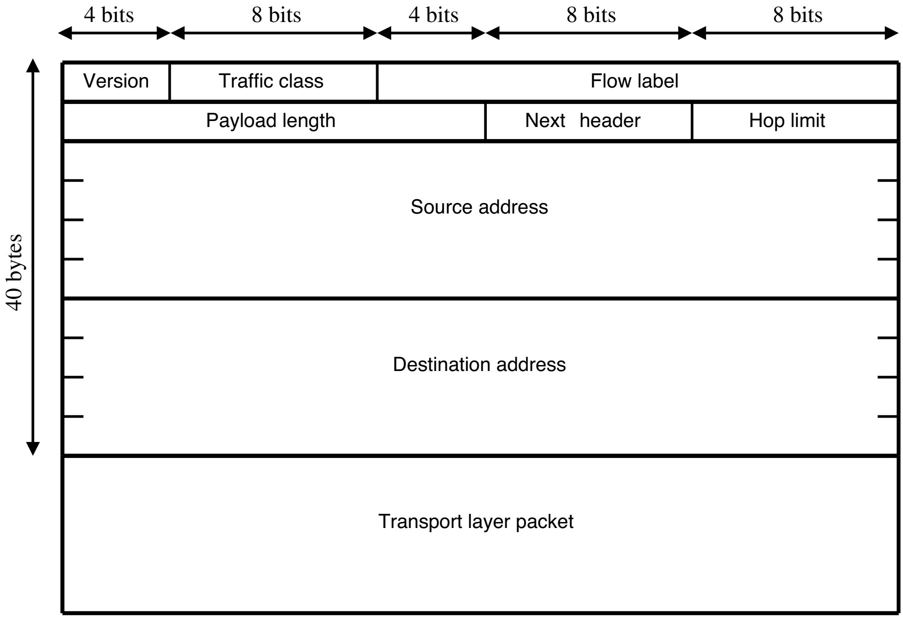


FIGURE 6.35 Frame format for the IP-v6 packets with details of the header.

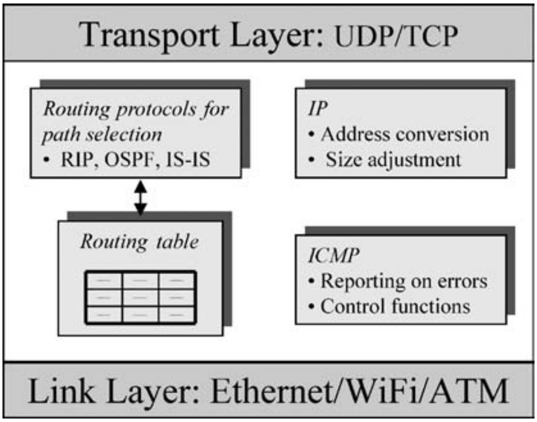


FIGURE 6.36 Overall view of the functionalities provided at the network layer.

The main change brought by IP-v6 is a much larger address space that allows greater flexibility in assigning addresses. The large number of addresses allows a hierarchical allocation of addresses that may make routing and renumbering simpler. Since using IP-v6 has an impact on the design of the routers and endstations, the transition from IP-v4 to IP-v6 was much less successful than expected.

Figure 6.36 provides an overview of different functionalities of the network layer used within a router. As we discussed earlier in this section, IP is a routable protocol used for adjusting the packet lengths and support a structured addressing scheme for routing among different subnetworks. The IP basically supports host-to-host datagram services in a system of interconnected networks. Routers also need to communicate between themselves for control purposes and with the host terminals; for example, to report an error in datagram processing. The Internet control message protocol (ICMP) messages are used to implement these functions. The objective of ICMP control messages is to provide feedback about problems in the communication environment, but it does not fix the problem and IP packets will remain unreliable. When a reliable transmission is needed, higher levels are responsible to provide reliable communication procedures and they can use ICMP messages. The IP address is used by routing algorithms, such as RIP, OSPF, and intermediate system to intermediate system (IS-IS), to prepare a routing table inside each router. The RIP was one of the most commonly used IRPs allowing routers dynamically adapt to changes of network connections using distance-vector algorithm. Although RIP is still actively used, it is generally considered to have been made obsolete by routing protocols such as OSPF and IS-IS using a link-state algorithm. In the next section we discuss algorithms used for routing datagrams in the Internet.

6.6.3 Routing Algorithms

A routing protocol is a protocol that specifies how routers communicate with each other to exchange information, allowing a router to find out the next router it has to deliver an arriving packet to so that the packet is directed towards its destination address. This communication is necessary so that routers can learn the network topology and its changes throughout the time. Since each router can only communicate directly with its immediate neighbors, a routing protocol provides a means for sharing this information with distant routers so that, ultimately, each router has knowledge of the entire network topology. Therefore, a routing protocol can be defined as a protocol operating at the network layer to disseminate topology information among routers. The routing protocols differ from one another by the way they determine preferred routes from a sequence of hop costs and other preference factors, as well as the method they use to handle routing loops. Routing protocols rely on the addressing of the routable protocols, such as IP, to implement their algorithms. There are two major types of routing algorithm: distance–vector routing algorithms and link-state routing algorithms.

In link-state algorithms, also known as short path first algorithms, each router floods portions of the routing table that describes the states of its own links to all routers in the entire network, allowing each router to build a picture of the entire network in its routing tables. In distance-vector algorithms, also known as Bellman–Ford algorithms, each router sends all or some portion of its routing table only to its neighboring routers. In other words, link-state algorithms send small updates everywhere, while distance-vector algorithms send larger updates only to neighboring routers. Routers using link-state algorithms need larger memories and longer computational time than routers using distance–vector algorithms, increasing the implementation costs of the router. However, link-state algorithms converge faster, scale better, and create smaller numbers of routing loops than distance–vector algorithms.

Routing algorithms use a metric to describe the “cost” of a certain route and, based on the cost, they decide how to direct a packet towards a certain node. The metric can be a combination of the number of hubs to the destination, the rate of the links, which governs the delay for transmission of the packet, or a value that is assigned by an administrator to discourage use of a certain route. Path length, reliability of the path, delay, bandwidth, communication cost, and the load of the link have been used as metrics in different routers. The alternative to using routing protocols is that the network administrator manually enters the route information. Manually entering routes is time consuming, subject to human error, and demands manual reconfiguration when the network topology changes. This approach is referred to as static routing and it is sometimes used in very small networks.

Distance–Vector Routing Algorithm. With a distance–vector algorithm, each router calculates the distance and direction of all routers in the network by exchanging information only with the neighboring routers. The distance or cost of reaching a destination is determined by using mathematical calculations comparing costs of directing a packet towards different directions or vectors. Different implementations of the algorithm may use a simple hop count or other information such as data rate, traffic density, and economical or political preference factors to calculate the costs. Neighboring routers using the same distance–vector protocol update their tables periodically to adjust to changes in the network. The frequency with which routers send route updates depends on the routing protocol and it is usually between 10 and 60 s. At each update the entire routing table of a router is sent to all of its neighboring routers. Other routers check the received information against the existing routing table to make appropriate changes if necessary. A simple example helps in understanding the algorithm better.

Example 6.6: Distance–Vector Routing Figure 6.37 shows a simple network with four routers and five links interconnecting them. At the start of operation each router knows

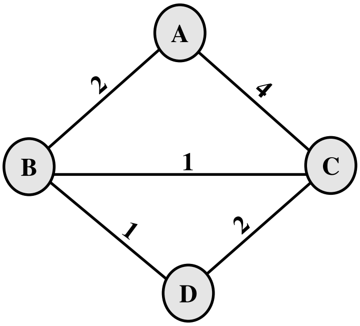
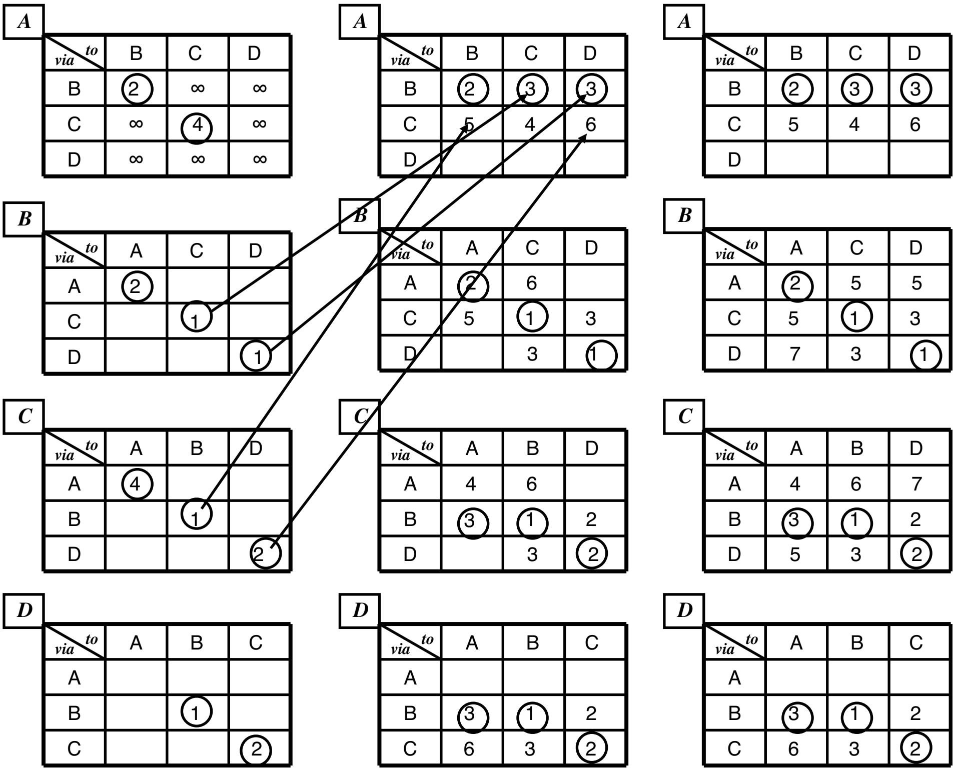


FIGURE 6.37 A simple network with four routers.

only about the immediate neighbors and the cost of connection with them and based on that each router forms its own table. The four tables in the left side of Fig. 6.38 shows the table associated with the start of the operation. In the first iteration of the algorithm, all routers broadcast these tables to all of their immediate neighbors: A to B and C; B to A, C and D; C to A, B, and D; and D to B and C. In the second iteration, each router uses the tables arriving from its neighbors to update its own table and find the shortest path to get to all other routers. The second column of the tables in Fig. 6.38 shows the updated tables



A

via *to*	B	C	D
B	2	∞	∞
C	∞	4	∞
D	∞	∞	∞

B

via *to*	A	C	D
A	2		
C		1	
D			1

C

via *to*	A	B	D
A	4		
B		1	
D			2

D

via *to*	A	B	C
A			
B		1	
C			2

A

via *to*	B	C	D
B	2	3	3
C	5	4	6
D			

B

via *to*	A	C	D
A	2	6	
C	5	1	3
D		3	1

C

via *to*	A	B	D
A	4	6	
B	3	1	2
D		3	2

D

via *to*	A	B	C
A			
B	3	1	2
C	6	3	2

A

via *to*	B	C	D
B	2	3	3
C	5	4	6
D			

B

via *to*	A	C	D
A	2	5	5
C	5	1	3
D	7	3	1

C

via *to*	A	B	D
A	4	6	7
B	3	1	2
D	5	3	2

D

via *to*	A	B	C
A			
B	3	1	2
C	6	3	2

FIGURE 6.38 Formation of routing tables for the network shown in Fig. 6.35 using a distance-vector protocol.

after completion of the second iteration. To go over the details, consider the updating process for the table of router A, a neighbor of the B and C routers, during the second iteration. Using the information arriving in the table from router B, router A discovers that (1) there is another router, D, which was unknown to A previously and (2) router B is connected to C as well. Now A can complete all elements of the row of distances via B. If A wants to approach C via router B, then the distance is the sum of the distance between A and B and distance between B and C, or $2 + 1 = 3$. Similarly, the distance between A and D is 3 if the next router is B. Then router A uses information from router C to complete the second row of its own table associated with distances between A and other routers when the connection is made through router C. Since the distance between A and C is 4, the route going from A to C via B is $4 + 1 = 5$ and the distance between A and D when connection is through C would be $4 + 2 = 6$. These details are also shown in Fig. 3.38. Similarly, packets from A, C, and D are used to update the table for router B and the information from A, B, and D to update the table for router C, and the information from routers B and C to update router D. In each of the four tables in the middle column of the tables in Fig. 6.38, the circled number under each column of a table shows the minimum distance to connect to a specific router. The third column of four tables show the results obtained from the third iteration of the algorithm with the same approach used for previous updates. In this iteration, tables associated with B and C make some changes and the tables for A and D remain unchanged. After a third iteration all the tables stay the same and converge to their final value before any change is made in the network. This way, the information in the final table for each router shows the next router and its associated shortest distance for a packet destined to any other router.

The distance–vector algorithm sometime causes routing loops resulting in a count-to-infinity problem. This problem is can also be shown by an example.

Example 6.7: Looping using Distance–Vector Routing The network topology shown in Fig. 6.39 provides a simple example to explain this problem. This network originally has four routers and for some reason suddenly the link between D and C goes down. When C updates its table it notices that the distance from D has changed from “1” to infinity, indicating a break in the connection. Now assume that C gets an update from B indicating that it has a distance “2” from D because B is not aware that D is not connected anymore. In this situation C receives false information from B, but it has no means to examine its correctness. Therefore, C assumes that it is at distance “3” from D and propagates that information. When that update arrives at B, this router adjusts its distance to “4” and the process gets repeated slowly until the distance to D goes to infinity.

The core of the count-to-infinity problem is that if B tells C that it has a path somewhere, then there is no mechanism in the algorithm to let C know if it is on the path. This problem calls for a method to handle the situation. There are two techniques that are used to prevent routing loops in distance–vector routing protocols. These techniques are referred to as split horizon and split horizon with poison reverse. The split horizon algorithm solves the problem of routing loops and counting to infinity by avoiding advertising routes on the interfaces from which they are learned. In principle, this approach is based on the argument that if a distance is learned from a router then that router must be closer to the destination. Considering our example shown in Fig. 6.39, router B would not advertise back to router C any route that it learned from that router, which breaks the loop. The split horizon with poison reverse technique uses a hop count of infinity if the route was found from the other

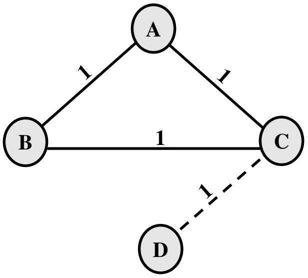


FIGURE 6.39 Infinite loop in distance–vector routing algorithm.

router. In other words, routers advertise routes back on the interfaces from which they were learned, but they change the hop count to infinity. In our example of Fig. 6.39, router B would advertise to router C but it would change the hop count to infinity (a very large number). This way, router B is informing router C that I have router D in my table but I have received that information from you and you had better not use it to update your table. These techniques avoid the formation of routing loops and count to infinity in a number of small-sized networks, but it cannot solve the problem completely in larger networks. Another approach is the refusing of route updates for a few minutes after a route retraction; this works virtually for all cases, but it causes a significant increase in convergence times. We next describe link-state protocols, which provide another alternative to avoid distance–vector protocols problems.

Link-State Routing Algorithm. In the link-state routing algorithm, every router in the network builds a map of the entire network topology along with the link costs and uses that map to determine the shortest paths for routing packets. This is in contrast to the distance–vector protocol, in which each node only keeps the neighboring router information to forwarding a packet. Each router in a network using the link-state protocol broadcasts link-state advertisements which inform all other routers about what networks it is connected to. This way, all routers share the same information database, which shows all connections for all routers. Using the same database and the same algorithm, all routers can build the topology of the network for themselves. When building of the network topology in each router is complete, similar to the distance–vector protocol, the routers update each other periodically. However, the link-state protocol needs much less frequent updates than link-state protocols. Updates are also sent when a change in the network topology is detected. This feature, combined with the fact that routers hold maps of the entire network, results in a much faster convergence for the link-state protocol compared with the distance–vector protocol. Compared with the computational complexity of the distance–vector protocol, routers using the link-state protocol require higher computational speed to calculate the tables and much larger memory to store the entire table. A router using distance–vector protocols only maintains a small database of the neighboring routers and forms the table with much simpler calculations.

In the link-state protocol, the only information passed between the routers is the information used to construct the connectivity maps. This information is processed by every router in the network to construct the map of the connectivity of the network in the form of a graph showing all the connections and costs among the nodes in the network. Then, each node uses the graph to independently calculate the best next hop for every possible destination in the network. The collection of the best next hops forms the routing table for a router. In practice, the link-state protocol is implemented in a hierarchical structure dividing the network into smaller areas so that each router does not need to store the map of the entire network. In the rest of this section we describe the details of a technique for building the network topology in each router and an algorithm to process the tables to find the shortest paths.

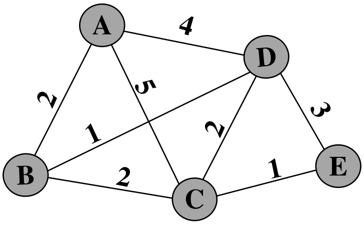
The link-state algorithm starts by building a map of the network in every router. This procedure is performed by a few simple steps. First, each router determines its neighboring nodes using a simple *reachability protocol*, which is executed separately for each direct connection of the router to its neighboring routers. Then each node prepares a message for the *link-state advertisement* and floods that throughout the network. This message is advertised periodically and it identifies the source router and all of its neighbors, as well as a sequence number showing the version of the message. The sequence number is increased each time the source router makes up a new version of the message. When a node receives a link-state advertisement, it looks up the sequence number of the message and its source router. If the sequence number for the source router is more than the sequence number of the existing sequence number for that source router, then the receiving node saves the message and sends a copy to each of its own neighbors.

This simple procedure quickly distributes the latest version of each router's link-state advertisement to every router in the network. After completion of the link-state advertisements, the algorithm builds up the topology of the network showing all nodes and their connections with each other. When a router detects a change in its connectivity to one of its neighbors, such as a link failure, a new link-state message is formed and flooded throughout the network. These changes in the connectivity are detected by the reachability protocol.

After the completion of the network topology map, in each router the link-state algorithm uses the map to construct the routing table in each router. To perform this operation, each node independently runs an algorithm similar to Dijkstra's algorithm for calculation of the shortest path from itself to every other router existing in the network topology map.

The link-state routing algorithm is an iterative process updating databases relating nodes and the distances obtained from link-state advertisements to calculate the shortest distance for each router. The algorithm is based on calculation of the distance of all nodes from a set of nodes which is gradually expanding. Using an example provides a better description of how this algorithm works.

Example 6.8: Link-State Routing Figure 6.40*a* shows a network topology and Fig. 6.40*b* shows the formation of the iterative algorithm to find shortest paths for router A in the graph to all other routers. The algorithm starts with forming the first row of the table for which the node set contains only the source router, {A}. The following columns of the first row show the path to the destination and the distance from the source router. Node D has no direct connection to the set, so the distance is shown by infinity. The closest node to the set is node B with distance 2. In the next iteration the node set includes the shortest distance node and expands to {A, B}. The rest of the columns of row 2 are then filled by closest distances to this node set and their associated paths. Nodes C and D are the neighbors of the {A, B} node set, the minimum distance to both to C using A–B–C path is 4, and the minimum distance to D using A–B–D is 3. Therefore, we add the closer neighbor D to our set to form the new node set {A, B, D} to be used in the next iteration for calculation of the third row of the table. The two nodes outside this set are C and E with minimum distances of 3 and 6 respectively.



(a)

Destination	B		C		D		E	
Nodes set	Path	Distance	Path	Distance	Path	Distance	Path	Distance
{A}	A-B	2	A-C	5	A-D	4	-	∞
{A,B}	A-B	2	A-B-C	4	A-B-D	3	-	∞
{A,B,D}	A-B	2	A-B-C	4	A-B-D	3	A-B-D-E	6
{A,B,C,D}	A-B	2	A-B-C	4	A-B-D	3	A-B-C-E	5
{A,B,C,D,E}	A-B	2	A-B-C	4	A-B-D	3	A-B-C-E	5

(b)

FIGURE 6.40 Example of link-state routing algorithm (*a*) topology of the network (*b*) table construction for calculation of shortest path for router A.

Therefore, the next row is formed for node set {A, B, C, D}, for which the distance to node E will change from A–B–D–E with distance 5 to A–B–C–E with distance 4. E is added to the set in the next iteration, making it {A, B, C, D, E}, which includes all the nodes. The last row of the table in Fig. 6.40*b* shows all the paths and minimum distances from A to all other nodes.

This procedure creates a tree containing all the nodes in the network, with the source node representing the *root* of the tree. The shortest path from that node to any other node and the sequence of routers to get to that node are identified by the algorithm. These paths identify the nodes one traverses to get from the root of the tree to any desired node in the tree. In the link-state protocol, each router determines its own routing table and shortest-path tree independent from other routers. If, for some hardware or software reason, two routers start with different maps, then it is possible to have scenarios in which routing loops may occur, but this event is much more unlikely than with the distance–vector protocol. Examples of link-state routing protocols using Dijkstra's algorithm are the OSPF and IS-IS protocols. OSPF is an IGP, it runs directly over IP, and it has its own reliable transmission mechanism. IS-IS is also an IGP, but it is neutral regarding the type of network addresses and it runs over the DLL.

6.6.4 Multiprotocol Label Switching

MPLS transfers IP packets over a virtual circuit technology which has been designed to work with routers. As shown in Fig. 6.1, part of the long-haul IP traffic is carried through the

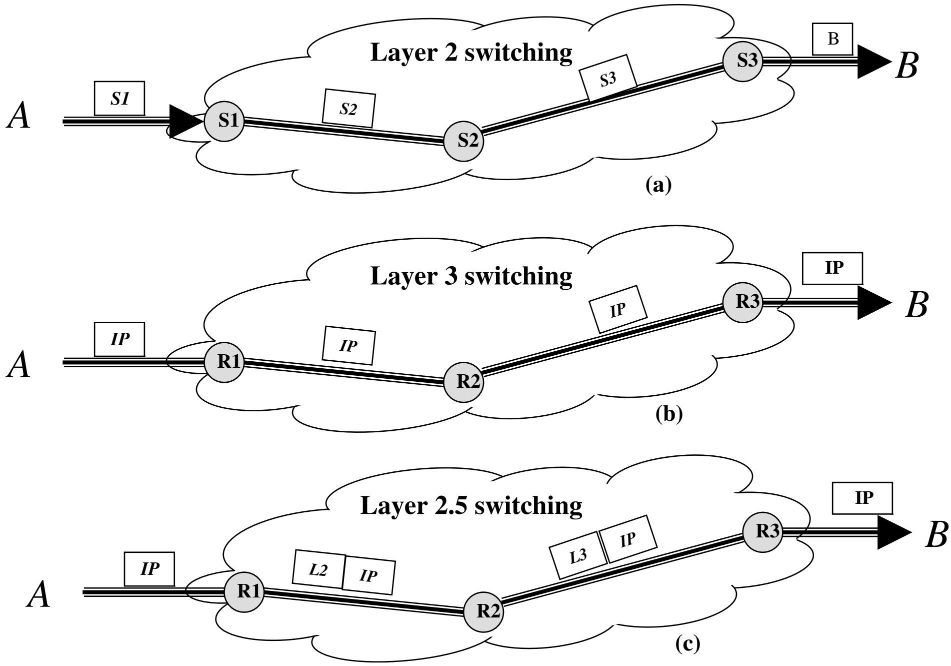


FIGURE 6.41 Packet switching approaches: (a) circuit switching; (b) packet switching; (c) label switching.

virtual circuit-switched ATM which has better handling of the QoS. Figure 6.41*a* shows more details of this transmission approach. A variable-length IP packet is broken into short fixed-length ATM cells at the egress to the long-haul network. The short ATM packets, each carrying a label addressing the next switch on the path, are carried through the network with minimal delay because these switches operate at layer 2 or the link layer by reading the label and directing that according to a preexisting table. In IP packet forwarding using layer 3 routers, shown in Fig. 6.41*b*, the IP address of the packet remains unchanged throughout the transmission. In this approach, the router processes the IP address of each packet to find its next routing address using an IP address look-up table. This process takes more computational time than a label-switching look-up table, resulting in longer delays, which is not desirable for delay-sensitive real-time applications such as VoIP or IP-TV. This approach is more complex and takes more time. In MPLS, shown in Fig. 6.41*c*, each packet carries a virtual circuit identifier, called *label*, enabling virtual circuit switching over the routers. Since it is a combination of layer 3 datagram routing and layer 2 virtual circuit switching, it is also referred to as layer 2.5 switching. MPLS handles labels similar to ATM virtual circuit identifiers. An IP packet arrives at an MPLS-enabled network at location A. The packet is forwarded through an MPLS network or *domain* between A and B. As the IP packet arrives at the first router of the MPLS-enabled network, the egress router uses the source and destination addresses of the packet to establish a virtual circuit for transmission between locations A and B. The address of the next router in the virtual circuit is then added as a part of the MPLS header to the packet. Upon arrival of the packet in the next router, the label is updated by swapping the address of the next router. The last router in the MPLS domain removes the MPLS header and delivers the IP packet to location B.

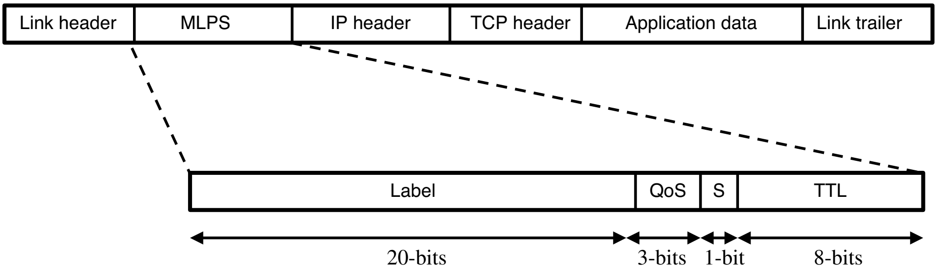


FIGURE 6.42 Details of MPLS header.

Figure 6.42 shows the implementation details of the MPLS header. The header is 32 bits long and it is inserted between the network layer and link layer headers carrying layer 3 (IP) and layer 2 (MAC) routing addresses. The label address takes 20 bits of the 32-bit header. The first 3 bits after the label provides a QoS to handle eight different QoS classes. The next bit "S" indicates whether MPLS has a stacked hierarchy or not. This allows implementation of multiple headers for multi-MPLS networks stacking on top of one another. In other words, using this bit, a packet within an MPLS domain can enter another MPLS domain and carry more than one MPLS label. The "S," or bottom-of-the-stack bit, indicates whether a label is the last in the stack or not. The last 8-bit field in the MPLS header is the Time to Live (TTL) field. Each MPLS router decrements this field and discards packets when the value of this field, initially set by the ingress router, reaches zero. The purpose of this field is to provide a means to avoid packets from indefinitely looping in case a circular virtual circuit was created by mistake. Just like any other virtual circuit-switched network, such as ATM, handling multicast in MPLS is a challenging task.

QUESTIONS

1. What are the differences among ISDN, MAC, and IP addressing techniques?
2. Use Fig. 6.7 to explain how mobile ISDN addressing works.
3. What are broadcast and multicast addressing and how do they get implemented in the Internet?
4. What is ARP and how does it work?
5. In Fig. 6.12, how does the home agent intercepts the packets sent for the mobile host when the mobile host is away from the home network?
6. Why is the PSTN better suited than the Internet for implementation of QoS?
7. What is the role of IEEE 802.1Q in implementation of VLAN tags?
8. What are the similarities and differences between IEEE 802.1D transparent bridges?
9. What is the meaning of filtering in 802.11 bridges?
10. What is the difference between the bridge address and port addresses and, are these general addresses defined by the manufacturers or they are addresses defined by standardization committees?
11. How does one handle mapping of the priorities when they bridge an 802.3 and an 802.5 LAN?
12. How can one assign a cost to a connection between two bridges or two routers?
13. How does the number of queues in a bridge relate to the handling of the priorities?
14. What are the seven traffic types and their associated priorities in 802.1D?

15. What is the purpose of IP-v6 and how does it differ from IP-v4?
16. How does mobile-IP work?
17. What is the difference between functionality of a router in fixed, mobile, and ad-hoc environments?
18. Draw the protocol stack of an ATM end station using LANE to run a TCP/IP application.
19. Name the four classes of bit rates that are provided by ATM switches and identify their differences.
20. What are UNI, NNI, VCI, and VPI in ATM networks?
21. How can LANE handle connections in the ATM network to support connection less LANs?
22. How can the broadcast/multicast functions be implemented in LANE?
23. How many levels of priority and how many different VLANs are supported by 802.1Q tag? How many bytes does this tag add to the 802.3 traditional MAC frames?
24. Draw the details of the MPLS tag and explain what its significance is.
25. Explain the difference between the absolute delay and delay jitter. How we can control the delay jitter for jitter sensitive real-time applications?

PROBLEMS

Problem 1:

(a) What is the maximum number of addresses using IEEE 802 format for MAC addresses?
(b) How many different serial numbers can be assigned for the MAC address by a specific manufacturer?
(c) What is the broadcast MAC address in the IEEE format?

Problem 2:

(a) What is the maximum number of IP-v4 addresses?
(b) How many different Class A, B, and C IP addresses can be produced?
(c) What is the maximum number of IP-v6 addresses?

Problem 3:

(a) A network on the Internet has a subnet mask of 255.255.240.0. What is the maximum number of hosts it can handle?
(b) Give the broadcast IP address for this network.
(c) Give the loopback address for this network.

Problem 4:

We want to assign IP addresses at 198.16.0.0. to three organizations requesting 5000, 3000, and 6000 IP addresses.

(a) Give the first and last IP address assigned to each of the three organizations.
(b) What are the masks used by each of these organizations to filter their traffic?

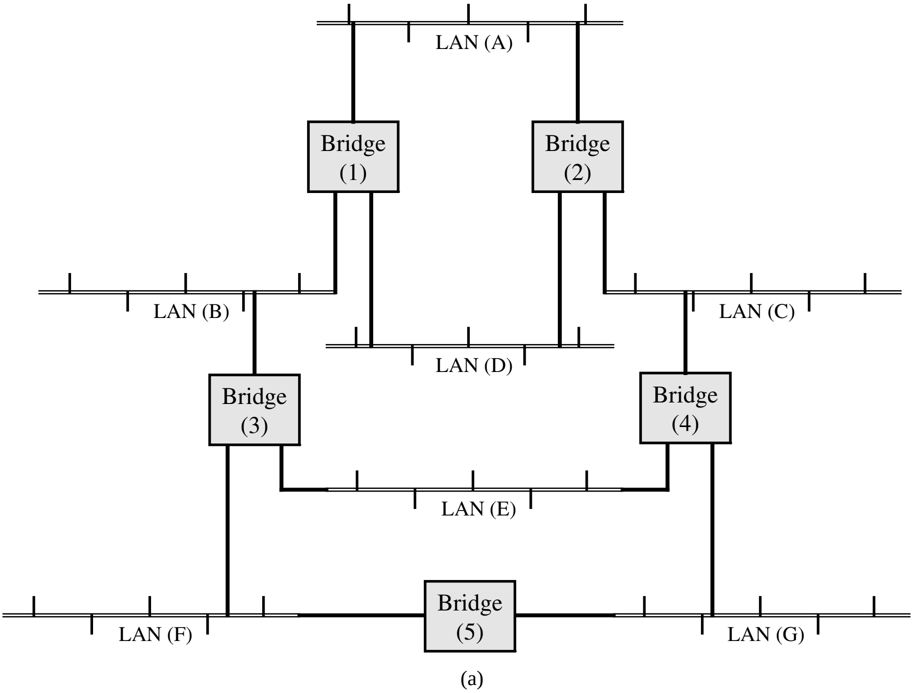


FIGURE P6.1 A local network with seven LANs and five bridges.

Problem 5:

In Fig. 6.19 we assumed bridges are numbered sequentially. In practice each bridge has a unique Bridge ID. Consider Fig. P6.1 and assume that Bridge 1's ID is 45, Bridge 2's 57, Bridge 3's 36, Bridge 4's 62, and Bridge 5's 59. Then apply the STA to the network using the following steps.

(a) Identify the root-bridge and explain how you found it.
(b) Identify the "root port" for each bridge and label that with letter "R".
(c) Identify the "designated port" for each LAN and label that with letter "D".
(d) Use parts a, b, and c to draw the final spanning tree implementation and identify each port by "R", "D" or "B" for blocked.
(e) Is your result the same as Figure 6.20? If not, explain why.

Problem 6:

For the direct graph and link costs shown in Fig. P6.2:

(a) Provide the least-cost routing algorithm table using OSPF when A is the source node.
(b) Show the steps of the OSPF least-cost algorithm for the first two iterations (rows in the table in part (a).

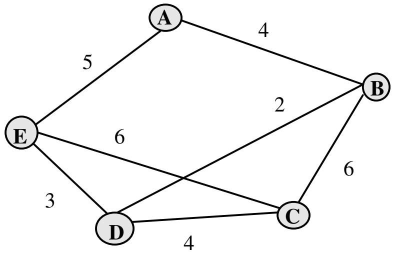


FIGURE P6.2 Graph and link costs of a network.

PROJECT 1: CLIENT-SERVER PROGRAMMING

In this project you will implement a one-way TCP communication channel between two terminals with a server program and a client program. For this purpose you will use the Unix TCP socket programming API to write the server program and the client program. The following is the normal operation of the communication programs:

1. Server continuously listens for the connection request from the client.
2. As the request arrives, the server establishes a TCP connection, then upon successful connection establishment a message "Connection established successfully with *IP_address*:*port_ID*!" will be displayed at both the server and the client sides. The *IP_address* and *port_ID* are the IP address and Port ID of the peer program, respectively, that is, the server displays the IP address and Port ID of the client, and vise versa.
3. client receives keyboard input from user and upon the input of a line of text the client sends the text to the server. Once the text is received, the server displays the text, and then echoes the text back to the client. The client displays the text received from the server.
4. Upon the input of the text "quit" by a user, the server and client first echo and display the text "quit", and then both exit after displaying a message "Program is terminated by user!"

Deliverables:

1. A well commented source code of the server program and the client program.
2. A detailed description of the procedure and format to compile and run the programs.
3. Screen captures of the server and client with complete run-time messages from the start to the end, including "Connection established successfully with *IP_address*: *port_ID*!", "Program is terminated by user!", and a few lines of echoed text that include your name and department.
4. Find tutorials on TCP socket programming and UNIX programming from the supplementary materials provided and find more of your interest on the Internet and in books in the library.
5. You may find prewritten codes on the Internet for client/server applications. However, you are expected to make necessary modifications and deliver the required package.

Extension of the Project:

You can extend this simple project to a "chat room" or "FTP" or many other more elaborate projects which are still suitable for this course. You can come up with a write up proposal to do the project for extra credit. In your proposal you need to justify the educational value and complexity.

7

CELLULAR NETWORKS

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

7.1 INTRODUCTION

In this chapter we discuss example cellular telephone systems to provide the reader with a deeper understanding of the details of how a variety of cellular networks operate. The chapter is focused on a general description of cellular networks followed by the design of the air interface for TDMA and CDMA networks originally designed with a primary focus on voice applications. Then we examine how these air interface designs were modified to accommodate high data rates for emerging data applications.

A *cellular network* uses a number of BSs, which are radio transceivers (you can see them as towers along highways), each serving a separate *cell* to provide a comprehensive wide area coverage to a wireless MS user. The 1G cellular networks were designed for circuit-switched voice applications using analog modulation techniques. The 2G cellular networks resorted to digital technology using TDMA and CDMA technologies. Although the pan-European GSM TDMA technology became the most popular 2G technology, the technology of choice for 3G networks became CDMA, originally designed at QUALCOMM in the USA. The objective of 3G cellular networks was to support higher data rates of up to 2 Mb/s, increase system capacity, and provide further integration of diversified services. Format flexibility, wider carrier bandwidth, higher capacity, and better quality of voice were the major reasons behind the selection of CDMA technology for 3G networks.

7.1.1 The Cellular Concept

Cellular topology is a special case of an infrastructure multi-BS network configuration that exploits the *frequency reuse* concept. The radio spectrum is one of the scarcest resources available, and every effort has to be made to find ways of utilizing the spectrum efficiently and to employ architectures that can support as many users as theoretically possible with the available spectrum. This is extremely important, especially today, in light of the huge demand for capacity. Spatially reusing the available spectrum so that the same spectrum can support multiple users separated by a distance is the primary approach for efficiently using the spectrum. This is called *frequency reuse*. Employing frequency reuse is a technique that has its foundations in the attenuation of the signal strength of electromagnetic waves with distance. For instance, in a vacuum or free space, the signal strength falls as the square of the distance. This means that the same frequency spectrum may be employed without any interference for communications or other purposes, provided the distance separating the transmitters is sufficiently large and their transmit powers are reasonably small (depending on the reuse distance). This technique has been used, for example, in commercial radio and television broadcast, where the transmitting stations have a constraint on the maximum power they can transmit so that the same frequencies can be used elsewhere. The cellular concept is an intelligent means of employing frequency reuse. Cellular topology is the dominant topology used in all large-scale terrestrial and satellite wireless networks. The concept of cellular communications was first developed at the Bell Laboratories in the 1970s to accommodate a large number of users with a limited bandwidth.

By cellular radio, we mean deploying a large number of low-power BSs for transmission, each having a limited coverage area. In this fashion, the available capacity is multiplied each time a new BS or transmitter is set up, since the same spectrum is being *reused* several times in a given area. The fundamental principle of the cellular concept is to divide the coverage area into a number of contiguous smaller areas each served by its own radio BS. Radio channels are allocated to these smaller areas in an intelligent way so as to minimize the interference, provide an adequate performance, and cater to the traffic loads in these areas. Each of these smaller areas is called a *cell*. Cells are grouped into *clusters*. Each cluster utilizes the entire available radio spectrum. The reason for clustering is that adjacent cells cannot use the same frequency spectrum because of interference. So, the frequency bands have to be split into chunks and distributed among the cells of a cluster. The spatial distribution of chunks of radio spectrum (which are called sub-bands) within a cluster has to be done in a manner such that the desired performance can be obtained. This forms an important part of network planning in cellular radio.

Two types of interference are important in such a cellular architecture. The interference due to using the same frequencies in cells of different clusters is referred to as *co-channel interference*. The cells that use the same set of frequencies or channels are called *co*-channel cells. The interference from frequency channels used within a cluster whose side lobes overlap is called *adjacent channel interference*. The allocation of channels within the cluster and between clusters must be done so as to minimize both of these.

The cellular concept can increase the number of customers that can be supported in the available frequency spectrum, as illustrated by the following examples, by deploying several low-power radio transmitters.

Example 7.1: Cellular Concept Consider a single high-power transmitter (see Figure 7.1) that can support 35 voice channels over an area of 100 km^2 with the available spectrum. If seven lower power transmitters are used so that they support 30% of the channels over an area of 14.3 km^2 each, then a total of ~80 voice channels are now available in this area instead of 35. In reality, channels will have to be allocated to BSs in such a way as to prevent interference between one BS and another. In Fig. 7.1, BSs 1 and 4 could use the same channels, as their coverage areas are sufficiently far apart, and so also BSs 3 and 6.

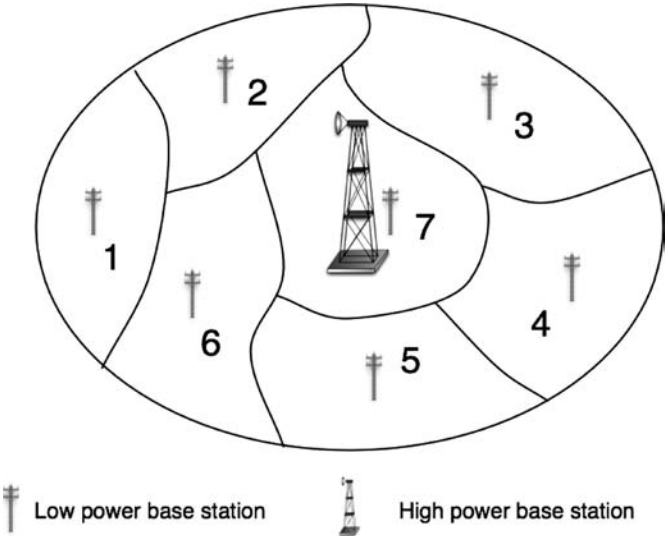


FIGURE 7.1 The cellular concept.

Suppose the cells labeled 1, 2, 5, 6, and 7 use disjoint frequency bands and the channels used in 1 and 6 are reused in 3 and 4; the set of cells $\{1, 2, 5, 6, 7\}$ then forms a cluster, and cells 3 and 4 form part of another cluster. In the limiting case, the density of BSs can be made so large that the capacity is infinite. However, in practice, this is impossible for several reasons, which include drastic increases in the network and signaling load, number and frequency of handoffs, and cost of infrastructure and planning.

Example 7.2: Importance of Cellular Topology We want to provide a radio communication service to a city. The total bandwidth available is 25 MHz and each user requires 30 kHz of bandwidth for voice communication. If we use one antenna to cover the entire town, then we can only support 25 MHz/30 kHz = 833 simultaneous users. Now let us employ a cellular topology where 20 lower power antennas are opportunistically located to minimize both kinds of interference. We divide our frequency band into four sets and assign one set to each cell. Each cell has a spectrum of 25 MHz/4 = 6.25 MHz allocated to it. We have a *cluster* of four cells in this example. The number of simultaneous users supported per cell is 6.25 MHz/30 kHz = 208.

The number of users per cluster is 4 × 208 = 832. The total number of simultaneous users is now 832 × 5 = 4160, since we have five clusters of four cells each. The new capacity is roughly five times the capacity with a single antenna.

Examples 7.1 and 7.2 illustrate the main benefits and elements of cellular network planning by relating the number of users, bandwidth, number of users per carrier, frequency reuse factor, and capacity of the network. As we discussed in Section 5.2.4, if W is the total available spectrum, B is the bandwidth needed per user, N_{f} is the frequency reuse factor or cluster size, and m is the number of users per carrier, then the number of simultaneous users per cell is given by

$$M = \frac{W}{B}\frac{m}{N_{\mathrm{f}}} \tag{7.1}$$

The capacity of the network for a service provider is the number of simultaneous users per cell multiplied by the number of cells. A service provider operating in a geographical region can increase the capacity by reducing the size of the cells, which increases the number of cells. The service provider can also increase the capacity by increasing m, the number of users per carrier, and decreasing the frequency reuse factor (number of cells per clusters N_{f}). A major remaining question at this point is how to assign the groups of sub-bands to individual cells so that interference between different users using the same sub-bands is acceptable. We address this issue in Section 7.8.1.

A cellular topology reduces the coverage requirements of both the mobile terminal and the BS. The reduction of the size of coverage lowers the required transmitted power by the mobile terminal, since mobile terminals are located closer to the BSs and they require less power to communicate with the network. This increases the battery lifetime and reduces the size of a terminal. These issues are extremely important to the user of a hand-held terminal. Therefore, the larger the number of cells, the larger the capacity is and the smaller the size of the hand-held terminal is. However, we need a fixed network infrastructure to interconnect the cells and ensure that the entire system works in a coordinated manner. As we increase the number of cells, the cost and the time for deploying the network increase. In addition, the smaller the size of the cell, the larger is the frequency of a mobile changing its connection from one BS to another (rate of handoffs). Therefore, a reduction in the size of the cells increases the complexity of the design and deployment of the network, as well as the signaling load in the fixed part of the infrastructure. The art of designing a cellular topology involves striking a balance between all these elements, and this is the subject of details that follow later in this chapter.

7.1.2 Cellular Hierarchy

There are three reasons to use a hierarchical cellular infrastructure supporting cells of different sizes. One is to extend the coverage to the areas that are difficult to cover by a large cell. For example, cells designed to cover suburban areas have antennas with tall towers and cover a large area. Signals from these antennas, however, cannot propagate sufficiently into urban canyons or indoor environments. For urban canyons we need to install antennas at lower heights, and in indoor areas we may mount the antennas on walls to provide a comprehensive coverage. Antennas mounted in these locations are of low power and cover a smaller area, resulting in the creation of a smaller sized cell. The second reason to have a cellular hierarchy is to increase the capacity of the network for those areas that have a higher density of users. Imagine the number of cellular phone users in the World Trade Center and compare it with the number of mobile users on an interstate highway. To support the larger subscriber demand and higher traffic in smaller areas we need to increase the number of cells by reducing their sizes. The third reason is that sometimes an application needs certain coverage. Consider the increasing number of wireless devices that we are carrying in our bag these days and the increasing need for communication between these devices. This necessitates extremely small-sized cells that provide a wireless network for connecting laptops or notepads to cellular phones.

In a modern deployment of a cellular network, a number of cell sizes are used to provide a comprehensive coverage supporting traffic fluctuations in different geographical areas and supporting a variety of applications. One way of dividing the cells into a hierarchy is to define the following cell sizes:

- *Personal cells* These are the smallest unit of the cellular hierarchy, used for connection of personal equipment such as laptops, notepads, and cellular telephones. These cells need to cover only a few meters, where all these devices are in the physical range of the user.
- *Picocells* These are small cells inside a building that support local indoor networks, such as WLANs. The size of these networks is in the range of a few tens of meters.
- *Microcells* These cells cover the inside of streets with antennas mounted at heights lower than the rooftops of the buildings along the streets. They cover a range of hundreds of meters and are used in urban areas to support PCSs.
- *Macrocells* Macrocells cover metropolitan areas and they are the traditional cells installed during the early phases of cellular telephony. These cells cover areas on the order of several kilometers and their antennas are mounted above the rooftops of typical buildings in the coverage area.
- *Megacells* Megacells cover nationwide areas with ranges of hundreds of kilometers and are mainly used with satellites.

Figure 7.2 illustrates the relationship between different cells with example applications. An ideal network has a hierarchy of these cells to cover airplane travelers with megacells,

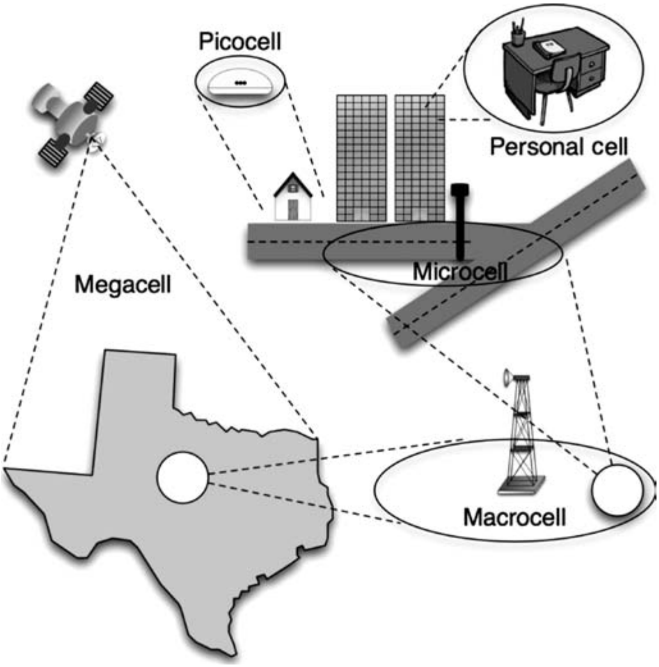


FIGURE 7.2 Cellular hierarchy.

car drivers in suburban areas with macrocells, pedestrians in the streets via microcells, indoor users with picocells, and connect personal equipment with personal cells. The focus of this chapter is on cellular telephone networks that typically deploy macrocells to cover their service areas.

The fundamentals of the wired infrastructure and architecture of the CDMA and TDMA networks are quite similar. Therefore, we start with a general description of a cellular network and then we resort to description of the air interface for the TDMA and CDMA networks. The 3G networks have standards for the core network (i.e. the wired part of the cellular network) that reuse the protocols and standards from the IETF that rely on IP as the underlying network layer protocol for all communications.

7.2 GENERAL ARCHITECTURE OF A CELLULAR NETWORK

Specifications of cellular wireless networks are complex, with detailed descriptions of the terminal, fixed hardware backbone, and software databases that are needed to support the operation. To describe such a complex system, at the beginning, a reference model or overall architecture is needed to provide an understanding of the network elements and operation, and later it is possible to divide the system into subsystems. Figure 7.3 presents an overall view of an example reference model with typical hardware and software elements used in a cellular network divided into three segments. These segments are the MS, the BSS, and the network and switching subsystem (NSS). Figure 7.4 provides a more physical representation

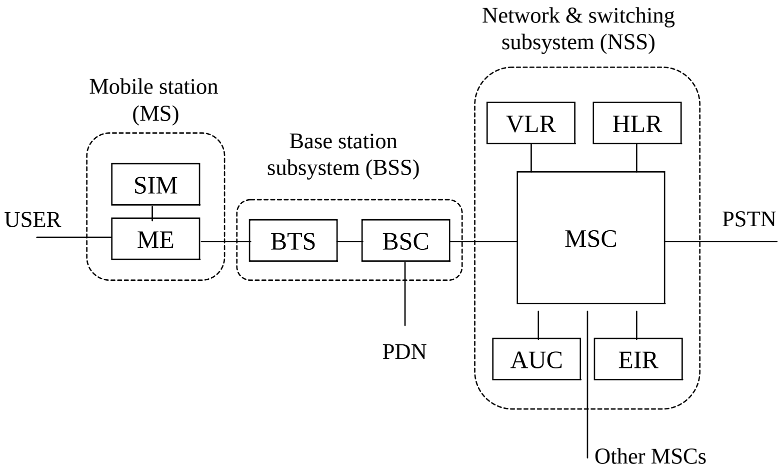


FIGURE 7.3 An overall reference model illustrating typical hardware and software (database) architectural elements of a cellular network.

of the architectural elements of network and the relationships among these elements. This division of the architectural elements was adopted from Haug [Hau94] and we will follow that for the description of the system elements. The description of the elements of this reference model now follows.

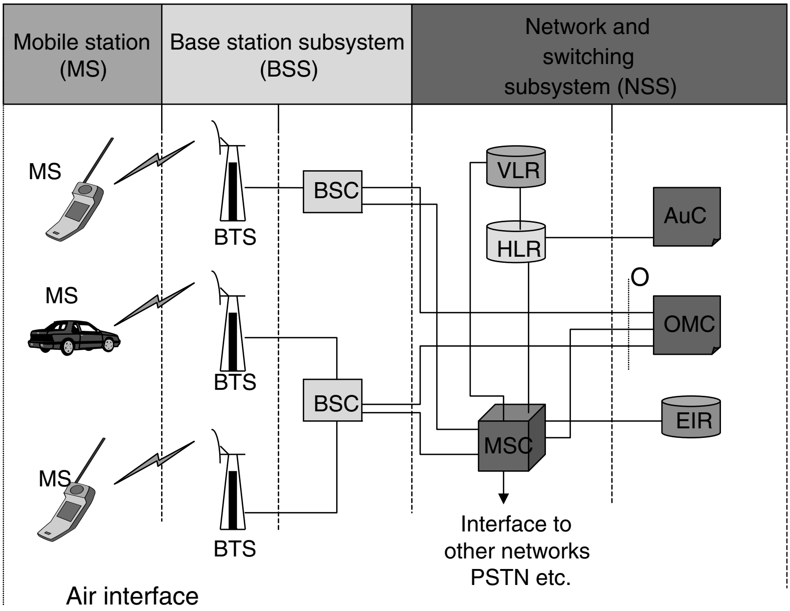

Mobile station (MS)
Base station subsystem (BSS)
Network and switching subsystem (NSS)
MS
MS
MS
BTS
BTS
BTS
BSC
BSC
VLR
HLR
AuC
O
OMC
EIR
MSC
Interface to other networks PSTN etc.
Air interface

FIGURE 7.4 A different view of the reference architecture for cellular networks.

7.2.1 Mobile Stations

MSs form the interface for exchanging information with the user and modify the information on to the transmission protocols of the air interface to communicate with the BS subsystem (BSS). The user information is communicated with the MS through a microphone and speaker for speech, keypad and display for short messaging, and cable connections to data terminals for other applications. The MS has two elements. The first element is the *mobile equipment* (ME), which is a piece of hardware that the customer purchases from the equipment manufacturer or their dealers. This hardware piece contains all the components needed for the implementation of the protocols to interface with the user and the air interface to the BSS. The components include speakers, microphones, keypad, and the radio modem. Therefore, the ME is an expensive piece of hardware. To encourage more users to subscribe to the wireless services, a number of service providers in the early days of the cellular industry, and even today, subsidized the price of ME.

The second element of the MS in the reference model of Fig. 7.3 is the *subscriber identity module* (SIM), which is a smart card issued at subscription time identifying the specifications of a user, such as address and type of service. The calls in the system are directed to the SIM rather than the terminal. Short messages are also stored in the SIM card. Although implementing a SIM is a fairly simple concept, it has a significant impact on the way that a user transacts with the service provider. Each SIM card carries the user's personal information that enables a number of useful applications. For example, people visiting different countries but not keen on making calls with their home number can always carry their own terminal and purchase a SIM card in every country that they visit. This way they avoid roaming charges at the expense of having a different contact number. Since SIM cards carry the private information for a user, a security mechanism is implemented in the network that asks for a four-digit personal identification number (PIN) to make the information on the card available to the user. This can sometimes play a negative role, because users are not keen on remembering too many passwords. Using SIM cards was not possibility with analog cellular systems, and some systems, such as the North American digital cellular standards, have not implemented this option.

7.2.2 The Base Station Subsystem

The BSS communicates with the MS through the wireless air interface and with the wired infrastructure through wired protocols. In other words, it translates between the air interface and fixed wired infrastructure protocols. The needs of the wireless and wired media are different, since the wireless medium is unreliable, bandwidth limited, and needs to support mobility. As a result, protocols used in the wireless and wired mediums are different. The BSS provides for the translation among these protocols. For example, consider speech conversion. The user's speech signal at the MS is converted into a voice-encoding scheme around 10 kb/s (for bandwidth efficiency) with a speech coder and is communicated over the air interface. The backbone wired network uses 64 kb/s PCM digitized voice in the PSTN hierarchy. Conversion from analog to 10 kb/s voice takes place at the MS and the change from 10 to 64 kb/s coding takes place at the BSS. As another example, the signaling format to establish a connection in wired networks is the multitone frequency scheme used in the POTS. Digital cellular systems, on the other hand, establish the call through the exchange of a number of packets. The translation of this communication into a dialing signal is made in the BSS.

As with speech coding and dialing, explained above, data transmission protocols over the air interface are different from those in the wired infrastructure. All these translations are performed at the BSS. To implement packet data services on the same air interface, the BSS also separates packet-switching data from the circuit-switched PSTN traffic to direct them towards the PSTN and the Internet separately.

There are two architectural elements in the BSS. The *base transceiver system* (BTS) is the counterpart of the MS for physical communication over the air interface. The BTS components include a transmitter, a receiver, and signaling equipment to operate over the air interface and it is physically located in the center of the cells where the BSS antenna is installed. The second architectural element of the BSS is the *BS controller* (BSC), which is a small switch inside the BSS in charge of frequency administration and handover among the BTSs inside a BSS. The hardware of the BSC in a single BTS site is located at the antenna and in multi-BTS systems in the switching center, where other hardware elements of the NSS are located. One BSS may have anywhere from one up to several hundred BTSs under its control [Red95].

7.2.3 The Network and Switching Subsystem

The NSS is responsible for network operation. It provides for communications with other wired and wireless networks as well as support for registration and maintenance of the connection with the MSs. The NSS could be interpreted as a wireless specific switch that communicates with other switches in the PSTN and at the same time supports functionalities that are needed for a cellular mobile environment. The NSS is the most elaborate element of the cellular network; it has one hardware element, i.e. an MSC, and four software elements, namely a visitor location registration (VLR), a home location register (HLR), an equipment identification register (EIR), and an authentication center (AuC).

The MSC is the hardware part of the wireless switch that can communicate with PSTN switches using the SS7 protocol, commonly used in the PSTN, as well as other MSCs in the coverage area of a service provider. The MSC also provides to the network the specific information on the status of the mobile terminals.

The HLR is database software that handles the management of the mobile subscriber account. It stores the subscribers' addresses, service type, current location, forwarding addresses, authentication/ciphering keys, and billing information. In addition to the telephone number for the terminal, the SIM card is identified with an international mobile subscriber identity (IMSI) number, which is totally different from the ISDN telephone number. The IMSI is used totally for internal cellular networking applications. For example, the telephone number of a subscriber in Finland could be 358-40-770–5246. The first three digits are the country code, the next two are the digits for the specific MSC, and the rest are the telephone number. The IMSI of the same user can be 244-91 followed by a 10-digit number that is totally different from the actual telephone number. The first three digits of the IMSI identify the country, namely Finland, and the next two digits are the billing company.

The VLR is temporary database software similar to the HLR identifying the subscribers that are visiting inside the coverage area of an MSC. The VLR assigns a temporary mobile subscriber identity (TMSI) that is used to avoid using IMSI on the air to protect the user billing number. Maintenance of two databases at home and at the visiting site allows a mechanism to support call routing and dialing in a roaming situation where the MS is visiting the coverage area of a different MSC. The mechanism of holding two databases (home and visiting) to support mobility is used in almost all mobile networks.

The AuC holds different algorithms that are used for authentication and encryption of the subscribers. Different classes of SIM cards have their own algorithms and the AuC collects all of these algorithms to allow the NSS to operate with different terminals from different geographical areas.

The EIR is another database managing the identification of the ME against faults and theft. This database keeps the international ME identity that reveals the manufacturer, country of production, and terminal type. Such information can be used to report stolen phones or check if the phone is operating according to the specification of its type. The implementation of the EIR is left optional to the service provider.

7.3 MECHANISMS TO SUPPORT A MOBILE ENVIRONMENT

Now that we have described all the hardware and software elements of a typical cellular network we can describe how the different functionalities of the network are implemented with these elements. Four mechanisms are embedded in all circuit-switched voice-oriented wireless networks that allow a mobile to establish and maintain a connection with the network. These mechanisms are *registration*, *call establishment*, *handoff* (or handover), and *security*. Registration takes place as soon as one turns the mobile unit on in a new environment, call establishment occurs when the user initiates or receives a call, handoff helps the MS to change its connection point to the network, and security protects the user from fraud and eavesdropping. In this section we describe the details of their implementation with examples using the reference architecture that was described in the last section. To illustrate the complexity of cellular wireless networks, when we discuss registration and call establishment, we compare these mechanisms with their counterparts in POTS.

7.3.1 Registration

When we subscribe to a POTS, the telephone company brings a pair of wires to our home that is connected to a port of a switch in a PSTN end office. Then our telephone number is registered in a database in the network and our registration is fixed. Therefore, the connection and registration process for a wired access to the network is a one-shot operation after which connection is active and registration is valid as long as subscription to the service is valid. With wireless access to a cellular network, each time that we turn on the MS, we need to establish a new connection and possibly establish a new registration with the network. We may actually connect to the network at different locations through a BS that may not be owned by our service provider. Therefore, a wireless network needs a registration process that is far more complex than the registration in wired networks.

Technically speaking, as we turn on an MS it passively synchronizes to the frequency; bit and frame timings of the closest BS get ready for information exchange with the BS. After this preliminary setup, the MS reads the system and cell identity to determine its location in the network. If the current location is not the same as before, then the MS initiates a *registration* procedure. During a registration procedure, the network provides the MS with a channel for preliminary signaling. The MS provides its identity in exchange for the identity of the network, and finally the network authenticates the MS. The simplest

Steps	MS	BTS	BSC	MSC	VLR	HLR
1. Channel request	→	→				
2. Activation response		←				
3. Activation ACK		→				
4. Channel assigned	←	←				
5. Location update request	→	→	→			
6. Authentication request	←	←	←			
7. Authentication response	→	→	→			
8. Authentication check				↔		
9. Assigning TMSI	←	←	←			
10. ACK for TMSI	→	→	→			
11. Entry to VLR and HLR				←	→	
12. Release a signaling channel	←	←				

FIGURE 7.5 Registration procedure in a typical digital cellular network.

connection takes place if the MS is turned on in the previous area, and the most complex registration process occurs when the mobile is turned on in a new MSC area which needs changes in the entries of the VLR and HLR. The following example illustrates the complexity of the registration process when a mobile is turned on in a new MSC.

Example 7.3: The Registration Procedure Figure 7.5 shows the 12-step registration process in a typical digital cellular network that takes place when an MS is turned on in a new MSC area. During each step, a message is carried through certain elements of the overall infrastructure of the network. In the first four steps, a radio channel is established between the MS and BSS to process the registration. In the next four steps, the NSS authenticates the MS. In the next three steps, a TMSI is assigned and adjustments are made to the entries in the VLR and HLR. In the final step, the temporary radio channel for communication is released and transmission starts over a traffic channel.

7.3.2 Call Establishment

Call establishment in POTS starts with a dialing process that transfers the number to the nearest PSTN switch where a routing algorithm finds the best connection through intermediate switches to the destination. After establishment of the link, the last switch (end office) at the destination sends a signal back to the source to announce whether the destination is available or busy, which is signaled to the user at the source. When the destination POTS terminal is off-hook, another signal is sent to the source end-office to stop the waiting tone and establish the traffic line. In the mobile environment we have two separate call establishment procedures for mobile-to-fixed and fixed-to-mobile calls. Mobile-to-mobile calls are a combination of the two. The following two examples provide the detailed procedure in a typical network for both types of call establishment.

Example 7.4: Mobile-originated Call The five-step procedure in POTS for call setup changes to a 15-step mobile-originated call establishment procedure in a typical digital cellular network. As shown in Fig. 7.6, the first five steps are similar to the registration

Steps	MS	BTS	BSC	MSC
1. Channel request	→	→		
2. Channel assigned	←	←		
3. Call establishment request	→	→	→	
4. Authentication request	←	←	←	
5. Authentication response	→	→	→	
6. Ciphering command	←	←	←	
7. Ciphering ready	→	→	→	
8. Send destination address	→	→	→	
9. Routing response	←	←	←	
10. Assign traffic channel	→	→		
11. Traffic channel established	←	←		
12. Available/busy signal	←			
13. Call accepted	←	←	←	
14. Connection established	→	→	→	
15. Information exchange	←	—	—	→

FIGURE 7.6 Mobile-originated call.

process, except that these are done to prepare for call establishment. The next two steps start ciphering (encryption) to provide a protection against eavesdropping. The rest of the steps are similar to those in wired networks, except that we have an additional traffic channel assignment procedure.

Example 7.5: Mobile-terminated Call The most complicated call establishment is for the situation where a fixed telephone dials a mobile visiting another MSC. As shown in Fig. 7.7, after dialing, the PSTN directs the call to the MSC identified by the destination address. The MSC requests routing information from the HLR. Since, in this case, the mobile is roaming in the area of a different MSC, the address of the new MSC is given to the MSC and it

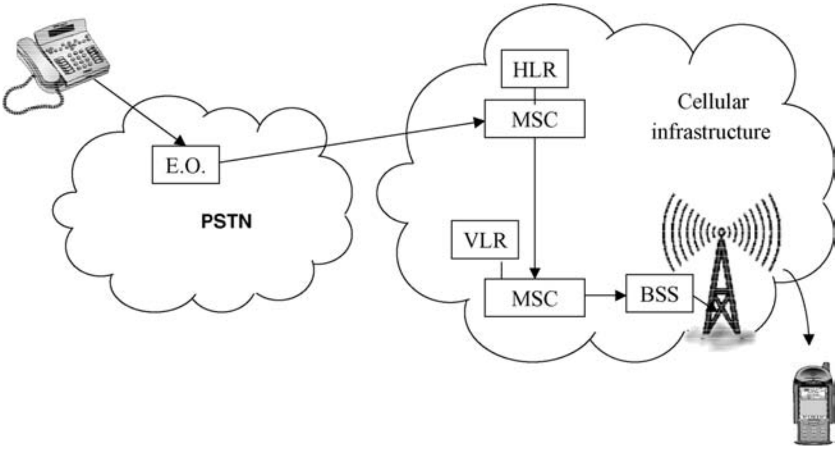


FIGURE 7.7 Mobile-terminated call in a visiting network.

contacts the new MSC. At the destination MSC, the VLR initiates a paging procedure in all BSSs under the control of the MSC holding the registration. After a reply from the MS, the VLR sends the necessary parameters to the MSC to establish the link to the MS.

7.3.3 Handoff

Handoff refers to the switching of connections by an MS from one point of connection to the fixed network to another. The MS is initially connected to a BS and then it moves out of the service area of the BS to that of another BS. When this occurs, there must be protocols in place to *detect* that the MS is in a new service area, to *initiate* the handoff, and finally to *complete* the switch from the old BS to the new BS. There are two types of handoff: internal and external. Internal handoff is between BTSs that belong to the same BSS and external handoffs are between two different BSSs belonging to the same MSC. Sometimes there are handoffs between BSSs that are controlled by two different MSCs. In such a case, the old MSC continues to handle call management. Roaming between two MSCs in two different countries is prohibited and the call simply drops.

Handoff is initiated for a variety of reasons. Signal strength deterioration is the most common cause for handoff at the edge of a cell. Other reasons include traffic balancing, where the handoff is network oriented to ease traffic congestion by moving calls in a highly congested cell to a lightly loaded cell. The handoff could be synchronous where the two cells involved are synchronized or it may be asynchronous. Since the MS does not have to resynchronize itself in the former scenario, the handoff delay is much smaller (100 ms against 200 ms in the asynchronous case). The decision to make a handoff typically is the responsibility of the BS controller or the MSC.

Figure 7.8 shows the details of messages for a handoff procedure between two BSSs that are controlled by one MSC in a typical cellular network. The details of handoff in different cellular networks vary substantially and depend on the technology. For example, GSM using TDMA technology adopts mobile-assisted handoff. The BTS provides the MS with a list of available channels in neighboring cells. The MS monitors the RSS from these neighboring cells and reports these values to the MSC. The BTS also monitors the RSS through

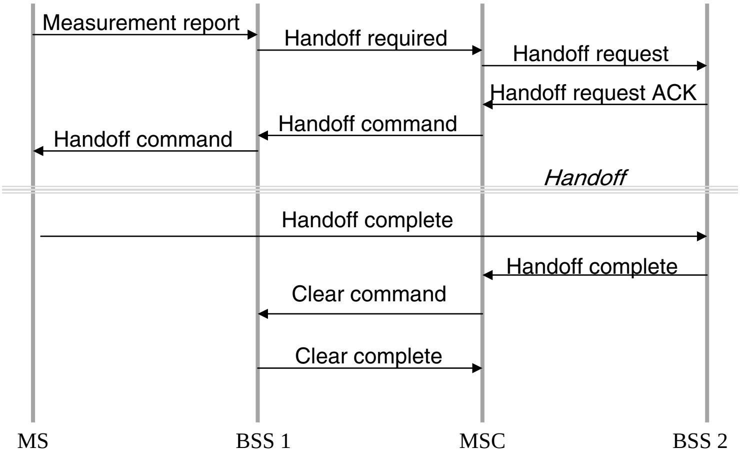


FIGURE 7.8 Handoff involving a single MSC but two BSSs.

messages received from the MS to make handoff decisions. Proprietary algorithms are used to decide when a handoff should be initiated. If a decision to make a handoff is made, then the MSC negotiates a new channel with the new BSS and indicates to the MS that a handoff should be made using a handoff command. Upon completion of the handoff, the MS indicates this with a handoff complete message to the MSC.

The CDMA networks employ soft handoff, which refers to the process by which an MS is in communication with multiple candidate BSs before finally communicating its traffic through one of them. The reason for implementing soft handoff has its basis in the near–far problem and the associated power control mechanism. If an MS moves far away from a BS and continues to increase its transmit power to compensate for the near–far problem, then it will very likely end up in an unstable situation. It will also cause a lot of interference to MSs in neighboring cells. To avoid this situation and to ensure that an MS is connected to the BS with the largest RSS, a soft handoff strategy is implemented. An MS will continuously track all BSs nearby and communicate with multiple BSs for a short while if necessary before deciding which BS to select as its point of attachment.

The soft handoff procedure involves several BSs. A controlling primary BS coordinates the addition or deletion of other BSs to the call during soft handoff. Figure 7.9 shows the setup and ending of handoff in a two-way soft handoff. The MS detects a pilot signal from a new BS and informs the primary BS. After a traffic channel is set up with the new BS, a frame selector join message is used to select a signal from both BSs at the BSC/MSC. After a while, the pilot signal from the old BS starts falling and the MS will request its removal, which is achieved via a *frame selector remove* message.

The pilot channels of each cell are involved in the handoff mechanism. The reason behind this is that the pilot channel provides MSs with a measure of the RSS. The MS maintains a list of pilot channels that it can hear and classifies them into different sets. Based on RSS values, thresholds, and timers, a handoff algorithm will decide which BSs or cell sectors to connect to at which time.

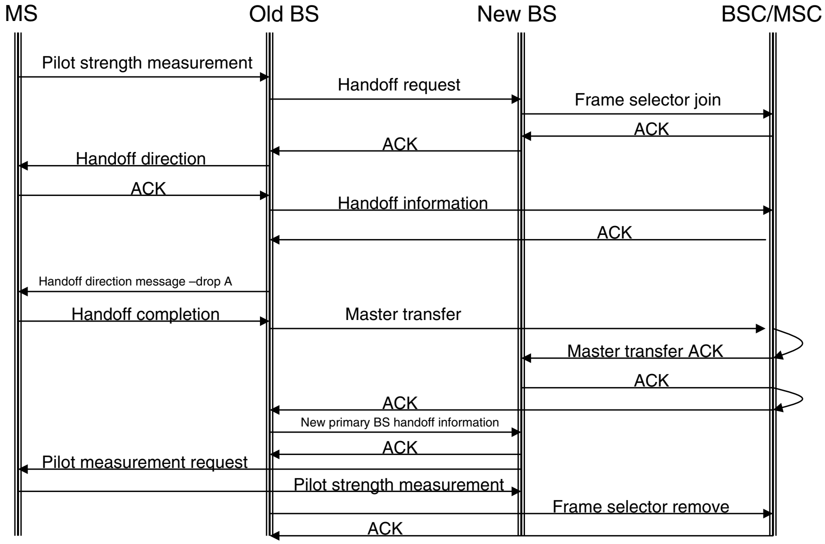


FIGURE 7.9 Setup and ending of soft handoff.

7.3.4 Security

Security in cellular systems is usually employed to prevent fraud via authentication, to avoid revealing the subscriber number over the air, and to encrypt conversations where possible. All of these are achieved using proprietary (secret) algorithms.

Security requirements for wireless communications are very similar to the wired counterparts (see Chapter 11 for details of security and encryption), but they are treated differently because of the applications involved and the potential for fraud. Different parts of the wireless network need security. Over the air, security is usually associated with privacy of voice conversations in cellular systems. This is changing with the increasing use of wireless data services. Message authentication, identification, authorization, etc. also become issues with cellular networks and WLANs. Wireless networks are inherently insecure compared with their wired counterparts. The broadcast nature of the channel makes it easier to be tapped. Analog telephones are extremely easy to tap, and conversations can be eavesdropped using an RF scanner. Digital systems such as TDMA and CDMA are much harder to tap and RF scanners can no longer do the trick; but since the circuitry and chips are freely available, it may not be hard for someone to break into a system that does not employ security.

Privacy Requirements of Cellular Networks. A variety of control information is transmitted over the air in addition to the actual voice or data, including call setup information, user location, user ID (or telephone number) of both parties, etc. These should all be kept secure, since there is potential for misusing such information. Calling patterns (traffic analysis) can yield valuable information under certain circumstances. A flurry of calls between the CEOs of two major companies may indicate certain trends if it was discovered, even if the actual information in the calls was secure. Hiding such information is also important. Various levels of privacy are defined for voice communications [Wil95b].

We commonly assume that all telephone conversations are secure. This is not true. Whereas it is possible to detect a tap on a wireline telephone, it is impossible to detect taps over a wireless link because of the broadcast nature of the medium. To provide privacy that is equivalent to that of a wired telephone, for routine conversations it may be sufficient to employ some sort of an encryption that will take more than simple scanning and decoding to decrypt. Wilkes [Wil95b] defines two levels of security: Level-0 and Level-1. Level-0 privacy is when there is no encryption employed over the air, so that anyone can tap into the signal. Level-1 privacy provides privacy equivalent to that of a wireline telephone call, one possibility being encrypting the over-the-air signal. In order to alert wireline callers about the insecure nature of a wireless call that is *not at all* encrypted, a “lack-of-privacy” indicator was suggested. For commercial applications, a much stronger encryption scheme would be required that would keep the information safe for more than several years. Secret-key algorithms with key sizes larger than 80 bits are appropriate for this purpose. This is referred to as Level-2 privacy in [Wil95b]. Encryption schemes that will keep the information secret for several hundreds of years are required for military communications and fall under Level-3 privacy. For wireless data networks, a bare minimum level would be to keep the information secure for several years. The primary reason for this is that wireless electronic transactions are becoming common. Credit card information, dates of birth, social security numbers, e-mail addresses, etc. can be misused (fraud) or abused (junk messages for example). Consequently, such information should never be revealed easily. A Level-2 privacy will be

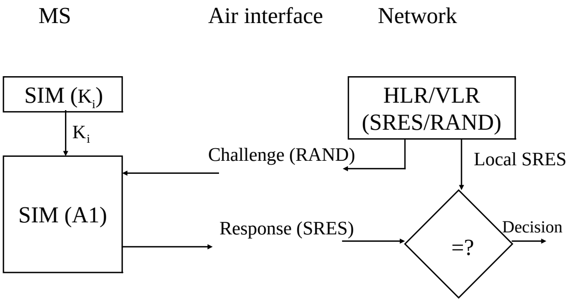


FIGURE 7.10 Basic principles of authentication in a cellular network.

absolutely essential for wireless data networks. In certain cases, a Level-3 privacy is required. Examples are wireless banking, stock trading, mass purchasing, etc. Most cellular systems and WLANs employ strong encryption today, and this is becoming a moot point.

Authentication and Integrity. While privacy and confidentiality continue to be the important issue in wireless networks, other security requirements are becoming significant in recent times. There has been widespread fraud and impersonation of analog cellular telephones in the past. Although this is more difficult with digital systems, it is not impossible. Thus, there is a need to *identify* and *authenticate* a mobile terminal correctly. Control messages need to be checked for integrity to ensure that spoofed messages do not cause the network to behave abnormally, leading to widespread disruption of communications.

Implementation. The SIM cards have a microprocessor chip that can perform the computations required for security purposes. Figures 7.10 and 7.11 show the principles of operation for the authentication and ciphering in a typical cellular network. We have used GSM as an example, but similar procedures are adopted in other systems. A secret key K_i is stored on the SIM card and it is unique to the card. This key is used in two algorithms, A1 and A2, for authentication and confidentiality respectively. For authentication purposes, shown in Fig. 7.10, the secret key K_i is used in a challenge response protocol between the BSS and the MS. The secret key K_i is used to generate a privacy key K_c that is used to encrypt messages (voice or data) as the case may be using the A2 algorithm. The control channel

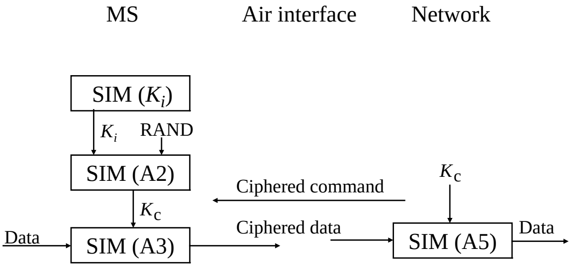


FIGURE 7.11 Basic principles of ciphering in a typical cellular network.

signals are encrypted using a third encryption algorithm. The size of a typical secret key K_i and the response to the challenge play an important role in the robustness of the security. Figure 7.11 illustrates the basic principles of ciphering in a typical cellular network. The challenge random string and the secret key K_i are used with algorithm A2 to generate a new ciphering key K_c which is used with a third algorithm, A3, to cipher the data. Another aspect of security in cellular networks is that the secret key information is not shared between systems. Instead, a triple consisting of the random number used in the challenge, the response to the challenge, and the data encryption key K_c is exchanged between the VLR and the HLR. The VLR verifies if the response generated by the MS is the same. The algorithms A2 and A3 are secret and not shared between different systems.

In 3G systems, message authentication (see Chapter 11) is used to ensure the integrity and authenticity of control messages.

7.4 PROTOCOL STACK IN CELLULAR NETWORKS

In the previous sections of this chapter we introduced the architectural elements and an overview of the mechanisms that allow this architecture to support mobile operation for a cellular network. In this section we will provide the description of how these elements and mechanisms are integrated with one another to implement a cellular network. Elements of a network communicate with each other through a protocol stack that is specified by a standardization committee. The standard specifies the interfaces among all the elements of the architecture that was discussed in Section 7.2. As an example, Fig. 7.12 shows the typical protocol architecture for communication between the main hardware elements and the associated interfaces.

The air interface, which specifies communication between the MS and BTS, is the most detailed and wireless-related interface. The interface between the BTS and BSC and the interface between BSC and MSC draw on existing wired protocols. In Fig. 7.12, for

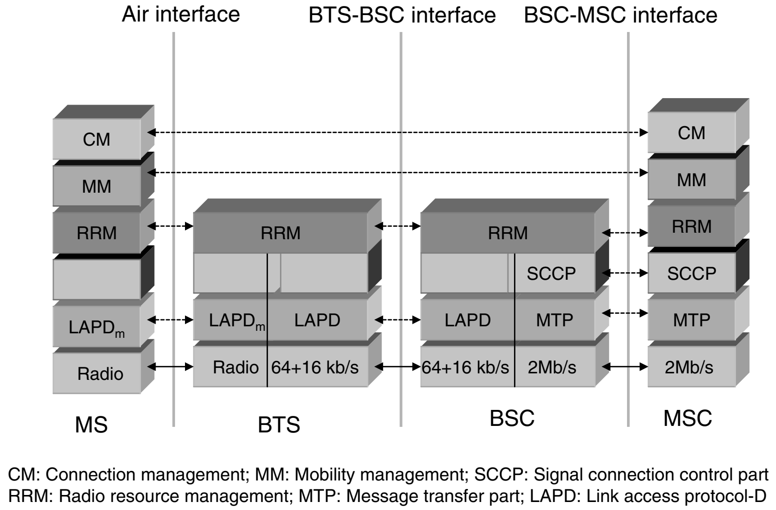

Air interface
BTS-BSC interface
BSC-MSC interface
CM
MM
RRM
$LAPD_m$
Radio
RRM
$LAPD_m$
LAPD
Radio
64+16 kb/s
RRM
SCCP
LAPD
MTP
64+16 kb/s
2Mb/s
CM
MM
RRM
SCCP
MTP
2Mb/s
MS
BTS
BSC
MSC
CM: Connection management; MM: Mobility management; SCCP: Signal connection control part
RRM: Radio resource management; MTP: Message transfer part; LAPD: Link access protocol-D

FIGURE 7.12 A typical protocol stack for a cellular network.

example, ISDN protocols are used for the physical wired connection. The protocol stack can be divided into three layers:

- layer 1 – physical layer
- layer 2 – DLL
- layer 3 – networking or messaging layer.

The support for the interface between the BTS and BSC is for voice traffic at 64 kb/s and data/signaling traffic at 16 kb/s. Both types of traffic are carried over link access protocol-D (LAPD), which is the data link protocol used in ISDN. The interface between different BSCs to the MSC takes place on a physical layer at 2 Mb/s and employs PSTN's SS7 protocols for communication. The message transport protocol (MTP) and the signaling connection control part (SCCP) of SS7 are used for error-free transport and logical connections respectively. The applications that employ the SS7 protocols deal with direct transfer of data and management information for radio resource handling and operation and maintenance information for the operation and maintenance communication messages.

In the following three sections we cover more details of the three layers with specific examples to provide the readers with an understanding of how a typical cellular network operates to support different services.

7.4.1 Layer 1: Physical Layer

The physical layer of the wired parts of the wireless network usually follow the traditional PSTN protocols (e.g. ISDN standard, with 64 kb/s digital data per voice user) or resort to IP services like those provided over the Internet. The physical layer defined in a cellular network that specifies the air-interface changes – which is either TDMA for GSM or CDMA for IS-95 or 3G digital cellular networks.

Communication between the terminal and the BS through the air interface involves both information traffic and signaling and control. The entire communication system can be thought as a distributed real-time computer that uses a number of instructions to transfer information packets from one location to another. As we described in Section 7.3, we have several major tasks to make such a system work. We need initial signaling for registration and call establishment, we need to maintain synchronization among the terminals, we need to manage mobility, and we need to transfer data traffic. In a manner similar to computers, we need a set of instructions and ports to instruct different elements of the network to perform their specified duties. In a cellular network these ports are referred to as logical channels, or simply as channels. In GSM, for example, logical channels use a physical TDMA slot or a portion of a physical slot to specify an operation in the network. In a CDMA network, such as cdmaOne, these channels use different spreading codes associated with different physical channels.

To implement a particular process in the cellular network (e.g. see the registration process shown in Fig. 7.14) we have a number of short messages, which are carried by different logical channels. To make it easier to define all these channels, we usually classify them into different categories. The principal categories of channels are *traffic channels* and *control channels*. Traffic channels carry voice and data traffic between the MS and BTS. We discuss the traffic channels on the air with examples of specific TDMA and CDMA technologies later. Control channels can be divided into three categories of channels: *broadcast channels*, *common control channels,* and *dedicated control channels*. The main duties of the broadcast channels are to prepare the physical layer to read other logical channels and establish a communication link. The information provided by these channels can be divided into two categories:

- *Synchronization and frequency control* provided by the BTS by broadcasting synchronization signals. An MS in the coverage area of a BTS uses these broadcasts to synchronize its carrier frequency, bit timing, frame synchronization, and signal strength to those specified by the network.
- *System identification* provided by BTS by broadcasting synchronization parameters, available services, and cell and network ID. Once the carrier, chip or bit, and frame synchronization between the BTS and MS are established, such information provides the MS the environment parameters associated with the BTS covering that area.

  The common control channels are also one-way channels used for call establishment and they can be divided into two categories:
- *Paging channel* used by the BTS to page the MS for incoming call.
- *Random access channel* used by the MS to access the BTS for registration and call establishment. The random access channel can be implemented on an ALOHA-type MAC protocol that allows the MS to contend for registration to receive a dedicated control or traffic channel.

The dedicated common control channels are two-way channels supporting signaling and control for individual users. Typical duties of these channels are to transfer network control information for call establishment and mobility management and to exchange necessary parameters between the BTS and the MS to maintain the link.

We leave the detailed description of air interface for TDMA and CDMA approaches for the following sections and continue our brief discussion on layer 2 and layer 3 to show how the physical layer packets are used to transfer a message at higher layers.

7.4.2 Layer 2: Data Link Layer

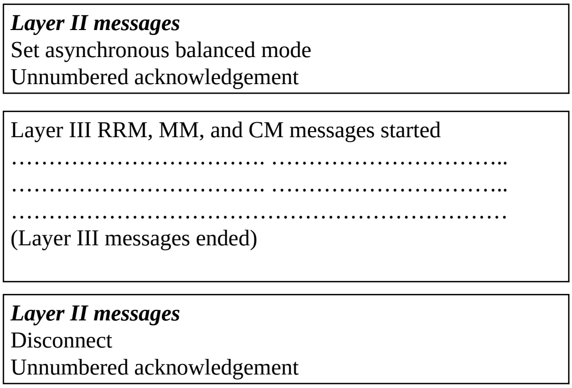
Cellular networks were originally designed for connection-based voice and data services. Any connection-based network can be considered to be two networks: one used for traffic and the other for signaling and control. The signaling and control may be through the same physical channels or through separate physical channels. In traffic channels, the information bits are encoded with strong error detection and correction codes to form the transmitted data stream, which is then sent over the TDMA or CDMA physical layers. Signaling and control data are conveyed through layer 2 and layer 3 messages. Similar to wired networks, the overall purpose of the DLL (layer 2) is to check the flow of packets for layer 3 and allow multiple service APs (SAPs) with one physical layer. The DLL checks the address and sequence number for layer 3 and manages acknowledgements for transmission of the packets. The DLL can be separated into two SAPs for signaling and SMS. This way, unlike other data services that are carried through traffic channels, the SMS is transmitted through a fake signaling packet that carries user information over signaling channels. The DLL provides this mechanism for multiplexing the SMS data into signaling streams.

In our example protocol stack shown in Fig. 7.12, the ISDN data link protocol LAPD is used for the DLL of the wired connections in the network connecting the BTS to BSS and BSS to MSC respectively. The DLL for the air interface is LAPD$_m$ where "m" refers to the modified version of LAPD adapted to the mobile environment. The length of the LAPD$_m$ packets is the same as LAPD, but the format is slightly adjusted to fit the mobile environment. One may eliminate the synchronization bits and CRC codes in LAPD because a cellular network has the time synchronization and strong coding at the physical layer. In peer-to-peer layer 2 communications, such as DLL acknowledgements, there is no layer 3 payload, but other DLL packets carry layer 3 messages. Typical "pure" or peer-to-peer layer 2 messages are: set asynchronous balanced mode, disconnect, unnumbered acknowledgement, receiver ready, receiver not ready, and reject. All these messages do not have layer 3 information bits and are referred to as layer 2 messages. The information bits in layer 2 packets specify layer 3 operations implemented on the logical signaling channels. These information bits are different for different operations.

Figure 7.13 shows a typical format of layer 2 and layer 3 messages in a procedure between two elements of the network. They start with simple pure layer 2 messages without layer 3 information bits to initiate a procedure. Then a number of layer 2 messages with layer 3 information follow to complete the necessary operation for the procedure. At the end, a couple of pure layer 2 messages disconnect the session between the two elements.

7.4.3 Layer 3: Networking Layer

As we discussed in Section 7.3, there are a number of mechanisms needed to establish, maintain, and terminate a mobile communication session. The networking or signaling layer implements the protocols needed to support these mechanisms. The networking layer is also responsible for control functions for supplementary services and SMS. The traffic channels are mapped into different frame formats and carried by physical-layer channel formats associated with different speech or data services. The signaling information uses other physical channels and more complicated DLL packaging. A signaling procedure or mechanism or protocol, such as the registration process shown in Fig. 7.5, is composed of a sequence of communication events or *messages* between hardware elements of the systems



Layer II messages
Set asynchronous balanced mode
Unnumbered acknowledgement

Layer III RRM, MM, and CM messages started
……………………………. …………………………..
……………………………. …………………………..
…………………………………………………………
(Layer III messages ended)

Layer II messages
Disconnect
Unnumbered acknowledgement

FIGURE 7.13 Typical format of the messages in a procedure used for implementation of a network operation mechanism.

that are implemented on the logical channels encapsulated in the DLL frames illustrated in Fig. 7.13. Layer 3 defines the details of implementation of messages on the logical channels encapsulated in DLL frames. As we mentioned before, among all messages communicated between two elements of the network, only a few, such as DLL acknowledgement, do not carry layer 3 information.

Information bits of the layer 2 packets, shown in Fig. 7.13, specify the operation of a layer 3 message. These bits are further divided into several fields to allow multiple procedures to operate in parallel, to identify the category of the operation, to identify the type of message, and to provide a time stamp for the instruction. The number of layer 3 messages is much larger than the number of pure layer 2 messages. To simplify the description of the layer 3 messages further, they are usually divided into three subcategories or sublayers: radio resource management (RRM), mobility management (MM), and connection management (CM) messages or layers shown in the upper part of Fig. 7.12.

The RRM sublayer of layer 3 manages the frequency of operation and the quality of the radio link. This sublayer does not have an equivalent in wired networks, because there is typically no frequency channel assignment issue in wired networks. The main responsibilities of RRM are to assign the radio channel and hop to new channels in implementation of the slow frequency hopping option, manage handoff procedures and measurement reports from MSs for handoff decisions, implement power control procedures, and adapt to timing advance for synchronization.

The MM sublayer handles mobility issues that are not directly related to the radio. Major responsibilities of this sublayer are location update, authentication procedures, TMSI handling, and attachment and detachment procedures for the IMSI. The CM sublayer establishes, maintains, and releases the circuit-switched connection and helps in SMS. Specific procedures for the CM sublayer are mobile-originated and -terminated call establishment, change of transmission mode during the call, control of dialing using dual-tones, and call re-establishment after MM interruption. Figure 7.14 shows the

Message name	Category
1. Channel request	RRM
2. Immediate assignment	RRM
3. Call establishment request	CM
4. Authentication request	MM
5. Authentication response	MM
6. Ciphering command	RRM
7. Ciphering ready	RRM
8. Send destination address	CM
9. Routing response	CM
10. Assign traffic channel	RRM
11. Traffic channel established	RRM
12. Available/busy signal	CM
13. Call accepted	CM
14. Connection established	CM
15. Information exchange	

FIGURE 7.14 Mobile-initiated call establishment procedure shown in Fig. 7.6 and breaking of different steps into three main categories of sublayer protocols.

15-step mobile-initiated call establishment procedure that was discussed earlier in Figure 7.6. The first column identifies the message and the second column identifies the sublayer of the layer 3 in which the message is implemented. Note that layer 3 only handles the signaling and control messages; the traffic messages do not have any sublayer.

Explanations of the details of coding of each message and a complete list of the messages for a cellular network are beyond the scope of this book. For a complete list of the messages used in layer 3 of the GSM, for example, the reader can refer to [Goo97] and for further detail of the operation to [Red95] or [Gar99].

7.5 PHYSICAL LAYER IN TDMA AIR INTERFACE

The physical layer in the air interface is the most complex part of a cellular network. This layer specifies how the information from different voice and data services are formatted into packets and sent through the radio channel. It specifies the radio modem details, structure of traffic and control packets in the air, and the packaging of a variety of services into the bits of a packet. This layer also specifies modulation and coding techniques, power control methodology, and time synchronization approaches to enable establishment and maintenance of the channels. This section provides the details of implementation of these functionalities in a cellular TDMA network using GSM as an example.

7.5.1 Modulation Technique

Since cellular networks cover longer distances and they were primarily designed for telephone application with a fixed rate and QoS, usually they use simpler modulation techniques than multirate wireless data networks operating in local areas. For example, GSM uses the GMSK modulation for implementation of the modem to carry layer 1 data packets. This modulation technique draws on the popular analog FM technique commonly used in radio broadcasting, for which a number of inexpensive chip sets exist in the market. Indeed, FM was the predominant form of analog modulation used in the mobile radio industry and the 1G cellular networks. In general, digital FM is referred to as frequency-shift keying (FSK), which forms a simple and popular method for wireless communications.

Figure 7.15*a* shows the basic concept behind binary FSK modulation. The binary baseband data stream is encoded into two different frequencies before transmission in the channel. To implement this modulation in its simplest form, as shown in Figure 7.15*b*, one can input the binary data stream directly to a traditional analog FM transmitter and use an analog FM receiver to demodulate the signal at the receiver. In its ideal form an analog FM transmitter linearly maps the instantaneous amplitude of the message signal to a constant-amplitude sinusoid with varying frequency at the output of the transmitter. A binary input signal takes only two levels of amplitude, so the output would be a constant-envelope signal with two frequencies (tones) associated with the two different levels of signal.

An important parameter in the design of FSK modems is the frequency spacing between the tones. This distance is representative of the occupied bandwidth of an FSK signal, and to maintain optimal detection at the receiver it should take specific values that ensure orthogonality of the transmitted symbols. For noncoherent detection (when the receiver is not locked to the phase of the transmitted carrier), FSK modems use a minimum distance between the tones of $1/T$, where T is the duration of the transmitted data symbols. For

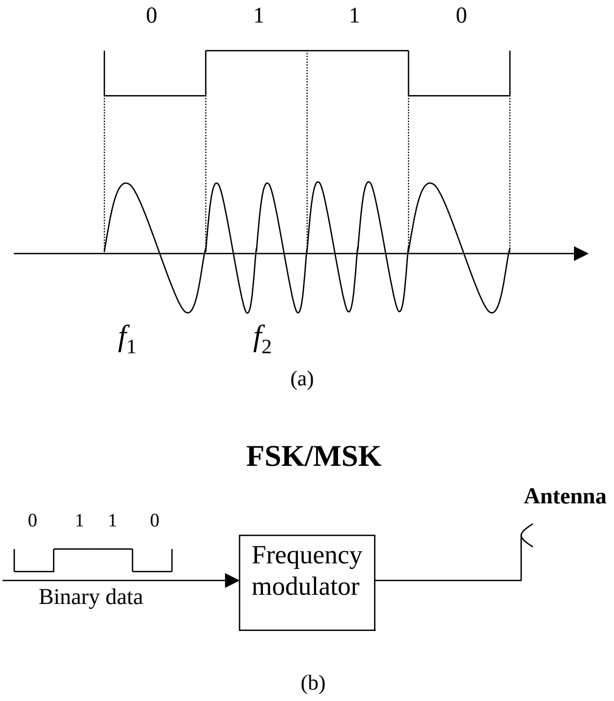


FIGURE 7.15 (*a*) Basic concept of FSK; (*b*) implementation of FSK using an FM transceiver.

coherent demodulation, the distance between the tones can be reduced to $1/2T$, which is the minimum acceptable distance between the tones that ensures the orthogonality of the transmitted symbols. The FSK modulation with minimal tone distance of $1/2T$ is referred to as *minimum shift keying* (MSK), which was a very popular transmission technique in radio communications when GSM was designed.

To make an MSK signal even more attractive for radio communications, as shown in Fig. 7.16, the transmitted data signal is filtered before FM to further reduce the side lobes causing interference with neighboring channels. The most popular filters used for this implementation are Gaussian filters, and the associated modulation technique is referred to as GMSK, which has been one of the most popular modulation techniques in 2G wireless networks and was adopted by the GSM standard. In the time domain, the Gaussian filter smoothes sharp transitions of the voltage levels. As a result, rather than immediate changes of the tone frequency at the output of the FM modulator, we will have a smooth transition from one tone frequency to another that reduces the side lobes of the transmitted FM modulated signal. Using GMSK, the GSM system manages a data rate of 270.833 kb/s in a 200 kHz band, giving a spectral efficiency of 1.35 bits/(s Hz) for the transmission technique

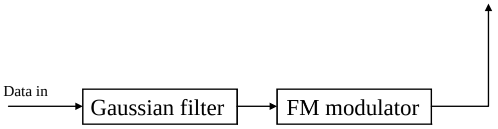


FIGURE 7.16 GMSK signals.

with more than 60 dB attenuation of the side lobes that interfere with adjacent channels. Over shorter distances for data applications, where variable data rates are acceptable to the user, one may increase the number of frequencies in the band to produce more symbols per second or more bits per symbol and, consequently, increase the data rate. This approach is taken for implementation of high data rates for TDMA networks, which is discussed later on in this chapter.

7.5.2 Power and Power Control

Power management is an important issue in wireless networks in general. Power management in cellular telephone networks helps the service provider to control the interference among the users and minimize the power consumption at the terminal. Therefore, power management has a direct impact on QoS and the life of the batteries, which are extremely important to the users.

There are two major classes of MSs: vehicle mounted and hand-held terminals. Mobile-mounted terminals use the car battery and handheld terminals use rechargeable batteries. The antenna for the mobile is mounted outside the car, which is away from the user's body, while the antenna in the handheld terminals is next to the ear and brain of the user, which raises health concerns for high-radiated powers. Cellular telephone networks have radii ranging from several hundred meters to several tens of kilometers. The size of the cells also plays a role in the required transmitted power for the BTS and the MS. To allow manufacturers and service providers to accommodate the diversified requirements for different MS and BSSs, the cellular standardization committees identify a number of radiated power classes.

For example, in the GSM standard there are five power classes for the mobile terminal from 29 dBm (0.8 W) up to 44 dBm (20 W) with a 4 dB separation between consecutive mobile classes. There are eight classes for the BTS power, ranging from 34 dBm (2.5 W) up to 55 dBm (320 W) in 3 dB steps. Transmitted RF power in the MS is always controlled to its minimum required value to minimize the co-channel interference among different cells and maximize the life of the battery. The MS is allowed to reduce its peak output power down to 20 mW in 2 dB steps. The BSS calculates the power level for individual MSs by monitoring the interference and received signal strength and sends this information through control signaling packets to the MS.

7.5.3 Physical Packet Bursts

Cellular telephone networks divide the available bandwidth into a number of carriers, each operating in a different center frequency. Usually, a simple modulation technique is used in each carrier to generate packet bursts to carry the traffic and control and signaling messages between the MS and a BTS. In a typical TDMA network, packet bursts have fixed lengths to simplify assignment of many different packet types in a time-divided stream. Cellular networks operating in wide areas use FDD to separate downlink or forward channels from the uplink or reverse channels, because the round-trip delay is relatively high and TDD needs unacceptably long time gaps between packet bursts. For example, GSM uses the 890–915 MHz for the uplink (reverse) and 935–960 MHz for the downlink (forward) channels. As shown in Fig. 7.17, the 25 MHz band for each direction is divided into 124 channels each occupying 200 kHz with a 100 kHz guard band at the two edges of the spectrum. Each carrier supports eight time slots for TDMA operation. The data rate of each

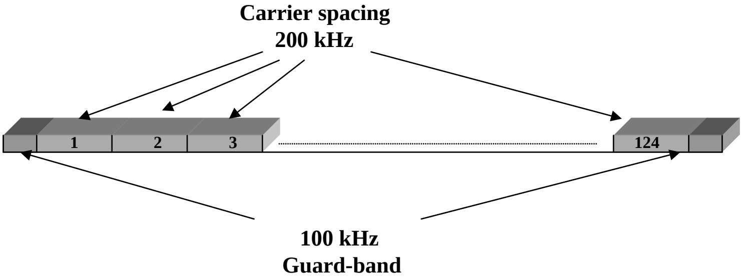


FIGURE 7.17 Division of spectrum into carriers for GSM cellular networks.

carrier is 270.833 kb/s, which is provided with a GMSK modem with a normalized bandwidth expansion factor of 0.3. With this data rate, the duration of each bit is 3.69 μs. The user transmission packet burst is fixed at 577 μs, which accommodates information bits and a time gap between the packets of duration equivalent to 156.25 times the bit duration of 3.69 μs.

GSM supports four types of bursts for traffic and control signaling. Figure 7.18 shows all four bursts types. The *normal burst* (NB), shown in Fig. 7.18*a*, consists of three tail bits (TBs) at the beginning and at the end of the packet, equivalent to 8.25 bits of gap period (GP), two sets of 58 encrypted bits (a total of 116 bits), and a 26-bit training sequence. The TBs are three 0 bits providing a gap time for the digital radio circuitry to cover the uncertainty period to ramp on and off for the radiated power and to initiate the convolutional decoding of the data. The 26-bit training sequence is used to train the adaptive equalizer at the receiver. Since the channel behavior is constantly changing during the transmission of the packet, the most effective place for the training of the equalizer is in the middle of the burst. The 116 encrypted data bits includes 114 bits of data and two flag bits at the end of each part of the data that indicate whether data is user traffic or for signaling and control.

TB (3)	Encrypted bits (58)	Training sequence (26)	Encrypted bits (58)	TB (3)	GP (8.25)

(a)

TB (3)	Fixed bit pattern (142)	TB (3)	GP (8.25)

(b)

TB (3)	Encrypted bits (39)	Synchronization sequence (64)	Encrypted bits (39)	TB (3)	GP (8.25)

(c)

TB (8)	Synchronization sequence (41)	Encrypted bits (36)	TB (3)	GP (68.25)

(d)

FIGURE 7.18 The four burst types in GSM: (*a*) normal burst; (*b*) frequency correction burst; (*c*) synchronization burst; (*d*) random access burst.

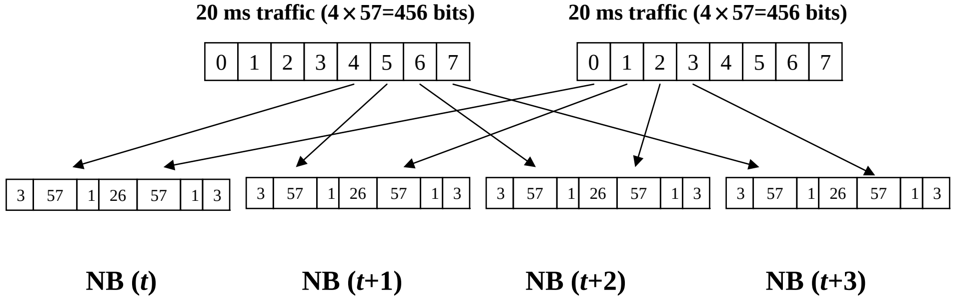


FIGURE 7.19 Interleaving traffic frames onto TDMA GSM frame in the air.

The user traffic data arrives in frames of length 456 bits; as shown in Fig. 7.19, they are interleaved into the transmitted NBs in blocks of 57 bits plus one flag bit. The purpose of interleaving is to improve the performance of users by distributing the effects of fade hits among several users. The 456 bits are produced every 20 ms. Therefore, the equivalent of 20 ms of arriving information is mapped into 456 bits. The standard specifies the method that maps the 20 ms of the traffic into the 456 bits.

Figure 7.20 shows how the 456-bit packets are formed from the speech signal in a GSM TDMA network. Each 20 ms of the coded speech at 13 kb/s forms a 260-bit packet. The first 50 most significant bits receives a 3-bit CRC code protection first and then they are added to the second group of 132 bits with lower importance and a 4-bit tail that is all zeros. The resulting $132 + 53 + 4 = 189$ bits are then encoded with a $^1/_2$ convolutional encoder that doubles number of bits to 378. The convolutional code provides for error correction capabilities. The 378 coded bits are added to the 78 least important speech-coded bits to form a 456-bit packet every 20 ms. The 456-bit packets are used to form transmission

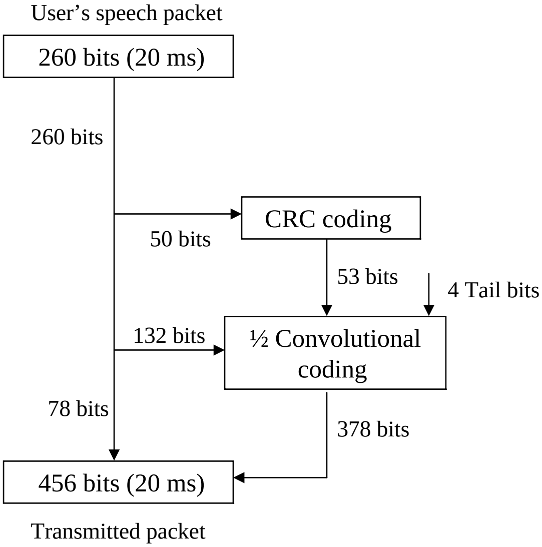


FIGURE 7.20 Coded speech packets in GSM.

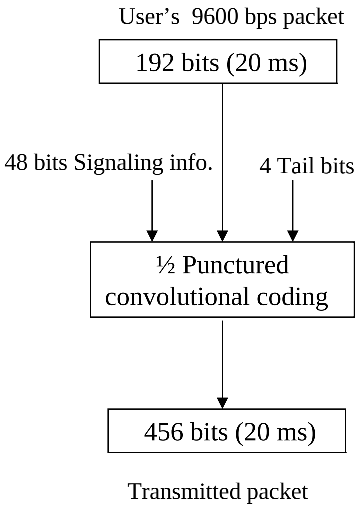


FIGURE 7.21 Coded 9600 bits/s data packets in GSM.

bursts, shown in Fig. 7.18. In this encoding scheme we have three classes of speech-coded bits; the first class of 50 bits receives both CRC error detection and the rate $^1/_2$ convolutional error-correcting coding protection. The second 132 bits receives only the convolutional encoding protection and the last 78 bits receives no protection. Therefore, the speech coder can protect the more important bits representing larger values of voltages by assigning them into different categories.

Figure 7.21 shows the formation of the 456-bit packets for 9600 bits/s data. The 192 bits of information are accompanied by 48 bits of signaling information and four TBs to form a 244-bit packet that is then expanded to 456 bits using a $^1/_2$ rate punctured convolutional encoder. Punctured coding can eliminate the need for doubling the number of transmitted bits by eliminating (puncturing) certain numbers of parity check bits [Pro00]. The resulting 456 bits are turned to NBs similar to the speech packets. The interesting point is that the 13 kb/s speech-coded signal and 9600 bits/s data modem both occupy the same transmission resources on the air interface. More channel coding bits are allocated to the data modem packets, which is expected to provide better error rate performance.

In addition to the traffic channels, we need a number of signaling or control channels, which are used to determine how the traffic packets should be routed in the network. Signaling channels using the NB as the channel over the air interface (shown in Fig. 7.18) use 184 signaling bits to convey the signaling message. These bits are first block coded with 40 additional parity check bits and four TBs to form a 228-bit block. The 228-bit block is then coded with a $^1/_2$ rate convolutional encoder to form a 456- bit packet occupying a 20 ms slot that is turned to a burst for transmission, as shown in Fig. 7.22.

The other three types of burst are simpler and designed for specific tasks. The simplest of all the remaining bursts is the *frequency-correction burst*, shown in Fig. 7.18*b*. It has three TBs at the start and the end of the burst. The rest of the packet contains all 0s, which allows simple transmission of the carrier frequency without any modulated information. An equivalent of 8.25 bits duration is used as the GP between this burst and others. This burst is used to implement the *frequency control channel* used by the BTS to broadcast

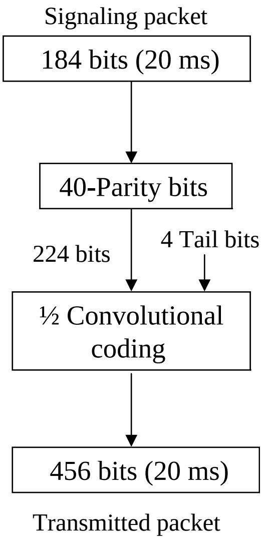


FIGURE 7.22 Coded signaling channel in GSM.

carrier synchronization signals. An MS in the coverage area of a BTS uses the broadcast frequency control channel to synchronize its carrier frequency and bit timing.

The *synchronization burst*, shown in Fig. 7.18*c*, is very similar to the NB except that the training sequence is longer and the coded data are used for the specific task of identifying the network. The BTS broadcasts the frequency-correction and synchronization bursts and the MSs use it for initial frequency and time-slot synchronization, as well as for training of the equalizer and initial learning of the network identity. This burst is used for implementation of *synchronization logical channel* used by the BTS to broadcast frame synchronization signals to all MSs. Using this channel, MSs will synchronize their counters to specify the location of arriving packets in the TDMA hierarchy.

The *random access burst* is used by the MS to access the BTS as it registers to the network. The overall structure is similar to NB, except that a longer start up and synchronization sequence is used to initiate the equalizer. Another major difference is the length of the much longer GP, which allows a rough calculation of the distance of the MS from the BTS. This calculation is possible from determining the arrival time of the random access burst. A GP of 68.25 bits translates to 252 μs. The signal transmitted from an MS should travel more than 75.5 km (at a speed of 300 000 km/s) before arriving at the BTS to exceed this GP. After calculation of the distance of the mobile when it first sends a packet to the BTS, the BTS calculates a time advance for the user so that packets arriving from different MSs are better aligned. In another words, the time jitter of arriving packets from different stations caused by differences among their distances reduces from 68.25 bits to 8.25 bits using the time advance calculated at the BTS.

Time-Division Multiple Access Frame Hierarchy. When a number of different slots carry the user traffic and a variety of control signals, a hierarchy is needed to identify the location of certain bursts among the large stream of bursts that are directed toward different

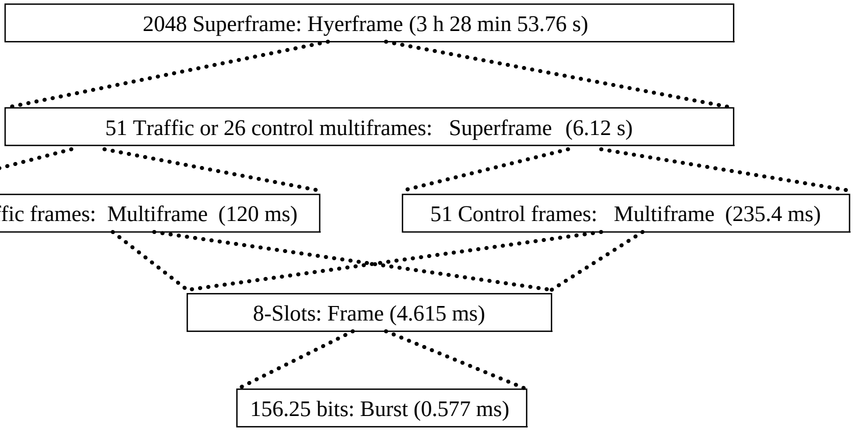

2048 Superframe: Hyerframe (3 h 28 min 53.76 s)
51 Traffic or 26 control multiframes: Superframe (6.12 s)
26 Traffic frames: Multiframe (120 ms)
51 Control frames: Multiframe (235.4 ms)
8-Slots: Frame (4.615 ms)
156.25 bits: Burst (0.577 ms)

FIGURE 7.23 GSM frame structure.

terminals. Each terminal needs a number of counters to track the related packets at different levels of the hierarchy.

For example, the GSM radio-interface standard provides a variety of traffic channels and control channels defined in a hierarchy built upon the basic eight-slot TDMA transmission format. The frame hierarchy, depicted in Fig. 7.23, shows the TDMA hierarchy of the GSM network from a burst of 0.577 ms interval to a hyperframe of length of around 3.5 h. The basic building block of the frame hierarchy is a 4.615 ms frame. Each frame comprises eight bursts or time slots. The time-slot interval is equivalent to the transmission time for about 156.25 bits, for which, as we saw in Fig. 10.16, durations equivalent to 8.25 (68.25 for random access burst) bit times are used as guard times during which no signal is transmitted. The next level in the hierarchy is a GSM multiframe, shown in Fig. 7.23. Each 120 ms multiframe is composed of 26 frames, each containing eight time slots. In each multiframe, 24 frames carry user information, while two frames carry system control information related to individual users. The data rate per voice user is calculated by considering that, for each 120 ms, 24 voice-bursts each carrying 2×57 114 bits of information are transmitted. Therefore, the data rate per user is $24 \times 114/0.120 = 22\,800$ bits/s. The speech coder has a data rate of 13 kb/s and the addition of error-detection and error-correction coding and additional bits for framing and the gaps brings the transmission rate up to 22.8 kb/s.

Figure 7.23 shows that the eight-slot frames may also be organized into control multiframes rather than traffic multiframes. Control multiframes are used to establish several types of signaling and control channels used for system access, call setup, synchronization, and other system control functions. Either traffic- or control-multiframes are grouped into superframes, which are in turn grouped into hyperframes. Counters at the terminals need to track the packet numbers at hyperframe, superframe, and mutiframe levels to communicate with the network. The counter for multiframes in the mobile terminal needs to keep track of the traffic channel for the terminal. Another counter needs to track the traffic superframe to identify where the location of the two control frames is. A variety of control signaling information embedded in the control superframe is extracted from the appropriate location using the counter for those frames. This complex method for tracking information related to each user can be considered as a weakness for TDMA and an incentive for the pan-European GSM standard to be replaced by CDMA technology in the 3G UMTS networks.

7.6 PHYSICAL LAYER IN CDMA AIR INTERFACE

The CDMA air interface became the popular choice for 3G cellular networks because it provided an inherent flexibility for multimedia traffic, provided a better quality voice, consumed less power (about 10% of analog or early TDMA phones), and does not require frequency planning since all cells employ the same frequency at the same time. The description of air interface in CDMA systems is more complex than TDMA air-interface because depending on the accuracy of synchronization among the terminals it may use different coding techniques to spread the channel and the CDMA technology is based on spread spectrum transmission which is more complex than traditional transmission techniques.

In a manner similar to TDMA networks, CDMA carriers occupy a portion of the overall bandwidth available for a cellular operator, but the *carrier bandwidth* is much wider in CDMA. For example, each carrier of cdmaOne or IS-95 occupies 1.25 MHz of the total available band, while carriers of GSM occupy 200 kHz. The carrier bandwidths of the 3G networks are even wider; for example, cdma2000 can potentially use multi-carrier CDMA with at least three times the bandwidth of cdmaOne, and the pan-European CDMA system UMTS occupies 5MHz of bandwidth. Within each carrier is the implementation of different logical channels. Different channels in CDMA are separated using *orthogonal* codes.

Implementation of the CDMA PHY is based on the application of a number of different coding techniques. The two dominant spread-spectrum codes are PN sequences and Walsh codes. The application of these codes for spreading the spectrum and increasing the reliability of the link on the forward and reverse channels is different. If the synchronization among the terminals is precise (for example, in the downlink or forward channel where the transmissions originate at a BS for all users and are perfectly synchronized), Walsh codes with zero cross-correlation when synchronized are used to eliminate the interference between user channels. When the synchronization among the terminals is not adequate (for example, in the uplink or reverse channel of IS-95 or cdmaOne 2G networks), the spreading that is employed uses Walsh codes differently along with PN sequences with good cross- and auto-correlation properties. In the following sections, we look at the CDMA implementation using Walsh codes and PN sequences using IS-95 as our example. In IS-95, the forward channels within a cell are separated by orthogonal codes and the reverse channels are separated using PN sequences. The variations implemented in 3G standards such as cdma2000 and UMTS will be discussed afterwards.

7.6.1 CDMA Forward Channels

In the forward channel, when the signals aimed at different mobile users are all sent from the same BS, traffic channels for different users and other control channels are separated by Walsh codes. To see the details of how channel separation is implemented, consider the details of IS-95. The IS-95 forward channel consists of four types of channel: pilot channel, synchronization channel, paging channel, and traffic channels. As shown in Fig. 7.24, each *carrier* contains a pilot channel, a synchronization channel, up to seven paging channels, and

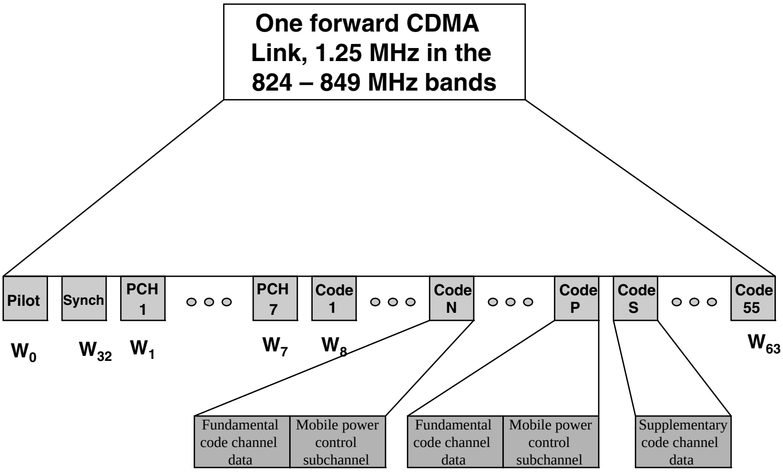

One forward CDMA Link, 1.25 MHz in the 824 – 849 MHz bands
Pilot
Synch
PCH 1
PCH 7
Code 1
Code N
Code P
Code S
Code 55
W0
W32
W1
W7
W8
W63
Fundamental code channel data
Mobile power control subchannel
Fundamental code channel data
Mobile power control subchannel
Supplementary code channel data

FIGURE 7.24 IS-95 forward channel.

a number of traffic channels. These channels are separated from one another using different Walsh spreading codes. The modulation scheme employed for transmission of spread signal in the forward channel is QPSK, which was described in Chapter 3.

The fundamental format of the spreading procedure for all channels is shown in Fig. 7.25*a*. Any information contained in the form of *symbols* (after coding, interleaving,

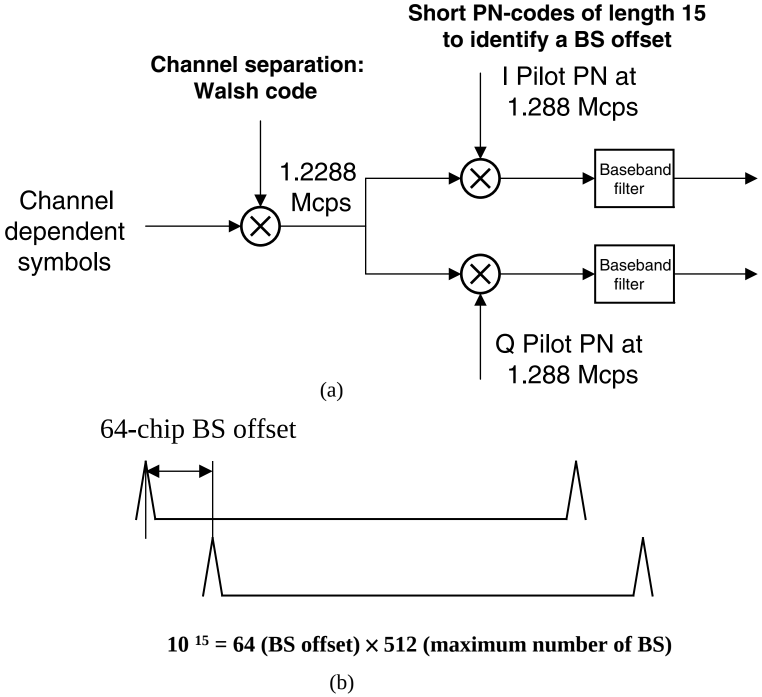

Short PN-codes of length 15 to identify a BS offset
Channel separation: Walsh code
I Pilot PN at 1.288 Mcps
1.2288 Mcps
Channel dependent symbols
Baseband filter
Baseband filter
Q Pilot PN at 1.288 Mcps
(a)
64-chip BS offset
10 15 = 64 (BS offset) × 512 (maximum number of BS)
(b)

FIGURE 7.25 (*a*) Basic spreading procedure on the forward channel in IS-95. (*b*) ACF of the PN sequence for different BTS.

etc.) is modulated by *Walsh codes* of length 64 that are obtained from *Hadamard matrices*, as discussed in Chapter 3. Each Walsh code identifies one of the 64 forward channels. After the channel symbols are spread using the orthogonal codes, they are further *scrambled* in the in-phase and quadrature phase lines by what are called *short PN spreading codes*. The PN spreading codes are not orthogonal, but possess excellent autocorrelation and cross-correlation properties to minimize interference among different channels. The PN spreading codes are M-sequences generated by LFSRs of length 15 with a period of 32 768 chips. The Walsh codes are used to isolate the transmissions between different channels *within* a cell and the PN spreading codes are used to separate the transmissions between different cells. In effect, the PN sequences are used to differentiate between several BSs in the area that are all employing the same frequency. The same PN sequence is used in all BSs; but, as shown in Fig. 7.25b, the PN sequence of each BS is *offset* from those of other BSs by some value. For this reason, BSs in IS-95 have to be synchronized on the downlink. Such synchronization is achieved using the GPS.

The rows of the Hadamard matrix form Walsh codes such that each row in the matrix is orthogonal to every other row. It is possible to generate a Hadamard matrix recursively, as described in Example 7.6.

Example 7.6: Recursive Generation of Hadamard Matrices and Walsh Codes The Hadamard matrix of order 2 is defined as

$$H_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

All other higher order Hadamard matrices can be obtained via the recursion

$$H_{2N} = \begin{bmatrix} H_N & H_N \\ H_N & \overline{H}_N \end{bmatrix}$$

Here, the matrix $\overline{H}_N$ is the matrix H_N with all 0s and 1s interchanged. Proceeding in this fashion, it is easy to generate H_{64}, which is employed in IS-95. Each row of the Hadamard matrix corresponds to a Walsh code. Consider

$$H_8 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

The first Walsh code from this matrix is $W_0 = [0\,0\,0\,0\,0\,0\,0\,0]$, i.e. the all-zero code. The last Walsh code is $W_7 = [0\,1\,1\,0\,1\,0\,0\,1]$. Note that all pairs of Walsh codes are orthogonal.

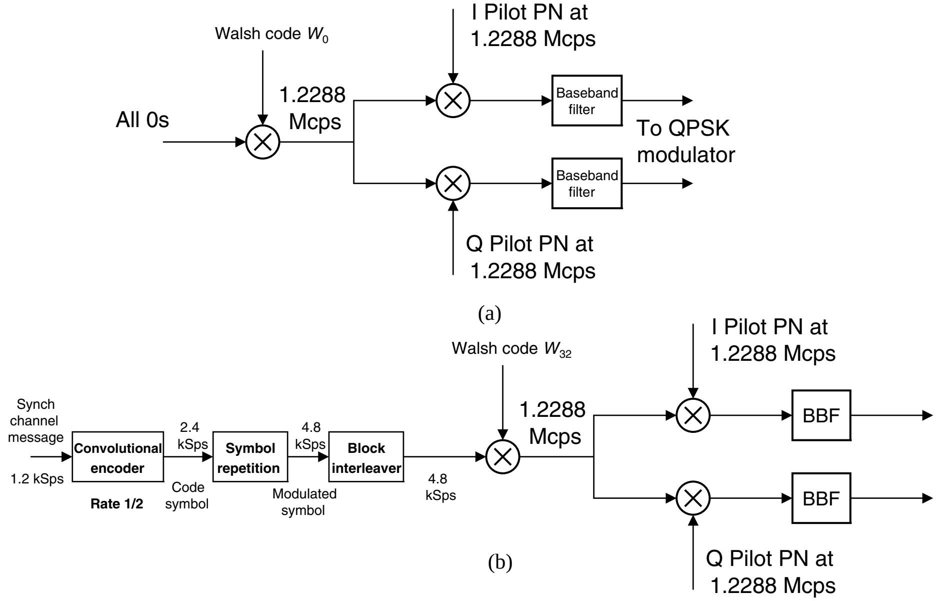


FIGURE 7.26 (*a*) Pilot and (*b*) sync channel processing in IS-95.

Various Walsh codes are used for spreading various logical channels in IS-95. The pilot channel employs the all-zero Walsh code W_0. The synchronization channel is assigned the Walsh code W_{32} and so on. The assignment of some Walsh codes is shown in Fig. 7.24.

The way the pilot channel is created is shown in Fig. 7.26*a*. The pilot channel is intended to provide a reference signal for all MSs within a cell that provides the phase reference for coherent demodulation. It is about 4–6 dB stronger than all other channels. The pilot channel is used to lock onto all the other logical channels. It is also used for signal strength comparison. It uses the all-zero Walsh code and contains no information except the RF carrier. It is also spread using the PN spreading code to identify the BS. The way to identify the BS is to *offset* the PN sequence by some number of chips. In IS-95, the PN sequences are used with offsets of 64 chips, which provide 512 possible spreading code offsets providing for unique BS identification in dense microcellular areas as well.

The sync channel is used to acquire initial time synchronization, and the way in which it is formed is shown in Fig. 7.26*b*. It uses the Walsh code W_{32} for spreading. Note that it uses the same PN spreading codes for scrambling as the pilot channel. The sync channel data operates at 1200 bits/s. After a rate $^1/_2$ convolutional encoding, the data rate is increased to 2400 bits/s, repeated to 4800 bits/s and then block interleaving is employed. The sync message includes the system and network identification, the offset of the PN short code, the state of the PN long code (see next section), and the paging channel data rate (4.8 or 9.6 kb/s).

The paging channel is used to page the MS when there is an incoming call, and to carry the control messages for call setup. Figure 7.27 shows how a paging channel message is created. It employs Walsh codes 1 to 7, so that there may be up to seven paging channels. There is no power control for the pilot, sync, and paging channels. The paging channel is additionally scrambled by the PN long code, as shown in Fig. 7.27. The long code is generated using a

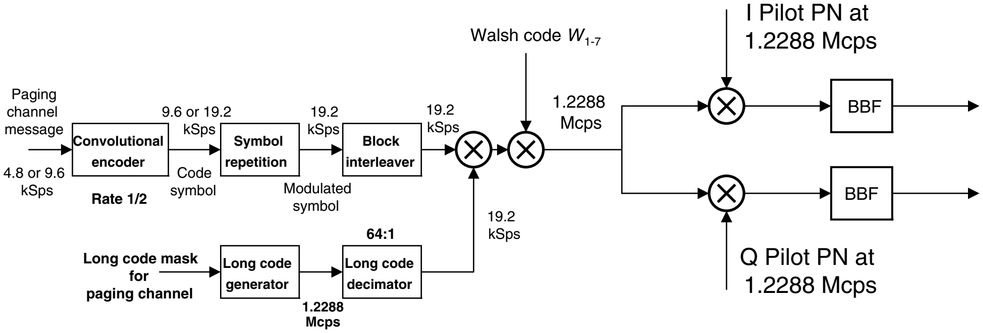


FIGURE 7.27 Paging channel processing in IS-95.

paging channel long-code mask of length 42. This means that the PN long code is generated by an LFSR of length 42 and has a period of 2^{42} chips.

The traffic channels carry the actual user information (i.e. digitally encoded voice or data). These data rates are typically 9.6, 4.8, 2.4 and 1.2 kb/s to support variable speech encoding or a variety of data applications. Walsh codes W_2–W_{31} and W_{33}–W_{63} can be used to spread the traffic channels, depending on how many paging channels are supported in the cell. As shown in Fig. 7.26, a rate ½ convolutional encoder is used that effectively doubles the data rate of the arriving traffic. Symbol repetition is used to increase the data rate to 19.2 kb/s. There is no repetition for 9.6 kb/s-encoded voice and there is a repetition of four times for voice at 2.4 kb/s. Regardless of the original rate of the data, the data rate of 19.2 kb/s is always delivered at the input of the block interleaver. The forward traffic channels are multiplexed with power control information for the reverse link, as shown in Fig. 7.28. Power control bits are multiplexed with the scrambled voice bits at 800 bits/s. Note that the traffic channels are scrambled with both the PN long code and the PN short codes to reduce interference among channels further.

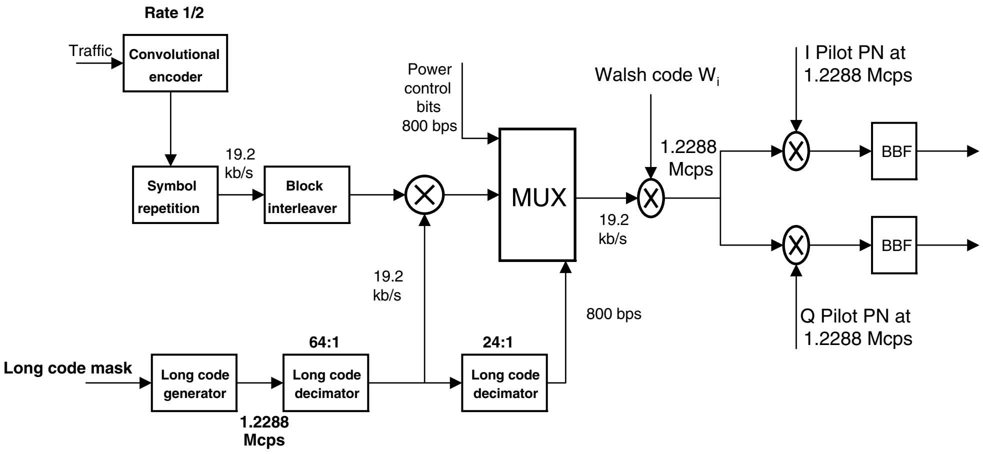


FIGURE 7.28 Forward traffic channel processing in IS-95.

7.6.2 CDMA Reverse Channels

The characteristics of the reverse channels in a CDMA system are different from those of the forward channel in many aspects. This arises for many reasons. Mobile users have smaller antennas, they are battery operated, they are used close to a human's head, demanding minimum transmission power, and the mobile terminals are not naturally synchronized.

The CDMA reverse channel is fundamentally different from the forward channel. It employs offset QPSK (OQPSK) rather than QPSK used in the forward channel. OQPSK is closer to a constant-envelope modulation. As described in Chapter 3, constant-envelope modulation techniques provide for a more power efficient implementation of the transmitter at the MS. In contrast, QPSK modulation is easier for demodulation again at the MS.

If the mobile users are synchronized, then orthogonal codes such as the Walsh codes are an excellent way for spreading the signal and separating different channels. However, unlike forward channels transmitted from the BTS, the timing of the user MSs are not aligned, and providing adequate synchronization among these terminals requires additional efforts in the design. If the designer of the CDMA network decides to avoid this extra effort, then the reverse channel can be implemented using PN sequences. This is the approach taken in IS-95, which is the example of our choice in this section. The overall structure of the reverse channels in IS-95 is shown in Fig. 7.29. Compared with the forward channel, there is no spreading of the data symbols using orthogonal Walsh codes. Instead, the orthogonal Walsh codes are used for *waveform encoding*, usually referred to as M-ary orthogonal coding or signaling. This means that the reverse link employs an orthogonal modulation scheme that consumes bandwidth but reduces the error rate performance of the system. To clarify this concept we resort to a simple example.

Example 7.7: M-ary Orthogonal Coding As a simple example of M-ary orthogonal coding, consider the example of the Hadamard matrix H_8 discussed in Example 7.6. There are eight orthogonal Walsh codes. We can perform a mapping between inputs of 3 bits to one

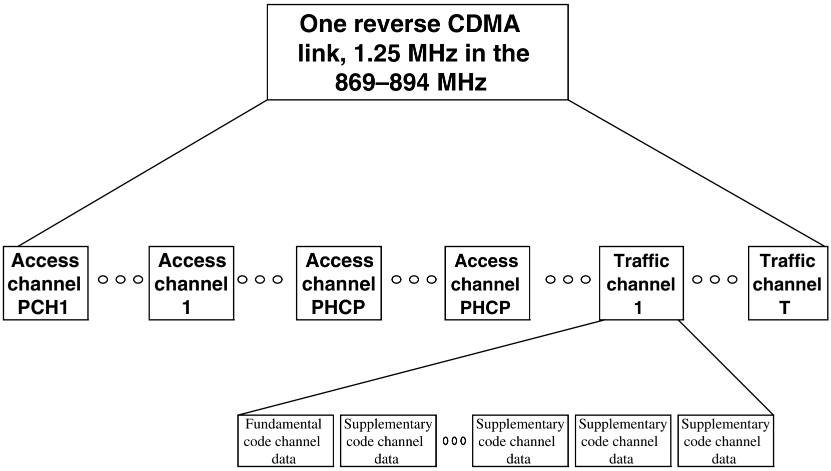

One reverse CDMA link, 1.25 MHz in the 869–894 MHz
Access channel PCH1
Access channel 1
Access channel PHCP
Access channel PHCP
Traffic channel 1
Traffic channel T
Fundamental code channel data
Supplementary code channel data
Supplementary code channel data
Supplementary code channel data
Supplementary code channel data

FIGURE 7.29 IS-95 reverse channel.

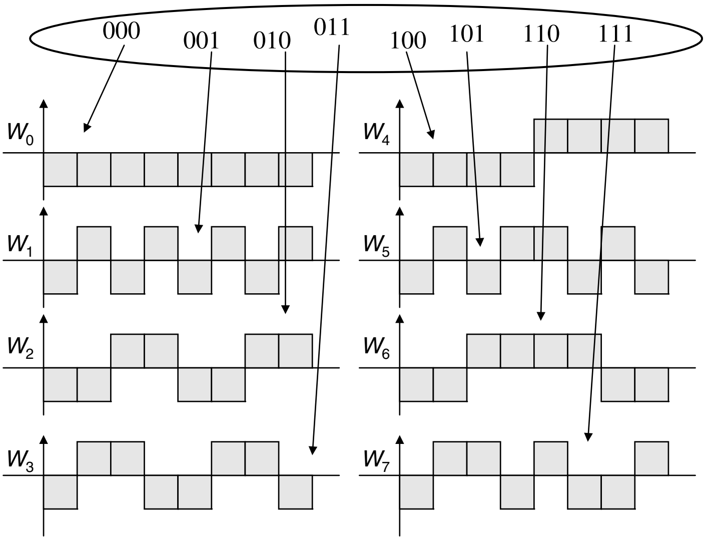


FIGURE 7.30 Mapping data bits to Walsh-encoded symbols.

of eight waveforms, as shown in Fig. 7.30. A different mapping scheme is employed in IS-95. Consider the Walsh codes of length 64. There are 64 such codes and they are orthogonal to one another. If these codes are used as waveforms to represent a group of information bits, then we can *encode* $\log_2 64 = 6$ bits using a Walsh code. For example, an input data stream 0 0 0 0 0 0 can be transmitted using the all-zero Walsh code W_0. This is a 64-ary modulation scheme where there are 64 symbols or alphabets for transmission and we use 6 bits to select one of them. Cross-correlation at the receiver is employed to detect the alphabets. In IS-95, the Walsh code that is used for encoding is determined by the equation

$$i = c_0 + 2_{c_1} + 4_{c_2} + 8_{c_3} + 16_{c_4} + 32_{c_5}$$

where c_5 is the most recent bit. For instance, if the input 6 bits are (1 1 1 0 1 0), then the Walsh code selected is $i = 1 + 2 \times 1 + 4 \times 1 + 8 \times 0 + 16 \times 1 + 32 \times 0 = 23$, i.e. W_{23} is transmitted.

The reverse channel of the IS-95 uses PN sequences to separate reverse channels and 64-ary orthogonal codes (see Example 7.7) to improve the reliability of transmission. There are basically two types of reverse channels in IS-95: the access channels and the reverse traffic channels. The MS transmits control information such as call origination, response to a page, etc. to the BTS via the access channels. Figure 7.31 shows the details of implementation of this channel in IS-95. The data rate over the access channels in IS-95 is fixed at 4800 bits/s. It is sent through a rate 1/3 convolutional encoder that increases the data rate to 14.4 kb/s. Symbol repetition is employed to increase the data rate to 28.8 kb/s. Every 6 bits is now mapped into 64 bits using the 64-ary orthogonal modulator. The long PN code is used to distinguish between different access channels. It spreads each of the bits at the output of the 64-ary orthogonal modulator by a factor of four, which yields a chip rate of 1.288 Mc/s.

Figure 7.32 shows the implementation of the traffic reverse channel in IS-95. The data burst after coding and interleaving, but just before the 64-ary orthogonal modulation, is at a

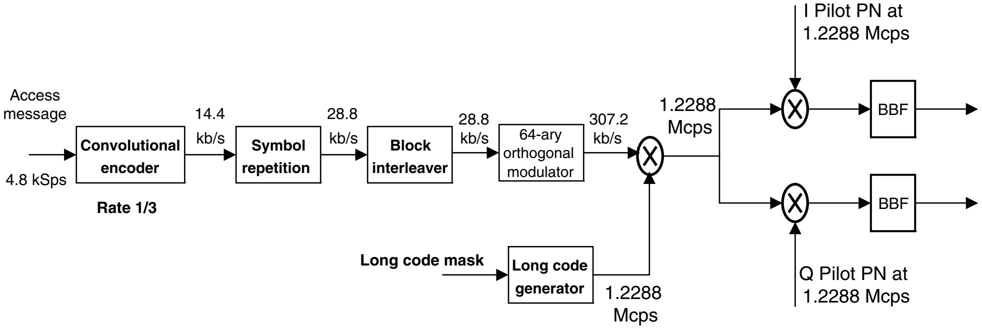


FIGURE 7.31 Access channel processing in IS-95.

rate of 28.8 kb/s. The output of the 64-ary orthogonal modulator is 28.8 × 64 /6 = 307.2 kb/s. After spreading by the long PN code by a factor of four, the final chip rate is 307.2 × 4 = 1.2288 Mc/s. A data randomizer is used in the fundamental code channel to mask out redundant data in case of symbol repetition. The reverse traffic channel sends information related to the signal strength of the pilot and frame error rate statistics to the BS. It is also used to transmit control information to the BS, such as a handoff completion message and a parameter response message.

7.6.3 Packet and Frame Formats in a Typical CDMA Network

We continue our discussion for packet formats using IS-95 as our example for CDMA networks. As discussed in the previous sections, the forward logical channels are of four types: the pilot, the sync, the paging, and the traffic channels. The reverse channels are either access channels or traffic channels. The forward traffic channel carries user data

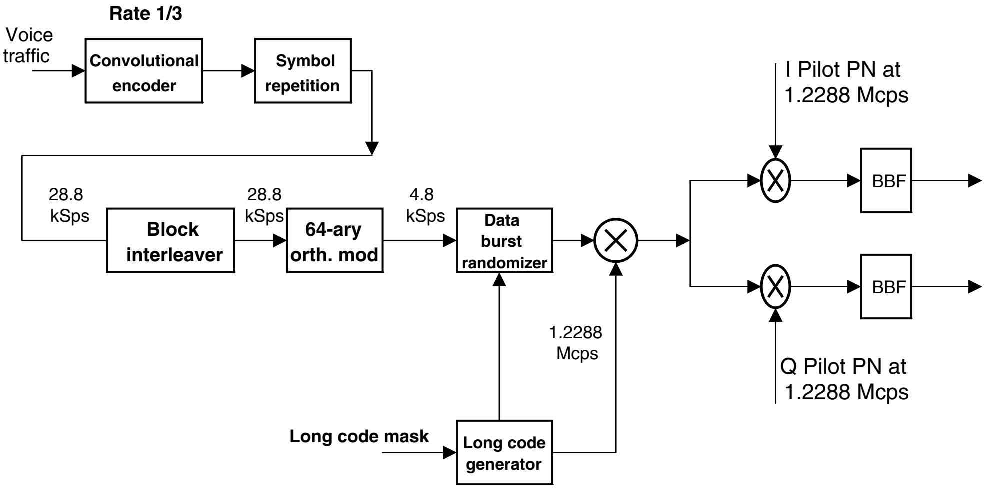


FIGURE 7.32 Reverse traffic channel processing for IS-95.

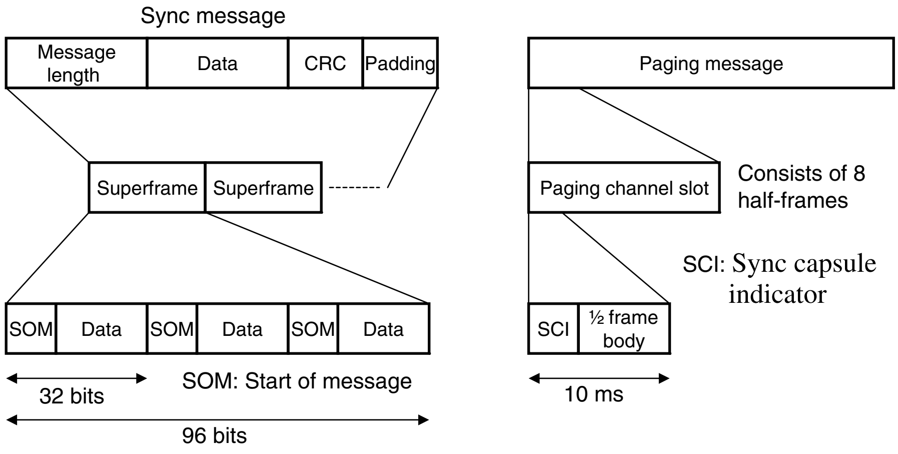
TABLE 7.1 Frame Contents for Forward Traffic Channels

Data rate (bits/s)	Information bits	CRC bits	Tail bits
9600	172	12	8
4800	80	8	8
2400	40	0	8
1200	16	0	8

(either data bits or encoded voice) at 9600, 4800, 2400 or 1200. The forward traffic channel frame length is 20 ms long. Table 7.1 shows the number of information bits, frame error control check bits, and TBs in each case.

The sync channel provides the MS information about the system network IDs, PN short sequence offset, the PN long code state, and the system time, among other things. Such *messages* can be long and are fragmented into *sync channel frames* of 32 bits, as shown in Fig. 7.33*a*. Three of the sync channel frames are combined into a sync channel superframe of 96 bits. The start of the message bit is a 1 for the first sync channel frame and 0 for subsequent ones that belong to the same message. The message itself (shown in the top part) consists of the message length, the data, an error checking code, and some padding. Padding with zeros is used to ensure that every new message starts in a new superframe.

The paging channel message, shown in Fig. 7.31*b*, announces a number of parameters to the MS, including the traffic channel information, the TMSI, response to access requests, and list of neighboring BSs and their parameters. Paging can be slotted or unslotted. In the former case, which enables the MS to save on battery power, the channel is divided into 80 ms slots. The paging channel message is similar in structure to the sync channel message (it has a message length, data, CRC, etc.). Since it is too long for



(a) Sync channel framing (b) Paging channel framing

FIGURE 7.33 Framing in IS-95 forward channels.

transmission in one slot, it is fragmented into 47 or 95 bits (data rate of 4800 or 9600 bits/s) and transmitted over a paging channel *half-frame* (10 ms long). The half-frame has one bit called the *synchronization capsule indicator* that has functionality similar to the start of the message bit. In this case, however, a message can start anywhere (not necessarily in a half-frame) and a zero value for the *synchronization capsule indicator* could indicate that one paging message ends and another starts within the same half-frame. Eight paging half-frames are combined into one paging slot of 80 ms.

Example 7.8: Number of Bits in the Paging Channel Half-Frame and Slot The number of bits depends upon the data rate. If the data rate is 9600 bits/s, then a 10 ms half-frame will carry 96 bits (one bit is the SCI) and 48 bits if the data rate is 4800 bits/s. Consequently, a paging slot, which has eight half-frames together, will contain $96 \times 8 = 768$ bits at 9600 bits/s and $48 \times 8 = 784$ bits at 4800 bits/s.

The access channel data rate is 4800 bits/s and each access channel message (very similar in structure to a sync message) is composed of several access channel frames lasting 20 ms. Thus, an access channel frame is 96 bits long. An *access channel preamble* always precedes an access channel message and it consists of several 96-bit frames with all bits in the frame equal to zero. The actual message itself is fragmented into 96-bit frames that have 88 bits of data and eight TBs set to zero.

The reverse traffic channel is once again broken into 20 ms traffic channel frames. The frame is further divided into 1.25 ms *power control groups*. There are thus 16 *power control groups* in one frame. A data burst randomizer randomly masks out individual *power control groups* depending on the data rate, which results in less interference on the reverse channel. For instance, at 4.8 kb/s (half the data rate), eight *power control groups* are masked. In addition to voice traffic, the traffic channel can also be used to transfer signaling or secondary data. In the *blank and burst* case, the entire frame carries data. In the *dim and burst* case, part of the frame carries voice and part of it data. The frame structures for the reverse traffic channel are very similar to that of the forward traffic channel.

7.6.4 Other Variations in CDMA Air Interface

So far in this section we have addressed technical aspects for implementation of forward and reverse channels in cdmaOne. Other variations of these basic concepts are used in different versions of CDMA networks used in the 3G cellular networks. Here we address these variations.

The primary standard for 3G systems is referred to as the International Mobile Telecommunications beyond the year 2000 (IMT-2000), the goal of which was to support higher data rates that can support multimedia applications, provide a high spectral efficiency, make as many of the interfaces standard as possible, and provide compatibility to services within the IMT-2000 [Zen00]. While voice traffic was expected to continue to be the main source of revenue, packet data for Internet access, advanced messaging services like multimedia e-mail, and real-time multimedia for applications such as telemedicine and remote security were envisaged in IMT-2000. The requirements for IMT-2000 include improved voice quality (wireline quality), data rates up to 384 kb/s everywhere and 2 Mb/s indoor, support for packet- and circuit-switched data services, seamless incorporation of existing 2G and satellite systems, seamless international roaming, and support for several simultaneous multimedia connections. Over 15 proposals were submitted and there were two major competing proposals the pan-European UMTS using wideband CDMA with FDD and TDD options and the CDMA2000 proposal that was backward compatible with IS-95. The main differences are summarized below [Zen00].

The primary requirements of 3G systems are that they should be able to support a variety of application data rates (from 384 kb/s circuit-switched connections to 2 Mb/s in indoor areas) and operation environments. This means that there must be support for quality of service and operation from megacells to picocells. The major modifications in the forward channel of the W-CDMA used in UMTS was that the BSs can operate in an asynchronous fashion that obviates the need of GPS availability to synchronize BSs. W-CDMA also employs what is known as an orthogonal variable spreading factor, allowing a variable spreading factor technique that maintains orthogonality between spreading codes of different lengths. CDMA2000 employs multiple carriers to provide a higher data rate than W-CDMA. It employs N carriers ($N = 1, 3, 6, 9$) for an overall chip rate of $N \times 1.2288$, which is 3.6864 Mc/s for $N = 3$. Alternatively, a single carrier can be employed to chip at the larger chipping rate. The former mode of operation is suitable for overlaying CDMA2000 over existing IS-95 systems. Walsh codes from 128 chips to 4 chips are employed to provide variable spreading and processing gains. All N carriers use the same single code for scrambling. The BSs still need to be synchronized and use the PN-code offsets for differentiation as before. Pilot channels are used for fast acquisition and handoff as before. In addition to the pilot, sync, and paging channels, auxiliary pilot channels can be used to supply beam-forming information if smart antennas are employed.

Support for variable data rates and operation in a variety of environments once again governs the implementation of the reverse link for 3G systems. In W-CDMA, more complex codes are used for scrambling on the uplink. These codes allow implementation of more signal processing techniques to improve the quality of the reverse channel by using a RAKE receiver in the BS and employing multiuser detection techniques to reduce the interference noise among different user channels. In CDMA2000, the reverse link is made more symmetrical with the forward link in many aspects. For instance, a *reverse pilot channel* is employed between each mobile and the BS for initial acquisition, time tracking, and power control measurement. More powerful codes are used, such as a rate $^1/_4$ convolutional code with a constraint length of 9. Variable-rate spreading is supported to enable better error correction capability and a variety of data rates. In W-CDMA, closed-loop power control is implemented in a manner similar to IS-95 with the power control bits transmitted 1500 times a second rather than 800 times in IS-95. This allows a very fast control of power and provides significant capacity gains in W-CDMA, especially at pedestrian speeds. Both inner and outer loop power control mechanisms are employed.

7.7 ACHIEVING HIGHER DATA RATES IN CELLULAR NETWORKS

In the mid 1990s, TDMA and CDMA technologies were the dominant 2G cellular networks when the 3G cellar network specifications for IMT-2000 demanded the data rates of 384 kb/s for everywhere coverage and a data rate of 2 Mb/s for local areas. Since each voice user produces a coded speech at around 10 kb/s, it is obvious that a high data-rate user needed to combine the time slots of several TDMA voice users or use many codes of a CDMA channel. As we described earlier in this chapter, each carrier of the dominant TDMA system, GSM, carries a data stream of 270 kb/s and the chip rate of the popular CDMA system, IS-95, is

1.25 Mb/s. Therefore, even if we allocate the entire carrier to one user we cannot satisfy the IMT-2000 basic requirement with the existing physical layers. The only solution to stay with the same carrier spacing would be to use multisymbol and/or multicarrier modulation that was proposed in CDMA2000. Otherwise, one would have needed wider bandwidth, as proposed in the wideband CDMA of the pan-European UMTS standard. In the meanwhile, since the main objective in the new trends for wireless cellular networks was to support higher data rates, a number of high data-rate systems emerged as an evolution of the existing legacy cellular networks. The most dominant ones in TDMA networks were GPRS and enhanced data for global evolution (EDGE) packet-switching networks and the dominant CDMA systems were 1xEV (single-carrier evolution) and 3xEV (three-carrier evolution) systems that are also referred to as HDR (High Data Rate) systems.

The principles of operation of these systems are important from three points of consideration: (1) how to modify the architecture to connect to Internet for a network which was originally designed to operate with connection-based PSTN; (2) how one can achieve high data rates using a voice-oriented network designed for lower speeds per user; (3) how one can assess the performance of such data networks for a mobile user. Architectural approaches combined with changes to the modem technology also form the basis for high-speed packet access (HSPA), which is being currently deployed in 3G systems, and long-term evolution (LTE), which is being considered for cellular systems beyond 3G. We discuss these three issues in the rest of this section using GPRS, EDGE, and HDR as our specific examples.

7.7.1 Changes in Reference Architecture to Connect to Internet

Figure 7.2 illustrates the general architecture of a cellular network to provide circuit-switched services using PSTN through a cellular telephone. To provide high data-rate packet-switched services, this architecture needs to be modified to support Internet connection. To implement this modification to the network we need a hardware device which steals the packet-switched data from the BSC and a mobile router which connects to other mobility-aware routers and to the Internet.

As we mentioned before, the new packet-switched network architecture is built on the existing circuit-switched architecture shown in Fig. 7.2. Figure 7.34 shows the general architecture of the modified Fig. 7.2 to support high data-rate packet switching using the Internet. There are a few new network entities needed, we can call them *support nodes*, which are now responsible for delivery and routing of data packets between the MS and external packet network. There are two types of support node: the *serving support node* and the *gateway support node*. There is also a new database for registration that is collocated with the HLR. It stores routing information and maps the cellular addresses to an IP address. Figure 7.34 shows this reference architecture. In addition to the two hardware and new database elements of the architecture, the standardization committee needed to define a number of new interfaces, identified by crossing dashed lines in Figure 7.34.

Before accessing the packet-switched data services, the MS must register with the packet-switched network to be identified with its new Internet addressing. The MS performs an attachment procedure with a service support node that includes authentication (by checking with the registration – “Reg” database). The gateway support node allocates the MS a temporary logical link identity and a packet data protocol carries the data between the MS and the gateway. The packet data protocol context is a set of parameters created for each session and contains the IP version, the address assigned to the MS, the requested QoS

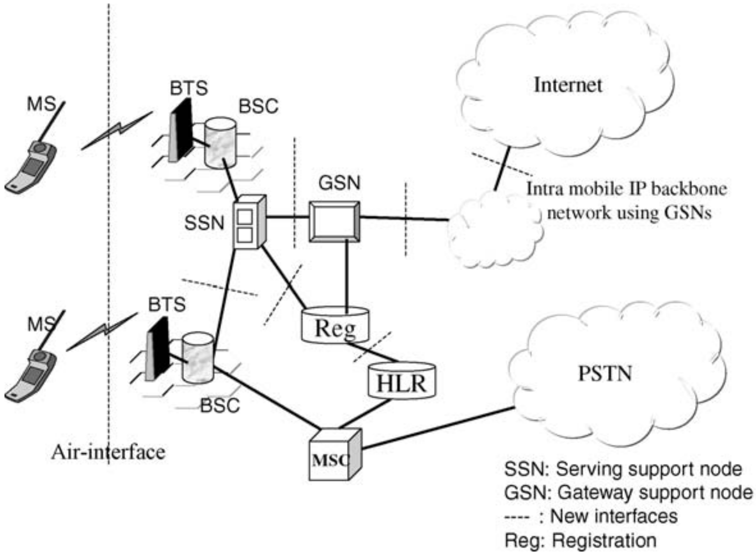


FIGURE 7.34 The system architecture to support packet switching.

parameters, and the gateway address that serves the point of access to the MS. The context is stored in the MS and the two new hardware elements of the mobile packet-switched network. A user may have several contexts enabled at a time. The address may be statically or dynamically assigned. The context is used to route packets accordingly through the Internet.

7.7.2 How to Achieve High Data Rates

Technical approaches to achieve higher data rates in CDMA and TDMA networks are quite different. Here, we provide examples of HDR and GPRS/EDGE to show the practical approaches to achieve higher data rates in CDMA and TDMA networks respectively.

Example 7.9: Implementation of HDR The data rate of the traditional IS-95 CDMA network from the BS to the MSs is 9600 bits/s, the spreading factor is 128, and the chip rate is 9600 bits/s × 128 chips per bit = 1.2288 Mc/s. This system supports 64 voice users that are separated by 64 orthogonal Walsh codes of length 64. To support HDR systems, let us suppose that all 64 Walsh codes are used for a single data user. Figure 7.35 illustrates the general concept of the IS-95 CDMA and the single HDR evolution (1xEV) that uses all 64 CDMA codes for one user and allows all users access through a TDMA scheme (different users transmit in different sets of time slots on a carrier within a cell). If BPSK were used for all channels, then the maximum achievable data rates would be 64 × 9600 bits/s = 614.4 kb/s. To further increase the data transmission rate, 1xEV uses multisymbol modulation to replace the existing BPSK modulation. The highest data rate in this system is achieved by a 16-QAM modulation with 4-bits per symbol, resulting in 614.4 kb/s × 4 = 2.4576 Mb/s [Ben00, Qua01]. As shown in Table 7.2, the 1xEV HDR system supports 12 data rates for packet data transmission. Depending on the received SNR from the MS, the BS adjusts the number of symbols of the modulation technique,

TABLE 7.2 Specification of Different Data Rates in the 1xEV HDR system [Qua01]

Physical layer parameters												
Data rate (kb/s)	38.4	76.8	153.6	307.2	307.2	614.4	614.4	921.6	1228.8	1228.8	1843.2	2457.6
Bits per encoder packet	1024	1024	1024	1024	2048	1024	2048	3072	2048	4096	3072	4096
Code rate	1/5	1/5	1/5	1/5	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
Encoder packet duration (ms)	26.67	13.33	6.67	3.33	6.67	1.67	3.33	3.33	1.67	3.33	1.67	1.67
Number of slots	16	8	4	2	4	1	2	2	1	2	1	1

number of bits per packet, and the number of times the packet is repeated (which is equivalent to changing the effective processing gain). The minimum data transmission rate is 38.4 kb/s, which is 64 times lower than the highest rate and is achieved by sending shorter packets four times in a 16 times longer time interval. To support this data rate a QPSK modulation is used. When the mobile is in the close vicinity of the BS, the highest data rate with 16-QAM and minimal coding is used. As the mobile moves away from the BS, the number of symbols in the signal constellation is reduced and the coding for error protection is increased, which results in a lower data rate.

As we described in Example 7.9, the MAC with HDR is really a TDMA/CDMA. As shown in Fig. 7.35, all 64 codes usually assigned to separate voice users or control channels are allocated to one HDR data user. Note that power control on the forward link in IS-95 results in some unused power margin (which is beneficial in the case of voice in reducing interference and improving voice quality). No such margin is left over for data transmissions in the case of HDR. On top of that, BPSK modulation used in the forward (downstream) channel is replaced with an optional QPSK modulation, 8-PSK modulation and 16-QAM to support the variety of data rates shown in the Table 7.2.

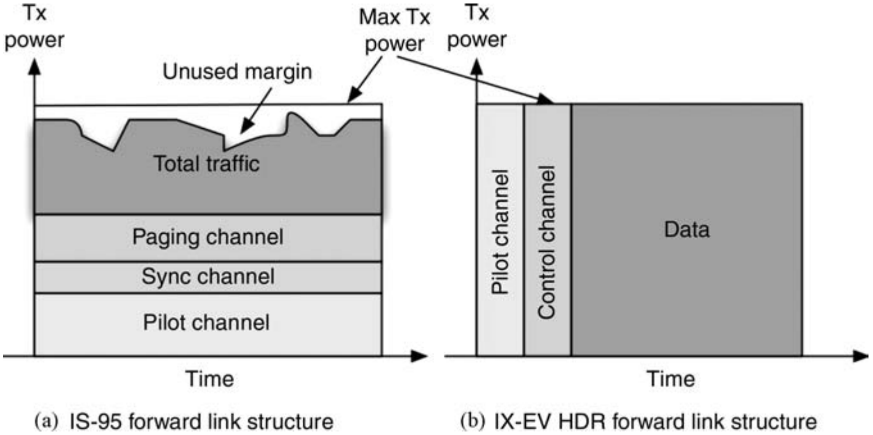


FIGURE 7.35 Distribution of transmitted power in IS-95 CDMA and 1xEV HDR forward channels using TDMA for data communications [Qua01].

TABLE 7.3 GPRS Data Rates for Channel-Coding Schemes 1–4

	Slot combination data rate (kb/s)		
Channel coding scheme	1 slot	4 slots	8 slots
CS1	9.2	36.8	73.6
CS2	13.55	54.2	108.4
CS3	15.75	63	126
CS4	21.55	86.2	172.4

GPRS and EDGE data services are based on the GSM infrastructure and TDMA air interface. In a manner similar to HDR, GPRS and EDGE use multiple voice user resources and multisymbol modulation schemes to achieve higher data rates. To be consistent with Example 7.9, we describe details of the GPRS and EDGE in the following example.

Example 7.10: High Data Rates in EDGE and GPRS GPRS uses GMSK, the same modulation technique used in GSM and creates different data rates by assigning different numbers of time-slots per user and changing the error correcting code used for error control in formation of the packets for packet data transmission. GSM has eight slots per carrier; therefore, GPRS (or EDGE) can use different number of slots to support higher data rates. A single 200 kHz carrier in GSM has eight time slots, each capable of carrying data at 9.2, 13.55, 15.75, or 21.55 kb/s (if FEC is completely omitted). The raw data rate for GPRS can thus be as high as $8 \times 21.55 = 172.4$ kb/s. Table 7.3 shows three sets of four data rates for one-slot, four-slot and eight-slot GPRS systems. The differences in the four categories of the data rates are created by using coding rates of $^1/_2$, $^2/_3$, $^3/_4$, and 1 for the formation of each packet. More details of the precise coding schemes are not our main concern here, but they are available from Cai and Goodman [Cai97]. We observe data rates ranging from 9.2 up to 172.4 kb/s supported by this system. As the distance of the mobile from the BS increases, the received signal strength reduces and the system falls to lower data rates. The main difference between EDGE and GPRS is that the EDGE system uses an optional 8-PSK modulation with 3 bits per symbol that can increase the maximum data rate another threefold to 473.6 kb/s. Table 7.4 shows four of the nine data rates supported by the EDGE system for one, four, and eight slots. The details of the coding for the GPRS and EDGE systems are slightly different;

TABLE 7.4 EDGE Data Rates Using $3\pi/8$ 8-PSK Modulation versus Different Channel-Coding Schemes and Slot Combinations

		Slot combination data rate (kb/s)		
Channel coding scheme	Modulation	1 slot	4 slots	8 slots
MCS1	GMSK	8.8	35.2	70.4
MCS4	GMSK	17.6	70.4	140.8
MCS5	8PSK	22.4	89.6	179.2
MCS9	8PSK	59.2	236.8	473.6

therefore, the data rates with the same modulation and the same numbers of slots do not match exactly. The details of different data rates and performance of EDGE is available in [Yal02].

7.8 DEPLOYMENT OF CELLULAR NETWORKS

Another important factor in deployment of wireless cellular networks is provision for expansion. The main investment of a wireless service provider is towards the cost of the fixed infrastructure, which includes the BS and connections between them. When a service provider starts an operation, they need to minimize the cost of infrastructure while continuously increasing the number of subscribers. As the number of subscribers increases, new income is generated and the service provider can afford to expand the network by increasing the complexity of its infrastructure to support a further larger population of subscribers. Therefore, there is a need for a plan to take into account the growth of the subscriber base and, thus, the entire wireless network.

In summary, we need to address to the following technical issues for planning a cellular network:

- selection of a frequency reuse pattern for different radio transmission techniques;
- physical deployment and radio coverage modeling;
- plans to account for the growth of the network.

7.8.1 Cell Fundamentals and Frequency Reuse

We looked at the cellular topology and the concept of employing a cellular architecture to increase the communications capacity and to cater to a large subscriber demand in hotspots in Section 7.1. We now consider quantitative means to characterize *interference* in a cellular topology. This, in turn, leads to quantitative means for determining the best cluster size and simple techniques for allocating the sub-bands of spectrum and within a cluster.

Even though in practice cells are of arbitrary shape (close to a circle) because of the randomness inherent in radio propagation, it is easier to obtain insight and understanding for system design by visualizing all cells as having the same shape. Also, it is easier mathematically to analyze a cellular topology by assuming a uniform cell size for all cells. Once some insight is obtained as to what the effects of interference are, then measurements, simulation, and a combination of these can be employed in actually determining the planning of a network.

For cells of the same shape to form a tessellation so that there are no ambiguous areas that belong to multiple cells or to no cell, the cell shape can be of only three types of regular polygon: equilateral triangle, square, or regular hexagon, as shown in Fig. 7.36.

A hexagonal cell is the closest approximation to a circle of these three and has been used traditionally for system design (see Fig. 7.37). The argument for a hexagonal shape comes from the fact that, among the three shapes mentioned, for a given radius (largest possible distance between the polygon center and its edge), the hexagon has the largest area.

In most of the literature and in back-of-the-envelope design, the hexagonal cell shape is chosen as the default cell shape. In particular cases that consider continuous distributions of

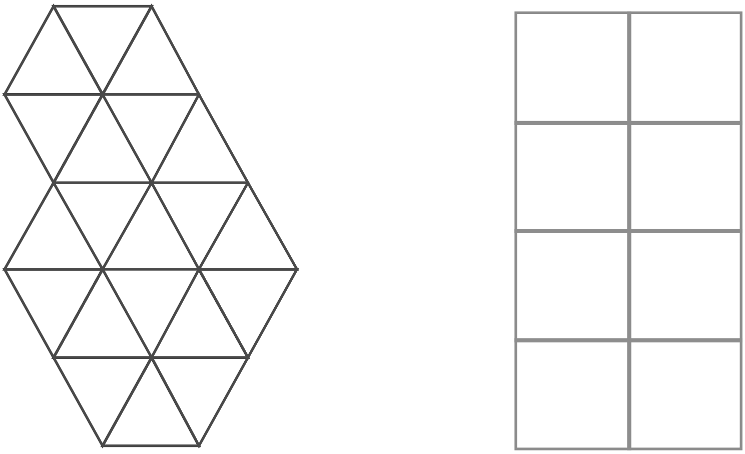


FIGURE 7.36 Triangular and rectangular cells.

traffic load and interference between different transmission schemes, a circular cell shape is employed for tractable mathematical calculations.

In order to investigate the effects of interference, which changes with distance, there is a need to come up with an elegant way of determining distances and identifying cells. Fortunately, it is possible to do this easily in the case of hexagonal cells [Mac79]. In order to

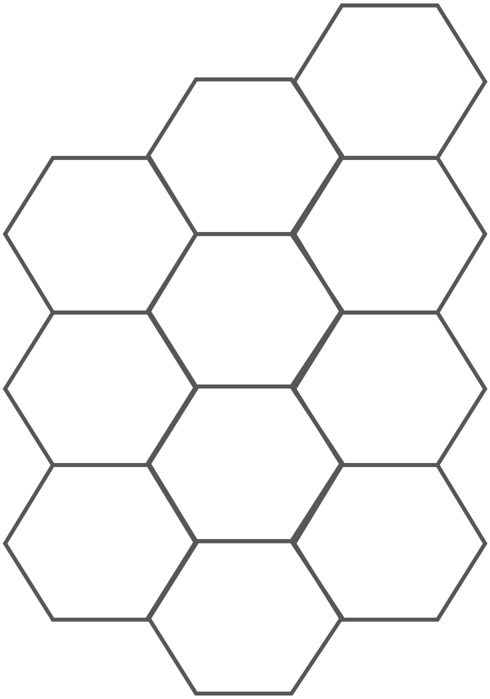


FIGURE 7.37 Arranging regular hexagons that can cover a given area without creating ambiguous regions.

maximize the capacity, co-channel cells must be placed as far apart as possible for a given cluster size. It can be shown that there are *only* six co-channel cells for a given reference cell at this distance. The distance between the co-channel cells can be shown to be $D_{\mathrm{L}} = \sqrt{3}R_{\mathrm{L}}$. Here, R_{L} is the radius of a cell. The relationship between the distance between co-channel cells, the cluster size, and the cell radius is given by

$$\frac{D_{\mathrm{L}}}{R_{\mathrm{L}}} = \sqrt{3N_{\mathrm{f}}} \tag{7.2}$$

This quantity is also referred to as the *co-channel reuse ratio*. Values for N_{f} can only take on values of the form $i^2 + ij + j^2$, where i and j are integers.

Example 7.11: Cluster Size of $N_{\mathrm{f}} = 7$ As described above, i and j can only take integer values. If we take $i = 2$ and $j = 1$, then we see that $N_f = 4 + 2 + 1 = 7$. Selecting a cell A, we can determine its co-channel cell by moving two units along one face of the hexagon and one unit in a direction 60° or 120° to this direction. Proceeding in this fashion, clusters of size $N_{\mathrm{f}} = 7$ can be created as shown in Fig. 7.38. A value of $N_{\mathrm{f}} = 7$ is employed in the USA in the AMPS.

The number of cells in a cluster N_{f} determines the amount of co-channel interference and also the number of frequency channels available per cell. Suppose there are N_{c} channels available for the entire system. Each cluster uses all the N_{c} channels. With fixed channel allocation, each cell is allocated $N_{\mathrm{c}}/N_{\mathrm{f}}$ channels. It is desirable to maximize the number of channels allocated to a cell. This means that N_{f} should be made as small as possible. However, reducing N_{f} increases the signal-to-interference ratio (as discussed in the following section). There is thus a tradeoff between the system capacity and performance.

Signal-to-Interference Ratio Calculations. In Section 7.1, we mentioned that a cellular architecture was essential in order to reuse the available spectrum while reducing interference caused by reusing the frequency spectrum. In this section, we will look in detail at the performance measures that are useful in system design, in particular the signal-to-interference ratio and its relationship with the path loss, and the grade of service.

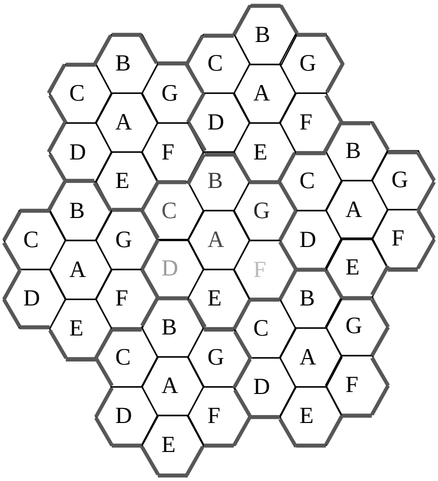


FIGURE 7.38 Hexagonal cellular architecture with a cluster size of $N_{\mathrm{f}} = 7$.

In general, the signal-to-interference ratio can be written as the ratio of the signal strength P_{desired} in the desired signal to the sum of the strengths of the signals from other users as follows:

$$S_\mathrm{r} = \frac{P_{\mathrm{desired}}}{\sum_{\mathrm{i}} P_{\mathrm{int},i}} \tag{7.3}$$

where $P_{\mathrm{int},i}$ corresponds to the strength of the ith interfering signal. In practice, the signal strength falls as some power of the distance α called the power–distance gradient or path-loss gradient (see Chapter 2). That is, if the transmitted power is P_t, then after a distance d, in meters, the signal strength of a radio signal will be proportional to $P_\mathrm{t} d^{-\alpha}$. In its most simple case, the signal strength falls as the square of the distance in free space ($\alpha = 2$). Suppose there are two BS transmitters, BS_1 and BS_2, located in an area with the same transmit power P_t and a mobile terminal is at a distance of d_1 from the first and d_2 from the second. If the mobile terminal is trying to communicate with the first BS, then the signal from the second BS is interference. The *signal-to-interference* ratio for this mobile terminal will be

$$S_\mathrm{r} = \frac{KP_\mathrm{t} d_1^{-\alpha}}{KP_\mathrm{t} d_2^{-\alpha}} = \left(\frac{d_2}{d_1}\right)^{\alpha} \tag{7.4}$$

The larger the ratio d_2/d_1 is, the greater is S (the smaller is the interference) and the better the performance. The objective in a cellular radio system is to allocate frequencies or channels to cells within a cluster so that the distance between interfering cells (co-channel or adjacent channel) is as large as possible. For urban land mobile radio, the distance–power gradient increases from 2 (in the case of free space) to roughly 4, so that the received signal strength falls as the fourth power of the distance. This further improves the signal-to-interference ratio. If there are J_s interfering BSs surrounding a given BS, then the general form of the SNR will be

$$S_\mathrm{r} = \frac{d_0^{\alpha}}{\sum_{n=1}^{J_\mathrm{s}} d_n^{\alpha}} \tag{7.5}$$

where the distance of the mobile from the given BS is d_0 and its distance from the nth BS is d_n.

Example 7.12: S_r in a Hexagonal Cellular Architecture Recalling that there are exactly six co-channel cells with a hexagonal cellular structure, it is clear that they will all cause similar levels of interference to a mobile terminal in the given cell. So $J_\mathrm{s} = 6$ here. Also, the distance at which the co-channel cells are located depends on the size of the cluster from Eq. (7.2). The farthest distance a mobile terminal can be from the BS of a given cell is the cell radius R_L. The approximate distance of the mobile terminal from the BSs of each of the co-channel cells is D_L. For land mobile radio, if only the six co-channel cells that make up the first tier of interferers are considered, then $J_\mathrm{s} = 6$ and the signal-to-interference ratio can be approximated as

$$S_\mathrm{r} \approx \frac{{R_\mathrm{L}}^{-4}}{J_\mathrm{s} D_\mathrm{L}^{-4}} = \frac{{R_\mathrm{L}}^{-4}}{6 D_\mathrm{L}^{-4}} = \frac{1}{6}\left(\frac{D_\mathrm{L}}{R_\mathrm{L}}\right)^4 = \frac{3}{2} N_\mathrm{f}^2 \tag{7.6}$$

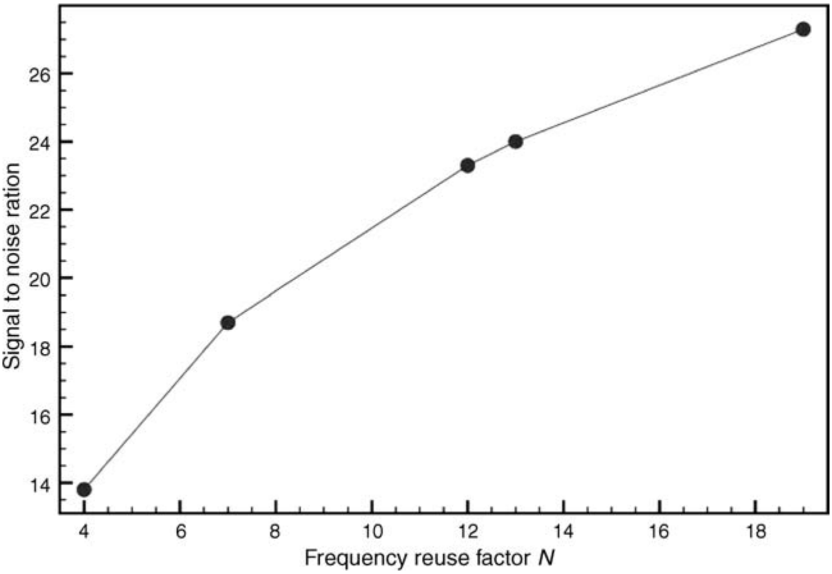


FIGURE 7.39 Signal-to-interference ratio as a function of *N*.

In terms of decibels, we can write the signal-to-interference ratio as

$$S_{\mathrm{r}} = -7.78 + 40\log(D_{\mathrm{L}}/R_{\mathrm{L}}) = 1.76 + 20\log N_{\mathrm{f}} \tag{7.7}$$

Figure 7.39 shows how the signal-to-interference ratio given by Eq. (7.7) varies with the cluster size N. Equation (7.7) is commonly used to determine the cluster size for an adequate performance. Note that the signal-to-interference ratio is influenced by the co-channel reuse ratio $D_{\mathrm{L}}/R_{\mathrm{L}}$, in that a given $D_{\mathrm{L}}/R_{\mathrm{L}}$ has to be maintained for a particular S_{r}. However, it is an approximation, since different BSs may employ different transmit powers and the path loss model may not be as simple as the d^{-4} model used here. The S_{r} calculation will be different for the uplink (mobile terminal to BS communication) compared with the downlink (BS to mobile terminal communication).

We have so far assumed that the received signal strength falls as the fourth power of the distance for land mobile radio. In decibels, as we described in Section 2.3, this translates to a path loss model of the form

$$L_{\mathrm{p}} = L_0 + 40\log d \tag{7.8}$$

The factor L_0 corresponds to the path loss at the first meter or kilometer, as the case may be, and d is in the same units. This path loss model may not be appropriate, especially since measurements of the received signal strength indicate that the path loss is dependent not only on the distance between the BS and the mobile, but also the RF of operation, terrain, and the antenna heights (see Chapter 2). The path loss is also dependent upon the scenario, whether the cellular architecture corresponds to land mobile radio or to a microcellular PCS application. However, this simple model is appropriate for first-cut approximations in system design.

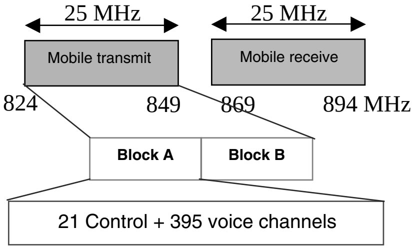


FIGURE 7.40 Frequency spectrum allocation for AMPS.

Let us consider an example of a real cellular system that tries to bring together many of the concepts that have been considered so far in this chapter.

Example 7.13: Cellular Architecture of AMPS As an example of cellular architecture, we consider the very first cellular radio telephone system in the USA, called *AMPS*, based on an analog FM scheme. Each voice channel in this first generation cellular system occupied 30 kHz of bandwidth and used FM. Figure 7.40 shows the spectrum allocations for AMPS.

A bandwidth of 25 MHz is allocated for both the uplink and downlink so that transmission is full duplex. The 25 MHz of spectrum is divided into two blocks of 12.5 MHz each. Block A is allocated to carriers who are not traditional telephone service providers. Block B is allocated to traditional telephone service providers. Each 12.5 MHz of spectrum can support 416 channels, each of which is 30 kHz wide. Of these, 395 are dedicated channels for voice and 21 are dedicated for call control.

Based on subjective voice quality tests, it was determined that a signal-to-interference ratio of 18 dB can be tolerated while providing a good voice quality to the user. From Eq. (7.7), this means that the cluster size has to be $N_f = 7$. Figure 7.38 shows the cellular architecture with this cluster size. Cells with the same label use the same frequency spectrum. They are separated by a distance $D_L = 4.58R_L$ in this case, which ensures that the signal-to-interference ratio is around 18 dB.

Let the 395 voice channels available for a service provider be numbered from 1 to 395. For example, on the downlink, 869–869.030 corresponds to channel 1, 869.030–869.060 to channel 2, and so on. Channels 1, 8, 15, . . . are allocated to cells labeled A. Channels labeled 2, 9, 16, . . . are allocated to cells labeled B, and so on. This ensures that there is a sufficient separation between channels used within a cell so that adjacent channel interference is minimized. In practice, the numbering scheme is different, since the entire 25 MHz of bandwidth was not available for AMPS. However, a separation of seven adjacent channels is maintained between channels used within a cell. It was also found in some cases that, since cells actually do not subscribe to a hexagonal shape and because of the assumptions made in coming up with the value for N, in reality, a cluster size of $N = 12$ has to employed for good voice quality.

7.8.2 Capacity Expansion Techniques for Frequency-/Time-Division Multiple Access Systems

In the past decade, the dominant source of income for the wireless telecommunication industry has been the cellular telephone service. This industry grew exponentially during the last decade of the past millennium. Numerous companies are in fierce competition to gain a portion of the income of this profitable and prosperous industry. The main investment in deploying a cellular network is the cost of the infrastructure, which includes the cost of BS and switching equipment, property (land for setting up the cell sites), installation, and links connecting the BSs. This cost is proportional to the number of BS sites. The income of the service is directly proportional to the number of subscribers. The number of subscribers should grow with time, and a cellular service provider has to develop a reasonable deployment plan that has a sound financial structure to account for many of these aspects. All service providers start their operation with the minimum number of cell sites to cover a service area that requires the least initial investment. As the number of subscribers increases, this generates a source of income for the service provider. At such a point of time, they can increase the investment on the infrastructure to improve service and increase the capacity of the network to support additional subscribers. Therefore, a number of methodologies have evolved to facilitate the expansion of cellular telephone networks.

There are basically four methods to expand the capacity of a cellular network. The simplest method is to obtain additional spectrum for new subscribers. This is a very simple but expensive approach. The so-called PCS bands were sold in the USA for around $20 billion. If we assume that each new subscriber generates a profit of approximately $1000 per year, then we will still need 20 million additional subscribers to recover this amount in a year. With the fierce competition to provide the lowest cost to the customer, this has proved to be suicidal. The reader should not, however, conclude that this is not an acceptable method. With our pessimistic scenario we are accentuating the vital importance of the need for other alternatives to expand capacity in addition to this simple approach of getting additional spectrum.

The second method to expand the capacity of a cellular network is to change the cellular architecture. Architectural approaches include cell splitting, cell sectoring using directional antennas, and using multiple reuse factors (called reuse partitioning). These techniques, described in detail in the rest of this section, change the size and shape of the coverage of the cells by adding cell sites or modifying the nature of antennas to increase the capacity. These techniques do not need additional spectrum or any major changes in the wireless modem or access technique of the system that will require the user to purchase a new terminal. These features of architectural approaches distinguish them as one of the more practical and less expensive solutions to expand the network capacity.

The third method for capacity expansion is to change the frequency allocation methodology. Rather than distributing existing channels equally among all cells, it is possible to use a nonuniform distribution of the frequency bands among different cells according to their traffic need. The traffic load of each cell is dynamically changed by the geography of the service area and with time, depending on the traffic load. In most downtown areas we have the largest traffic loads during rush hours and a relatively light traffic load in the evening hours and weekends. This situation is reversed in residential areas. If channels are allocated dynamically to different cells, then we can increase the overall capacity of the network. These techniques do not need any change in the terminal or physical architecture of the system and they are implemented somewhere inside the computational devices used for network control and management.

The fourth, and most effective, method to expand the network capacity is to change the modem and access technology. The cellular industry started with analog technology using FM, evolved towards TDMA, and then a CDMA air interface using digital modems. As we saw in the case of HDR and GPRS/EDGE, changing the coding and modem technology can also enhance the capacity of a system considerably. Digital technology increases the network capacity and also provides a fertile environment for integration of voice and data services. However, this migration requires the user to purchase new terminals and the service provider to install new components in the infrastructure.

We consider some architectural approaches for expanding capacity of cellular networks in more detail. In particular, we consider two approaches: cell splitting, which is useful for incremental capacity improvement, and cell sectoring, which is widely employed in today's cellular systems.

Cell Splitting. As the number of subscribers increases within a given area, the number of channels allocated to a cell is no longer sufficient for supporting the subscriber demand. It then becomes necessary to allocate more channels to the area that is being covered by this cell. This can be done be *splitting* cells into smaller cells and allowing additional channels in the smaller cells.

Consider Fig. 7.41. In this figure we have a cellular architecture where a cluster size of seven is employed. When the traffic load increases, a smaller cell is introduced such that it has half the area of the larger cells. This will ultimately increase the capacity fourfold (since area is proportional to the square of the radius). However, in practice, only a single small cell will be introduced such that it is midway between two co-channel cells. In this case, these are the larger cells labeled A. Thus, it is logical to reuse the channels allocated to these cells in the smaller cell to minimize the interference.

This approach gives rise to some problems. Let us suppose that the radius of the smaller split cell (labeled **a**) is $R_L/2$. Let the transmit power of the BS of the small cell be the same as the transmit powers of the larger cells. As far as the smaller cell is concerned, the signal-to-interference ratio is maintained because the maximum distance the mobile can be from the BS, which in this cell is $R_L/2$. So, though the distance between this cell and the co-channel cells A is reduced by half, the value of the signal-to-interference ratio S remains the same. On the other hand, this is not the case for the cells labeled A, since the co-channel reuse ratio for these cells is now $D_L/2R_L$ with respect to the smaller split cell. In order to maintain the same level of interference, the transmit power of the BS in the smaller cell should be reduced. But this will increase the interference observed by the mobiles in the smaller cell. The other alternative is to divide the channels allocated to cells labeled

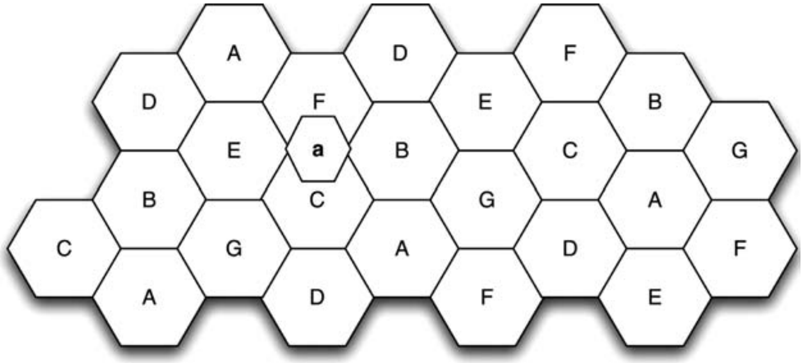


FIGURE 7.41 Cell splitting.

A into two parts: those used by **a** and those not used by **a**. The channels used by **a** will be used in the larger cells only within a radius of $R/2$ from the center of the cell so that the co-channel reuse ratio will be maintained as far as these channels are concerned. This is called the *overlaid cell concept*, where a larger *macrocell* co-exists with a smaller *microcell.*

The downside of this approach is that the capacity of the larger cells is reduced, which will ultimately lead to introducing split cells in their area, until such time that a chain reaction will result in the entire area being served by cells of a smaller radius. Also, the BSs in cells labeled A will become more complex and there will be a need for handoffs between the overlays.

Using Directional Antennas for Cell Sectoring. The simplest and the most popular scheme for expanding the capacity of cellular systems is cell sectoring using directional antennas. This technique attempts to reduce the signal-to-interference ratio and thus reduce the cluster size, thereby increasing the capacity. The idea behind using directional antennas is the reduction in co-channel interference that results by focusing the radio propagation in only the direction where it is required. In order to achieve this, the coverage of a BS antenna is restricted to part of a cell called a *sector* by making the antenna directional. In implementing this technique cell site locations remain unchanged and only the antennas used in the site will be changed. The main objective here is to increase the signal to interference ratio to a level that enables us to use a lower frequency reuse factor. A lower frequency reuse factor allows a larger number of channels per cell, increasing the overall capacity of the cellular network.

As we discussed earlier, see Eq. (7.5), the signal to interference ratio is given by

$$S_r = \frac{1}{J_s}\left(\frac{D_L}{R_L}\right)^4 = \frac{9}{J_s}N_f^2 \tag{7.9}$$

where J_s is the number of interfering cell sites. Using a sector antenna reduces the factor J_s, resulting in the interference and an increase in S_r. The most popular directional antennas employed in cellular systems are 120° directional antennas. In some cases 60° directional antennas are also employed. In the following two examples we evaluate the impact of these antennas that enable the reuse factor to be reduced from $N_f = 7$ to $N_f = 4$ and $N_f = 3$ respectively.

Example 7.14: Three-Sector Cells and a Reuse Factor of $N_f = 7$ Consider a seven-cell cluster scheme with 120° directional antennas shown in Fig. 7.42. Channels allocated to a cell are further divided into three parts, each used in one sector of a cell. As shown in the figure, the number of co-channel interfering cells is reduced from six to two. Thus, there is an improvement in the signal-to-interference ratio. For omnidirectional antennas (see Examples 7.12 and 7.13), the value of S for a cluster size of $N_f = 7$ is 18.66 dB. In this case, in a manner similar to Eq. (7.7), the signal-to-interference ratio is given by

$$S_r \approx \frac{R_L{}^{-4}}{J_s D_L^{-4}} = \frac{R_L{}^{-4}}{6 D_L^{-4}} = \frac{1}{2}\left(\frac{D_L}{R_L}\right)^4 = \frac{9}{2}N_f^2 \tag{7.10}$$

For $N_f = 7$, this will give us $S_r = 23.43$ dB. To see the importance of this gain, note that the required SNR for AMPS systems is 18 dB, which suggests $N_f = 7$. However, a larger S_r is required because of nonideal situations.

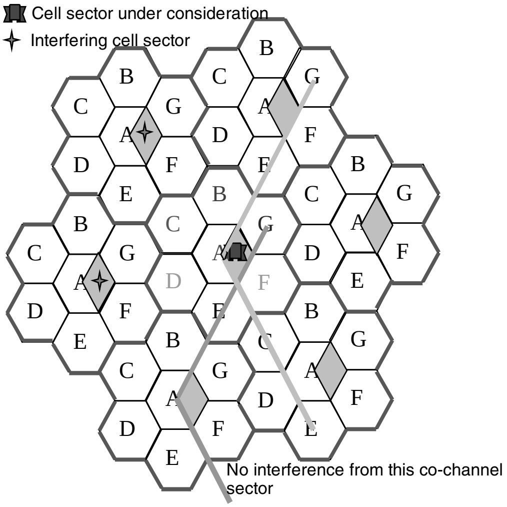


FIGURE 7.42 Seven-cell reuse with 120° directional antennas (three-sector cells).

Example 7.15: Three-Sector Cells and a Reuse Factor of $N_f = 4$ Equation 7.10 remains unchanged in this case, as there are only two interfering cells again (see Fig. 7.43). With omnidirectional antennas, $J_s = 6$ and for $N_f = 4$, we end up with $S_r = 13.8$ dB, which is woefully inadequate for AMPS.

It can be seen that the signal-to-interference ratio with three-sector cells is substantially better compared than omnidirectional antennas and no cell sectoring. With $N_f = 4$, the signal-to-interference ratio is 19.9 dB. This value is larger than the requirement of 18 dB based on subjective mean-opinion-score (MOS) tests of voice quality.

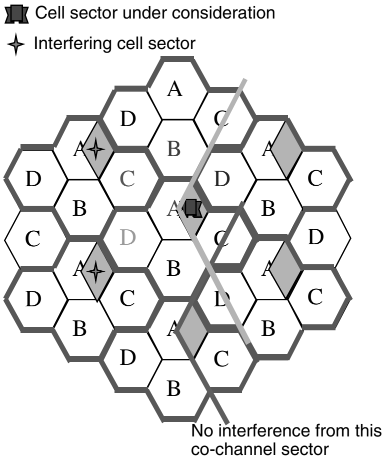

Cell sector under consideration
Interfering cell sector
A
D C
A B A
D C D C
B A B
C D C D
A B A
D C D C
B A B
D C
B
No interference from this co-channel sector

FIGURE 7.43 Four-cell reuse with 120° directional antennas (three-sector cells).

Example 7.16: Six-Sector Cells and Reuse Factors of $N_f = 4$ and 3 With 60° directional antennas, we have six sectors within a cell. The number of interfering co-channel cells reduces to one and the signal-to-interference ratio can be written as

$$S_r \approx \left(\frac{D_L}{R_L}\right)^4 = 9N_f^2 \tag{7.11}$$

It is possible to employ a cluster size of four or three with six-sector cells, since the signal-to-interference ratio will be 21.58 dB or 19.1 dB respectively, which has a sufficient margin for AMPS. The cellular layout and relation to sectors in this case is left as an exercise for the readers.

In practice, we cannot ideally sector a cell because ideal antenna patterns cannot be implemented. Therefore, the numbers obtained in the above examples for ideal cell sectors are optimistic. However, our conclusion from the above examples is that the use of sectoring increases the signal-to-interference ratio at the terminal. We should emphasize that in the particular examples we could reduce the frequency reuse factor from $N_f = 7$ to $N_f = 4$ or even $N_f = 3$ by using three and six-sector cells respectively. This reduction in frequency reuse from seven to four or even three would result in capacity increase of 1.67 and 2.3 respectively, allowing an equal increase in the number of subscribers and, consequently, in income of the service provider. The service provider needs to add these antennas to the BSs in the desired area. Compared with the cell-splitting method, using directional antennas is less effective in increasing capacity, but it can be significantly less expensive. The cost of additional cell sites, needed in cell splitting, includes the costs of the property and installing the antenna mounting tower, which are usually far more expensive than deploying directional antennas. Cell splitting also requires additional planning efforts to maintain interference levels in the smaller cells. If directional antennas are used without reduction in the frequency reuse factor, then the average required transmitted signal power from the MSs will be reduced, which can potentially result in longer battery life for the user.

7.8.3 Network Planning for Code-Division Multiple Access Systems

CDMA presents some unique features that are not present in traditional TDMA and FDMA systems. In TDMA and FDMA systems, the users operating in one channel are completely isolated from the users operating in other channels. The only interference comes from the fact that the same frequency bands are employed in spatially separated cells and this interference is the co-channel interference. Of course, leakage of signal from adjacent bands causing adjacent channel interference is also a factor, but intelligent design can reduce this effect greatly. However, in the case of CDMA, all users are operating on the same frequency channel at the same time, resulting in everyone causing co-channel interference. This problem is reduced on the downlink by employing time-synchronized orthogonal codes. On the uplink, a combination of convolutional coding, spreading, and orthogonal modulation are employed to combat the effects of this interference. Network planning in the case of CDMA is far more complicated than in the case of TDMA/FDMA in that sense; but at the same time, using CDMA completely eliminates the concept of conventional frequency reuse, since the same frequencies can be deployed in all cells.

Instead of defining an acceptable signal-to-interference ratio, in CDMA it is necessary to define the *quality of the signal* [Hal96]. Usually, this is expressed in terms of the acceptable energy per bit to total noise ratio E_b/N_t, which results in roughly a 1% data frame error rate. The reason for selecting this as a measure is that this frame error rate results in acceptable speech quality at the vocoder output. The value of E_b/N_t is usually between 6 and 11 dB, depending on the speed of the mobile terminal, propagation conditions, the number of multipath signals that can be used for diversity, etc. The value of N_t depends on the number of interfering signals and the transmit powers of the interfering users. Consequently, power control and appropriate thresholds play a very important role in the coverage of a CDMA cell and the soft handoff process associated with it.

As we discussed in Section 5.2.4, the number of active traffic channels (or calls) per cell that are possible at a given point of time is given in the case of CDMA by

$$M = \frac{W}{R_b}\frac{1}{S_r}\frac{G_A G_v}{H_0} \tag{7.12}$$

where W/R_b is the processing gain or spreading factor of the system, H_0 is the additional interference from adjacent cells (usually around 2 dB loss),[1] G_v is the voice activity factor (about 0.45), and G_A is the gain due to sectored cells (2.5 for a three-sectored cell). From this equation, it is clear that cell sectoring, discussed previously in the context of FDMA/TDMA-based cellular systems, has a bigger impact in improving the capacity of CDMA systems.

Many of the principles that apply to TDMA/FDMA systems also apply to CDMA systems, but there are important differences. For example, the path loss is very similar to TDMA systems, in that the signal strength drops roughly as the fourth power of the distance in macrocells and is quite site specific and terrain dependent. However, some design issues that differ are described below.

Managing the Noise Floor. In CDMA, managing the noise floor is very important. If the number of users in a particular area increases beyond that dictated by Eq. (7.12), then the system is interference limited and increasing the transmit power will not benefit any user or set of users, as the total interference also increases. It is quite possible that interference from many cells can raise the noise floor to such a level that *holes* may be created in the region where the coding/spreading gain is not sufficient to overcome the interference levels. This is illustrated in Fig. 7.44 [Hal96]. If there is an isolated three-sector cell, then most of the cell has an S_r larger than 7 dB, and in regions where there is soft handoff (where the mobile terminal can connect to more than one BS) the S_r value from each BS is around 3 dB, providing sufficient diversity gain to allow communication. If too many cells are deployed, then, as shown in the figure, there may be some regions where the noise level is so high that it is impossible to communicate. It is often possible to cover the same area with fewer cells to reduce the total interference levels, and it is usually not a very good idea to cover an area by more than three cells or cell sectors. The problem becomes more severe when terrain plays a role; then, in addition, to site selection, it will be important to use the downtilt of antennas and the use of minimum radiated power levels to manage the noise floor.

Cell Breathing. In CDMA, the boundary of a cell is not fixed; it depends on where the S_{rt} value is reached. For example, consider the uplink S_r value that is observed at a BS. As the number of traffic channels on the uplink is increased, this value also increases, and it is clear from Eq. (7.12) that the handoff boundary (where the mobile terminal has to move from one

[1] Sometimes a parameter α (between 0.5 and 0.9) is also included to count for the power control inaccuracy [Gar00].

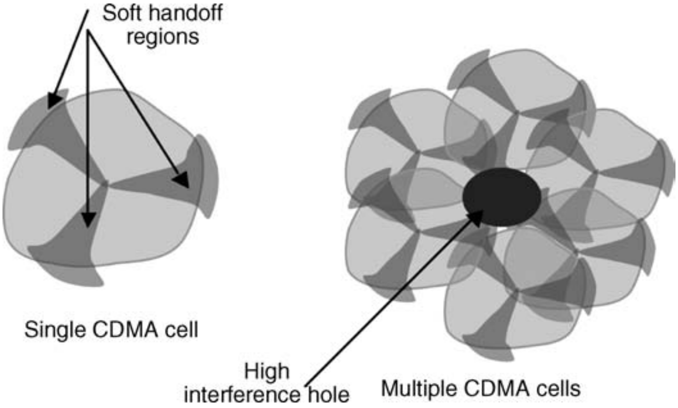


FIGURE 7.44 Noise floor management in CDMA.

BS to another) shifts closer to the BS. This effect is called cell breathing. In order to ensure that a correct handoff is performed, the transmit power of the pilot channel of the BS must also be reduced so that the forward link handoff boundary is also maintained at the same level as the reverse link boundary. In some cases, cell breathing can have a deleterious impact on the system performance, and this should be taken into account while planning the system, either by deploying more cells or offloading capacity to other carriers.

QUESTIONS

1. What are the differences between a mobile digital telephone and POTS?
2. Name the three subsystems in the general cellular architecture.
3. What are VLR and HLR, where they are physically located and why we need them?
4. What is the difference between registration and call establishment?
5. What are the reasons to perform handoff?
6. What is the difference between network decided and mobile assisted handovers?
7. What is the difference between a logical and physical channel?
8. Name five logical channels in cellular networks?
9. What is the difference between a pure “layer-2” message and messages that are otherwise carried by layer-2 in cellular networks?
10. What is the importance of the framing structure in GSM?
11. What are the incentives for power control in a TDMA network? Name the elements of the GSM system that are involved in handling power control.
12. How does GSM convert 456 bits of the speech, data, or control signal into a normal burst of 156.25 bits?
13. Name the forward and reverse channels used in IS-95.
14. How are Walsh codes employed in the cdmaOne forward and reverse channels? Explain the difference.
15. Why does W-CDMA use Walsh codes in forward and reverse channels for separating users and cdmaOne uses them only in the forward channel?
16. Handoff decision in wireless networks is performed using received signal strength measurements. Name the forward channel in IS-95 that is used for this purpose.

17. Why are several pilot channels monitored in IS-95?
18. Why are both architectural changes and changes in transmission schemes necessary for data transmission on cellular networks that were designed for voice?
19. Of the following, what values are possible for a cluster size in a cellular topology? Why? Assume a hexagonal geometry: 8, 21, 23, 30, 47, 61, 75
20. Name an architectural method that can be used to increase the capacity of cellular system without increasing the number of antenna sites.
21. How are high interference holes in CDMA deployment created?

PROBLEMS

Problem 1:

(a) Using the bit and time durations in Fig. 7.18 show that speech coding rate for GSM is 13 Kbps and the effective transmission rate to support one 13 Kbps coded voice channel is 22.8 Kbps.

(b) What is the required transmission bandwidth for eight slots of the GSM system?

(c) Give the overall overhead rate of the system - that is the difference between the required transmission rate for the traffic and the actual transmission rate of the GSM.

(d) Determine the efficiency of the system - that is the ratio of the overhead over raw transmission rate.

Problem 2:

(a) Consider the multiframe transmission in GSM depicted in Fig. 7.23. Use the overall structure of the multiframe, frame and slot to show that the transmission rate of the GSM is indeed 270.833 bps.

(b) In each GSM multiframe 24 frames are used for traffic and two for associated control signaling. Considering the detailed burst frame and multiframe infrastructure show that the effective transmission rate for each GSM voice traffic is 22.8 Kpbs.

(c) The slow association control channel uses 114 bits of one slot of each 26 slot traffic multiframe. What is the transmission rate for this channel in bits per second?

Problem 3:

The *stand-alone dedicated control channel* is a channel in GSM that uses four time slots per each 51-control multiframe shown in Fig. 7.23. Use the superframe timing to determine the effective data rate of this logical channel.

Problem 4:

Considering Fig 7.21 give the net data rate (data plus signaling) and the effective transmission rate of a 9600bps GSM data service.

Problem 5:

(a) What is the allowable power ramping time for GSM receivers? Hint: The time gaps of normal, frequency correction and synchronization bursts, shown in Fig 7.18, are designed to allow power ramping.

(b) The time gap of the random access burst, shown in Fig. 7.18, is designed to assure this packet does not collide with the normal bursts. What is the maximum coverage, the distance between the BS and MS, a GSM base station? Assume that this gap is reserved for two-way travel and radio wave travel at 300,000 Km/sec.

(c) The length of the synchronization sequence in synchronization burst is designed to allow time advance for two-way bit synchronization. Use this parameter to calculate the maximum coverage GSM. Compare your results with that of part (b).

Problem 6:

Sketch all sixteen of the sixteen-bit Walsh functions.

Problem 7:

Assume that you have six-sector cells in a hexagonal geometry. Draw the hexagonal grid corresponding to this case. Compute S_r for reuse factors of 7, 4 and 3. Comment on your results.

Problem 8:

We have an installed cellular system with 100 sites, frequency reuse factor of K = 7 and 500 overall two-way channels. Give the number of channels per cell, total number of channels available to the service provider, and the minimum carrier to interference ratio (C/I) of the system in dB. Assume hexagonal cells.

Problem 9:

(a) What is the number of RF channels per cell in the GSM network? The frequency reuse factor of the GSM is K = 4.

(b) What is the maximum number of simultaneous users per cell in this system?

(c) Assume that we want to replace this GSM system with an IS-95 spread spectrum system in the same frequency bands. What is the maximum number of users per cell? Assume an ideal power control and use the practical considerations for the IS-95 system.

Problem 10:

(a) Determine the carrier to interference ratio, in dB, of a cellular system with frequency reuse factor of K = 7.

(b) Repeat (a) for K = 4.

(c) If we consider multi-symbol QAM modulation for the digital transmission of the information, how many more bits per symbol can be transmitted with K = 4 as it is compared with K = 7 architecture?

Problem 11:

We want to use a GSM system with sectored antennas (3 sectors and reuse factor K = 3) to replace an existing AMPS system (with a reuse factor of K = 7) with the same cell sites. In the existing AMPS system the service provider owns 395 duplex voice channels operating in 12.5 MHz of band.

(a) Determine the number of voice channels per cell for the AMPS system.
(b) Determine the number of voice channels per cell for the GSM system.
(c) Repeat (b) if we were using a W-CDMA system with the bandwidth of 12.5MHz for forward channel. Assume a signal to noise ratio requirement of 4 (6dB) and include the effects of antenna sectorization (2.75), voice activity (2), and extra CDMA interference (1.67).

Problem 12:

(a) Considering the frequency allocation strategy of Fig 7.17 for the GSM systems, give the total number of traffic channels per 50 MHz of bandwidth used for two-way GSM communications.
(b) Give the total number of GSM channel per MHz of bandwidth.
(c) Give the number of channels per cell for frequency reuse factors of N = 4 and N = 3.

Problem 13:

The IS-136 system was the digital equivalent of GSM in the US, but used a channel structure derived from AMPS. Repeat Problem 11 for an IS-136 system assuming that this system replaces an AMPS system with 395 traffic channels and a frequency reuse factor of N = 7. Assume that each carrier in IS-136 can carry 3 voice calls.

Problem 14:

Using Table 7.2, show exactly how a data rate of 921.6 kbps can be achieved using HDR. Compare the provisioning of 236.8 kbps with EDGE and 307.2 kbps with HDR using 4 slots each (use Tables 7.2 and Table 7.4). What differences and (dis)advantages can you enumerate?

Problem 15:

A cellular service provider has 235 cells, a 12.5MHz of bandwidth, and uses AMPS technology with a frequency reuse factor of 7.

(a) Calculate the maximum number of subscribers for the service provide if the providers intendeds to keep the average blockage rate below 2%. Assume the average calls of the network are 2 calls/hour and each user has an average holding time of 3 min.

(b) Repeat (a) if the provider reduces the blockage rate to less than 1%.

(c) Repeat (a) if the service provider resorts to GSM with a frequency reuse factor of 3.

(d) Repeat (a) if the service provider uses 3G CDMA technology with the user data rate of 9600bps.

Note: you can use web resources to calculate the results needed for the Erlang equations.

Problem 16:

The maximum recommended coverage of GSM is 37Km. Use the Okumura-Hata model from Chapter 2 to compute the minimum receiver sensitivity in dBm to provide the maximum coverage at a frequency of 900MHz. Assume a large city environment, 1.5m mobile antenna height, 100 meters base station antenna height, and a maximum transmitted power of 3 W.

PART THREE

LOCAL AND PERSONAL-AREA NETWORKS

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

8

IEEE 802.3 ETHERNET

8.1 INTRODUCTION

LANs are used to connect computer terminals, mainframes, printers, and other equipment in small geographic areas in offices and homes. The main differences between LANs and WANs are the higher data transfer rates, smaller geographic range, and ownership of the network. A LAN is usually owned privately, while WANs are owned by a service provider leasing the service to different people or organizations. Today, the most popular wired LAN technology is the Ethernet operating over UTP cabling, but a variety of other technologies, such as token ring or token bus, have been used and competed with the Ethernet in the past several decades.

The cost of infrastructure in WANs is very high and the coverage is very wide. As a result, WANs are offered as a charged *service* to the user. The service provider invests a large amount of capital for the installation of the infrastructure and generates revenue through

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

monthly service charges. Local networks are sold as end products to the user and there is no service payment for local communications. Operation of LANs is very similar to a PBX, in that the user owns them and pays monthly charges to the wide-area Internet service providers for wide-area communications.

As we explained in Section 1.3.1, the LAN industry emerged during the 1970s to enable sharing of expensive resources like printers and to manage the wiring problem caused by the increasing number of terminals in offices. By the early 1980s three standards were developed: Ethernet (IEEE 802.3), token bus (IEEE 802.4) and token ring (IEEE 802.5), specified three distinct MAC and PHY layers and different topologies for networking over thick-cable medium but shared the same management and bridging (IEEE 802.1) and LLC (IEEE 802.2). With the growing popularity of LANs in the mid 1980s, the high installation costs of thick cable in office buildings moved the LAN industry toward using thin cables, which was also referred to as "cheapernet." Cheapernet covered shorter distances of up to 185 m compared with the 500 m coverage of thick cables. In the early 1990s, the star topology (often referred to as hub-and-spoke LANs) using easy-to-wire twisted-pair wiring with coverage of 100 m was introduced. Figure 1.9 graphically describes the early days of the evolution of the Ethernet. The interesting observation is that this industry has made a compromise on the coverage to obtain a more structured solution that is also easier to install. Twisted-pair wiring, also used by PSTN service providers for telephone wiring distribution in homes and offices for over 100 years, is much easier to install. The star network topology opened an avenue for structured hierarchical wiring, also similar to the telephone network topology. Today, IEEE 802.3 (Ethernet) using twisted-pair wiring is the dominant wired LAN technology and it is the focal point of this chapter.

The data rates of legacy LANs (thick, thin, and twisted pair) were all 10 Mb/s. The need for higher data rates emerged from two directions: (1) there was a need to interconnect LANs located in different buildings of a campus to share high-speed servers and (2) computer terminals became faster and capable of running high-speed multimedia applications. To address these needs, several standards for higher data-rate operations were introduced. The first fast LAN operating at 100 Mb/s was the FDDI that emerged in the mid 1980s as a backbone medium for interconnecting LANs. The ANSI published this standard directly. In the mid 1990s, 100 Mb/s fast Ethernet was developed under IEEE 802.3. In the late 1990s, IEEE 802.3 approved the gigabit Ethernet and today the 10 Gb/s Ethernet is standardized and it is striving to attract a market both in the local- and wide-area networking. All of these high-speed LANs use fiber optics, high-quality twisted pair, and multiple twisted-pair wirings to support faster transmission. Figure 8.1 shows an example of a hierarchical wiring of a LAN. A variety of 10 and 100 Mb/s terminals are connected with two levels of switches and repeaters to a router that connects the LAN to the rest of the world.

In the mid 1990s, when most people in the telecommunication industry believed that ATM would take over the entire emerging multimedia communications industry, ATM-LANE was initiated. The purpose of ATM-LANE was to adapt the existing legacy LAN infrastructures and applications to the then perceived end-to-end ATM network. The main technical challenge in implementing LANE was the adaptation of a connectionless legacy LAN to a connection-oriented ATM network. Details of the variety of LANs are available in Stallings [Sta00] with good descriptions. Around the same time the 100VG-AnyLAN (IEEE 802.12) tried to provide a single 100 Mb/s LAN solution for both Ethernet and its closest competitor of the early days, the IEEE 802.5 token ring.

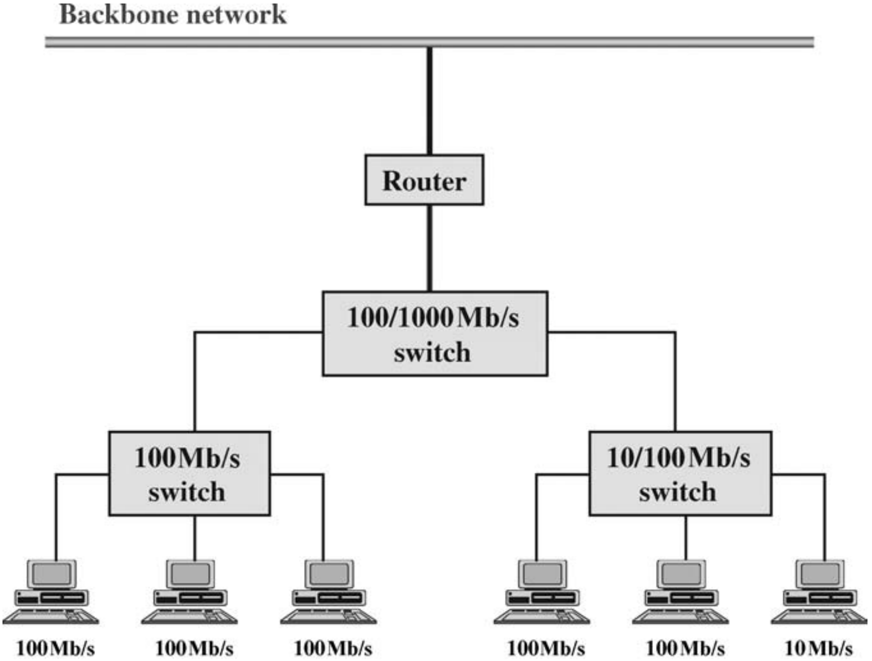


FIGURE 8.1 Hierarchical LANs.

In summary, the LAN industry has developed a number of standards, mostly under the IEEE 802 community. Figure 1.10 shows the overall structure of the IEEE 802 community standards, Table 8.1 lists the name of the standard series, and Fig. 8.2 relates the protocol stack of the IEEE 802 standards with those of IETF Internet and OSI protocol stack reference models. The 802.1 (higher level interfaces for bridging) and 801.2 (LLC) parts are common for all the standards, 802.3, 802.4 and 802.5 are wired LANs, and 802.9 is the so called ISO-Ethernet that supports voice and data over the traditional Ethernet mediums. IEEE 802.6 corresponds to metropolitan area networking and the IEEE 802.11, 15, and 16 are related to wireless local networks. IEEE 802.14 is devoted to cable-modem-based networks providing Internet access through cable TV distribution networks operating over coaxial cable wiring and fiber originally installed for TV distribution. IEEE 802.10 is concerned with security issues and operates at higher layers of the protocols. The existing LANs can be logically divided into four generations: *first-generation* legacy LANs, 802.3 and 802.5, that provided terminal-to-host connectivity and client–server architectures at moderate data rates of up to 10 Mb/s in offices; *second-generation* LANs, such as FDDI, that responded to the need for backbone LANs and support of high-performance workstations; *third-generation* LANs, such as ATM-LANE, fast Ethernet, gigabit Ethernet, and 100VG-AnyLAN, that were designed for high throughput with delay control for multimedia applications; and *fourth-generation* LANs supporting data rates of 1 and 10 Gb/s which compete with SONET for backbone of the MANs and WANs. The three major drives of this industry have always been the ease of installation, increase of data rate, and popularity of the technology to support mass production at lower costs.

TABLE 8.1 Overview of the Important IEEE 802 Standard Series

802.1:	Higher level interface (HILI)
802.2:	Logical Link Control (inactive)
802.3:	CSMA/CD Ethernet (10 Mb/s)
802.4:	Token Bus (disbanded)
802.5:	Token Ring (inactive)
802.6:	MAN (disbanded)
802.7:	Broadband Technical Advisory Group (disbanded)
802.8:	Fiber Optics Technical Advisory Group (disbanded)
802.9:	Integrated Service LAN Interface (disbanded)
802.10:	Standard for Interoperable LAN Security (disbanded)
802.11:	Wireless LANs
802.12:	Demand Priority (inactive)
802.14:	Cable TV Based Broadband Communication Networks (disbanded)
802.15:	WPAN
802.16:	LMDS and then WiMax
802.17:	Resilient Packet Ring (RPR) Working Group
802.18:	Radio Regulatory Technical Advisory Group
802.19:	Coexistence Technical Advisory Group
802.20:	Mobile Wireless Access Working Group
802.21:	Media Independent Handover Working Group
802.22:	Wireless Regional Area Network

Although switched Ethernet is now the most common data link layer protocol and IP as a network layer protocol, many different options have been used, and some continue to be popular in niche areas. Smaller LANs generally consist of a one or more switches linked to each other – often with one connected to a router, cable modem, or DSL modem for Internet access. Larger LANs are characterized by their use of redundant links with switches using the spanning tree protocol to prevent loops, their ability to manage differing traffic types via QoS, and to segregate traffic via VLANing. LANs may have connections with other LANs

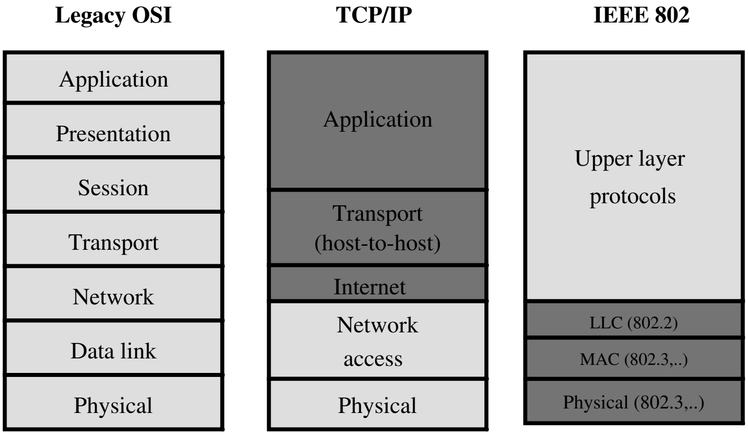


FIGURE 8.2 Three popular reference models.

via leased lines, leased services, or by "tunneling" across the Internet using virtual private network (VPN) technologies.

8.2 LEGACY ETHERNET

LANs were originally designed to solve the wiring problem in the offices where a huge number of terminal cables were drawn to a box to share expensive resources such as mainframe computers, printers, and mass storage devices. The original legacy Ethernet solution was based on the idea of a shared coaxial cable acting as a broadcast transmission medium to access all terminals. Figure 8.3 shows the overall architecture of the legacy Ethernet using a bus topology connecting all terminals and a server to a thick coaxial cable with a maximum length of 500 m. The end of the network needs a terminator at each end for proper operation; this is a drawback for maintenance of the network, because accidental removal of the terminator stops the network operation. The standard allows four repeaters, which extend the length to 2.5 km. When the signal is sent from one end of the cable it becomes attenuated as it travels through the cable. For thick coaxial cables used in the legacy Ethernet, at distances more than 500 m the strength of the signal against the background noise is so low that the error rate of the received bits is no longer acceptable and a repeater is needed to boost the level of the signal. Each time that the signal level is boosted the accompanying noise is also boosted, and accumulation of the noise for more than four repeaters does not allow proper operation over the cable. Another parameter affecting the decision on the length of the cable is the overall delay of the cable, which we discuss later when we talk about the MAC of the Ethernet.

Each terminal hung over the cable using a vampire connector, shown in Fig. 8.4. This approach is somehow similar to wireless systems, with the difference that the multipath fading in wireless channels make detection of collision between the packets much more difficult than with the cables. The common cable providing the communication channel played a role similar to the "ether," a substance once thought to fill all space to carry electromagnetic waves, and for that reason the network was referred to as "Ethernet."

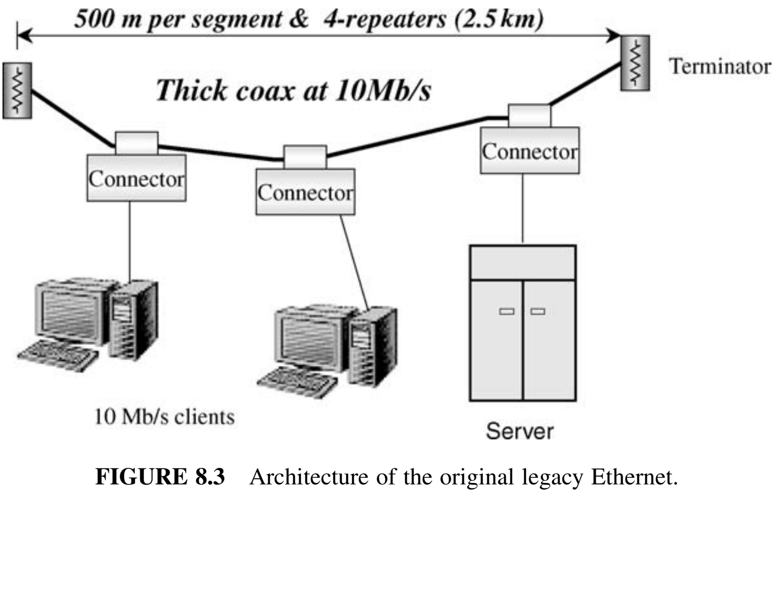


FIGURE 8.3 Architecture of the original legacy Ethernet.

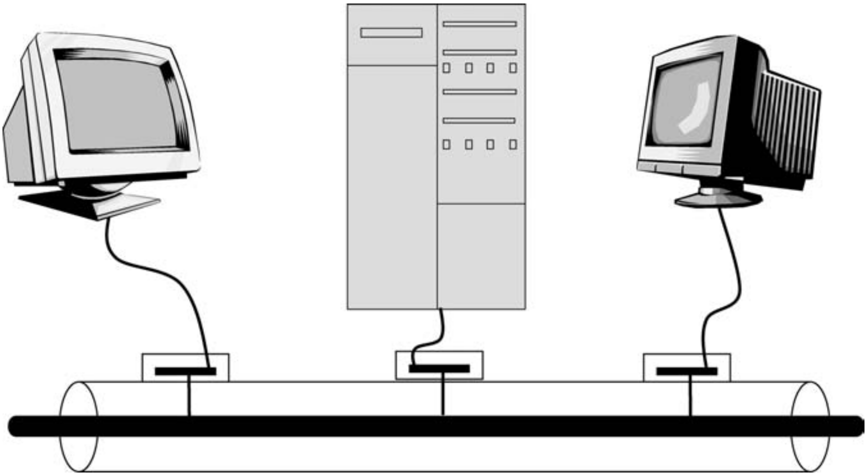


FIGURE 8.4 Vampire connectors in legacy Ethernet.

Each device is connected to the Ethernet vampire connector with an NIC, as shown in Fig. 8.5, whose role is to act as a buffer to regulate the data transmission among different terminals operating with different clocks. The computer data buses usually operate as parallel lines with a certain data rate; the NIC reads/writes the data from the terminal bus, buffers them in the buffers, adds some header to create a frame with a specific format, and transmits the frame at the specified data rate (10 Mb/s for legacy Ethernet) for the LAN in the cable medium. The legacy Ethernet is also known as 10Base5 reflecting 10 Mb/s baseband line coding covering up to 500 m per segment of the cable.

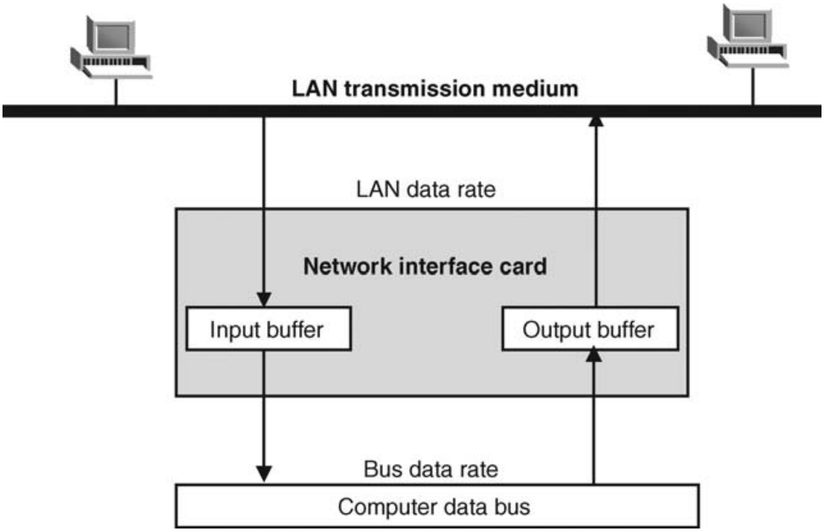


FIGURE 8.5 Connections to legacy Ethernet.

← - - - - - - - - - - - - - - - - - - - Minimum 64 bytes - - - - - - - - - - - - - - - - - - →

Preamble (7)	Start delimiter (1)	DA (2 or 6)	SA (2 or 6)	Length of data (2)	Data (0-1500)	Pad (0-46)	Checksum (4)

FIGURE 8.6 Frame format of the IEEE 802.3 Ethernet.

8.2.1 The Packet Format and the Physical Layer

The IEEE 802.3-recommended frame format for the Ethernet is shown in Fig. 8.6. The different parts of the frame are defined as follows:

Preamble – a sequence of 56 bits having alternating 1 and 0 values that are used for synchronization. They serve to give components in the network time to detect the presence of a signal, and synchronize for reading the signal before the frame data arrives.

Start frame delimiter – a sequence of 8 bits having the bit configuration 10101011 that indicates the start of the frame.

DA/SA – the destination/source MAC address field 802.3 standard permits either 2 bytes or 6 bytes, but virtually all Ethernet implementations use 6-byte addresses. A destination address may specify either an "individual address" destined for a single station or a "multicast address" destined for a group of stations. A destination address with all bits 1 refers to all stations on the LAN and is called a "broadcast address."

Length/type – if the value of this field is less than or equal to 1500, then the length/type field indicates the number of bytes in the subsequent MAC client data field. If the value of this field is greater than or equal to 1536, then the length/type field indicates the nature of the MAC client protocol (protocol type).

MAC client data – this field contains the data transferred from the source station to the destination station or stations. The maximum size of this field is 1500 bytes. If the size of this field is less than 46 bytes, then use of the subsequent "pad" field is necessary to bring the frame size up to the minimum length.

Pad – if necessary, extra data bytes are appended in this field to bring the frame length up to its minimum size. A minimum Ethernet frame size is 64 bytes from the destination MAC address field through the frame check sequence.

Frame check sequence – this field contains a 4-byte CRC remainder for dividing into a 33-bit number. When a source station assembles a MAC frame, it performs a CRC calculation on all the bits in the frame from the destination MAC address through the pad fields (that is, all fields except the preamble, start frame delimiter, and frame check sequence). The source station stores the value in this field and transmits it as part of the frame. When the frame is received by the destination station, it performs an identical check. If the calculated value does not match the value in this field, then the destination station assumes an error has occurred during transmission and discards the frame.

The minimum length of the packet is 64 bytes and the maximum is 1518 bytes, allowing 1500 bytes of data plus 18 bytes of overhead. As we explain later, the minimum length is needed to ensure the collision detection process in the MAC protocol. Ethernet frames allow a 96-bit (9.6 μs for 10 Mb/s or 96 ns for 1 Gb/s) interframe time gap or space to allow the device a brief recovery time and prepare for the next frame. The PHY layer of the legacy Ethernet uses differential Manchester coding, as shown in Fig. 8.7. As we described in Chapter 2, this line coding technique encodes the data stream in the transitions at the

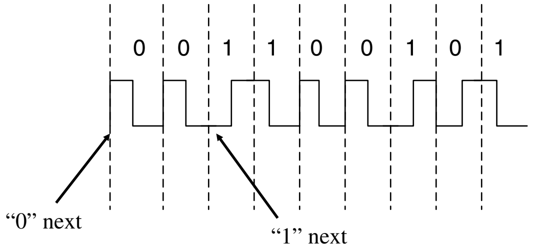


FIGURE 8.7 Differential Manchester coding used as the PHY of the legacy Ethernet.

beginning of each bit. If the transmitted bit is a 0 then we have a transition at the start of the bit, and if it is a 1 then we have no transitions at the start of the bit. To have adequate transitions for synchronization at the receiver, this line coding technique enforces one transition in the middle of every bit transition interval. This is a very simple coding technique which could support the desired 10 Mb/s data rate over the thick cable. The efficiency of this coding technique is 50%, because actually each bit is transmitted with two neighboring pulses at the symbol transmission rate of 20 MS/s. As we will see later on in this chapter, the need for higher data rates has introduced more efficient line-coding techniques for implementation of PHY during the evolution of the Ethernet technology.

8.2.2 Carrier Sense Multiple Access with Collision Detection for the Medium Access Control Layer

Legacy Ethernet used so-called "thick cable" as the shared medium winding around a building or campus to every attached terminal and machine. This terminal shared the medium using the CSMA/CD algorithm. This algorithm was simpler than the other two legacy LAN competing technologies, namely the IEEE 802.5 token ring supported by IBM for office areas and the IEEE 802.4 token bus supported by HP and others for a manufacturing environment. In a CSMA/CD MAC protocol, when a terminal wants to send a frame it starts to sense the channel by simply reading the voltage level on the line. If there is no activity on the medium and the medium is idle then it sends its frame. If not, then the terminal waits until the medium becomes ready and after the interframe gap period of 9.6 μs in 10 Mb/s legacy Ethernet it sends the frame. If a collision occurs, then the terminal goes through the collision detection process, stops the transmission of the packets, and uses a back-up algorithm to retransmit the packet with higher probability of success. Figure 8.8

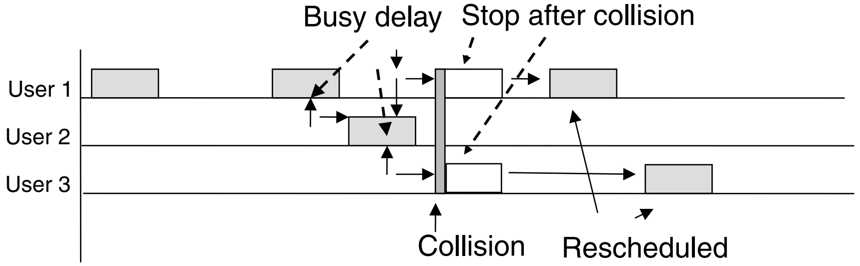


FIGURE 8.8 Principal of operation of the CSMA/CD specified as the MAC layer of the IEEE 802.3 Ethernet.

shows a simple example: User 1 senses the channel and it is idle, so it transmits its first frame; later on it again senses the channel and there is no activity, so it send the second frame. During the transmission of the second frame, User 2 senses the channel and finds it busy, so it waits until the channel becomes available; it prepares itself in 9.6 µs and sends its frame. While User 2 is transmitting its frame, both Users 1 and 3 sense the channel and find it busy. As soon as the medium becomes available, both users send their frames; shortly after they find out that there is a collision because the voltage levels across the cable have gone above the normal values. As soon as the transmitters discover the occurrence of the collision they stop transmission of the rest of the packet, send a 32-bit jamming signal, and reschedule the frame for the next slot. The jamming bits can have any value other than the CRC value of the colliding packet and they ensure that the collision lasts long enough to be detected by all stations on the network. If retransmission is done immediately, then, with unit probability, in the next slot the packets will collide again; to avoid this collision, Ethernet uses an algorithm referred to as binary exponential back-off algorithm.

Legacy Ethernet uses the exponential back-off algorithm, which is based on a random waiting time exponentially increasing with the number of collisions. The unit of waiting time is a slot and that is the minimum length of a packet, so that one can detect the collision before transmission of the entire packet is completed. In legacy Ethernet this slot time is specified as 51.2 ms, which is the time needed for transmission of 512 bits (64 bytes) of data at a rate of 10 Mb/s. Figure 8.9 shows the worst-case scenario for collision report. Two terminals, A and B, are located at two ends of the network with maximum length of 2.5 km. Terminal A sends a frame and when that frame is about to reach terminal B that terminal sends its own packet and a collision occurs. The collision signal must travel the 2.5 km back to terminal A before that terminal knows that a collision has occurred. Therefore, we need 25 ms to detect a collision for the worst-case scenario. Selection of 51.2 ms, which is almost double this value, allows another 25 ms for delay associated with the four repeaters. In other words, the maximum delay for collision detection and, consequently, the minimum length of packet to detect collision is almost twice the round-trip propagation time, and the extra time is needed for delay caused by repeaters.

The exponential back-off timing algorithm of the Ethernet operates based on the probability of transmission of the frame after a given number of collisions. After the first collision the frame is sent in the next time slot with a probability of $^1/_2$, which means we pick from (0,1); after second probability of transmission in the next slot reduces to $^1/_4$, which means we pick from (0,1,2,3), after the third we pick from (0,1,. . .,7) with

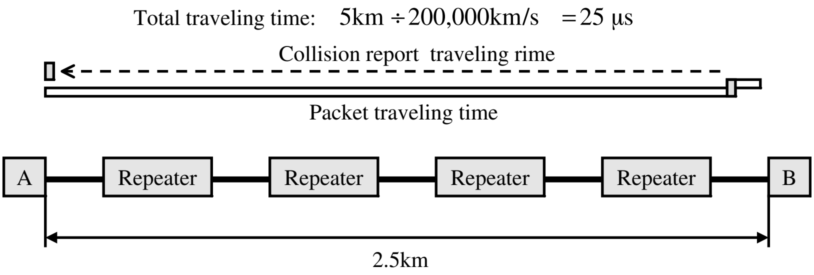

Total traveling time: 5km ÷ 200,000km/s = 25 µs
Collision report traveling rime
Packet traveling time
A
Repeater
Repeater
Repeater
Repeater
B
2.5km

FIGURE 8.9 Collision detection and the length of the packet.

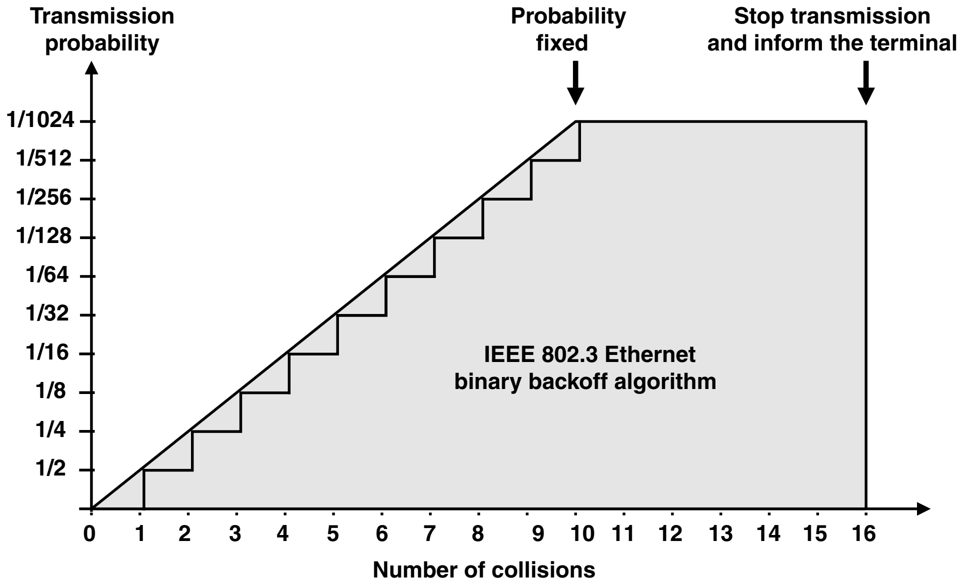


FIGURE 8.10 Back-off timing of the binary exponential back-off algorithm used in the Ethernet.

probability of $^1/_8$, and we continue reducing the probability of transmission after collision by a half until the tenth collision. After 10 collisions we send a packet with the same probability of 1/1024 by repeatedly picking from (0,1,. . .,1023) six more times. After 16 collisions, the MAC reports a failure to the computer and further recovery will remain up to higher layers. Figure 8.10 illustrates the relationship between transmission probability and the number of collision for the binary exponential back-off algorithm employed in the Ethernet. This algorithm automatically prevents instability of the network. Figure 8.11

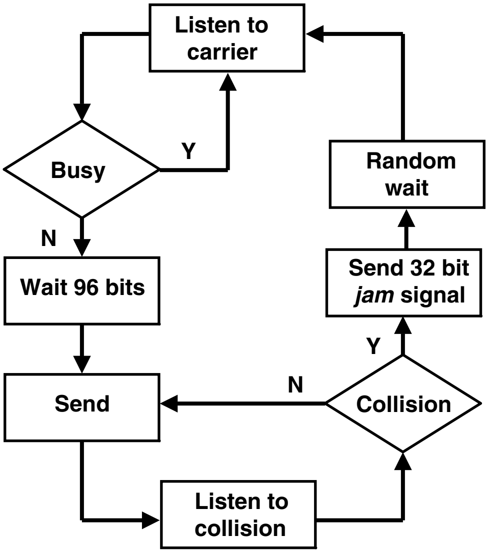


FIGURE 8.11 A flow chart for CSMA/CD operation.

provides a flow chart summarizing the CSMA/CD algorithm used in the Ethernet. The weakness of the exponential back-off algorithm used in IEEE 802.3 is that it is a first in, last out system. If a terminal arrives earlier and experiences a few collisions then it has a lower probability of access to the channel than a terminal which arrived later and has experienced its first collision. As we will see, this problem is resolved in IEEE 802.11.

8.2.3 Medium Access Control Performance

The performance of the MAC can be measured by calculation of the efficiency or channel utilization of the protocol. In general, channel efficiency or utilization for average packet lengths of T_{P} and the average idle time of T_{I} is given by

$$\text{Channel Efficiency (Utilization)} = \frac{T_{\mathrm{P}}}{T_{\mathrm{P}} + T_{\mathrm{I}}} \tag{8.1}$$

This value reflects the efficiency of the MAC protocol in utilizing the PHY resources. Although there have been numerous rigorous calculations of the throughput of the CSMA/CD for a finite and an infinite number of users, we resort to a relatively simple calculation presented in [Met76, Tan03]. This analysis is based on a fixed probability of contention p for N terminals contending to access a slot in the medium. Consider one terminal; for that particular terminal the probability of successful transmission in the slot is given by $p(1-p)^{N-1}$, which implies that the specific terminal is transmitting and the other $N-1$ terminals are not. Since we have N terminals each with probability $p(1-p)^{N-1}$ of occupying the slot, the probability of a packet from all users occupying the slot, which is actually the probability of successful transmission in the slot A, would be given by

$$A(N,p) = Np(1-p)^{N-1} \tag{8.2}$$

As $N \rightarrow \infty$ the function $A(N, p)$ maximizes for $p = 1/N$, for which

$$A(N,p) = \left(1-\frac{1}{N}\right)^{N-1}\Bigg|_{N\rightarrow\infty} = \frac{1}{\mathrm{e}}$$

As shown in Fig. 8.12, given the probability of successful transmission in a slot is A and the probability of having an idle slot is $1-A$, then one can determine the probability of successful transmission at the kth slots:

$$P(k) = A(1-A)^{k-1} \tag{8.3}$$

This is indeed the probability of having *k-slot* contention, from which we can calculate the average number of slots of waiting idle before successful transmission of a packet. Using the PDF of the contention slots we can calculate the average number of

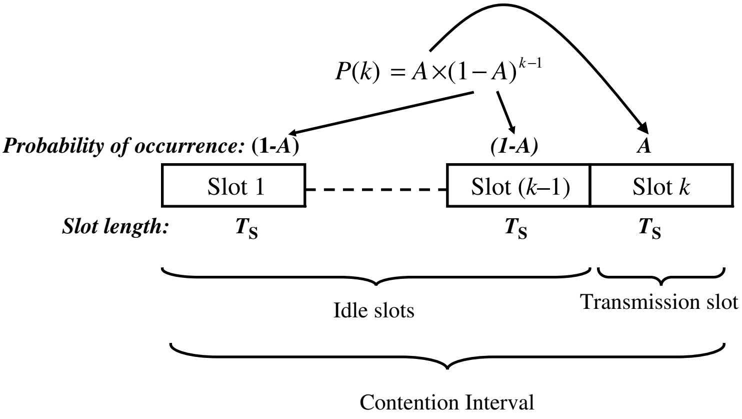


FIGURE 8.12 Contention interval and frame transmission.

slots per contention:[1]

$$\begin{aligned} E\{k\} &= \sum_{k=1}^{\infty} kP(k) = \sum_{k=1}^{\infty} kA(1-A)^{k-1} \\ &= A + 2A(1-A) + 3A(1+A) + \ldots = \frac{1}{A} \end{aligned} \tag{8.4}$$

If we assume the length of the slot is T_S, then the idle time is calculated from

$$T_I = \frac{T_S}{A} - T_S = T_S\frac{1-A}{A}$$

If we consider the bounds as the number of users approaches infinity, $A \to 1/\mathrm{e}$, and the average idle time is

$$T_I = T_S\frac{1-A}{A} < T_S(\mathrm{e}-1) = 1.72T_S \tag{8.5}$$

Substituting in Eq. (8.1) we have

$$\text{Channel Efficiency(Utilization)} = \frac{T_P}{T_P + \dfrac{T_S(1-A)}{A}} < \frac{T_P}{T_P + 1.72T_S} \tag{8.6}$$

[1] Note that $\sum_{i=0}^{\infty} ia^i = a/(1-a)^2$.

As shown in Fig. 8.6, each MAC packet which is longer than minimum requirement (no pads) has 26 bytes of additional overhead plus 12 bytes (96 bits) of interframe gap. In addition, the length of the slot is 64 bytes (512 bits). Therefore, the actual throughput can be calculated from

$$\text{Channel Efficiency (Utilization)} = \frac{L_D + 38}{L_D + 38 + \dfrac{64(1-A)}{A}} < \frac{L_D + 38}{L_D + 148} \qquad (8.7)$$

where

$$A = \left(1 - \frac{1}{N}\right)^{N-1}\Bigg|_{N\to\infty} = \frac{1}{e}$$

and L_D is the length of the data packet in bytes. Figure 8.13 shows the channel utilization of the legacy Ethernet for different lengths of data packets. For large packet lengths close to the maximum allowed length the efficiency is close to 90%, and for short packets it reduces approximately three times.

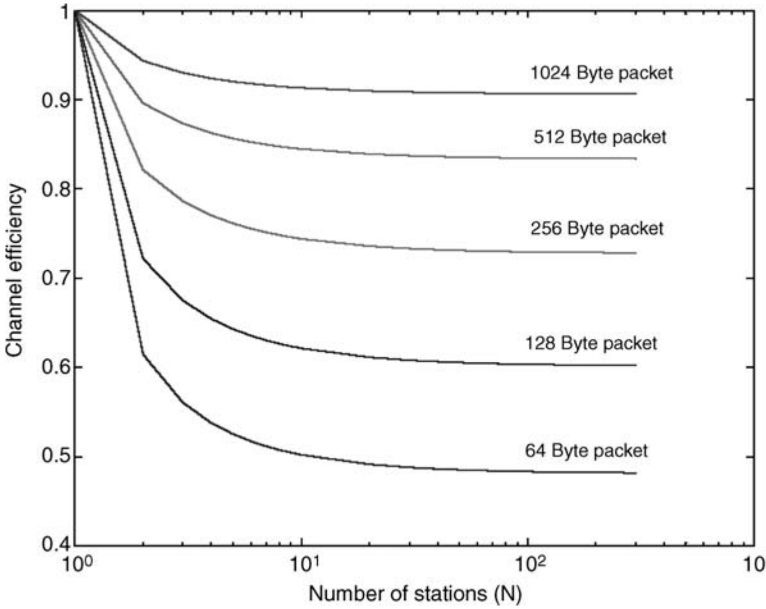


FIGURE 8.13 Channel utilization as a function of number of users.

8.2.4 Alternatives to Legacy Ethernet

In the early 1970s the main competitor of the Ethernet supported by Xerox, DEC, and Intel was the token ring introduced by IBM. This technology later became the IEEE 802.5 standard and was in the market up to a few years ago. Similar to Ethernet, the IEEE 802.5 token ring was intended for commercial and light industry. Figure 8.14*a* shows the basic topology of a token ring LAN; this approach provides for a collection of point-to-points rather than a broadcast medium. The operation is based on a token packet which is circulating in the network. When a packet arrives at a terminal buffer, the terminal captures the token to prevent others from transmission and puts its own frame in the network. When the frame travels along the ring, the terminal which matches the destination address of the frame copies the frame but it lets it continue its circulation around the ring. When the frame arrives at the source terminal it is captured from circulation and the token is released in the ring to allow others to have access to the medium. This approach for medium access provides for an efficient and fair MAC protocol under heavy load conditions. The physical layer was fully digital and, unlike Ethernet, it did not get involved in an analog collision detection mechanism. A monitor station oversees the operation by handling token loss using a timer, cleaning the invalid garbled frames, and keeping the length of the ring delay longer than the token time. The first terminal that is connected to the network sends a claim for a token frame; if there is no other monitor terminal then it resumes the rule of the monitor station. The IEEE 802.5 packets have a field reserved for applying priority which was not available with Ethernet.

In the early days of legacy Ethernet it was perceived that it was suitable for offices, but there were reservations with use in factories because the delay could be arbitrarily long, there was no priority assignment, and taking turns, like with token ring, appeared more suitable for factory applications. The token ring, however, was also perceived not to address issues related to manufacturing, because a break in a cable or a monitoring station would bring the network down. Therefore, people were thinking of a network without a central control station and no physical ring operation. The solution was the token bus topology shown in Fig. 8.14*b*. Unlike token ring, the token bus does not have a *monitor*

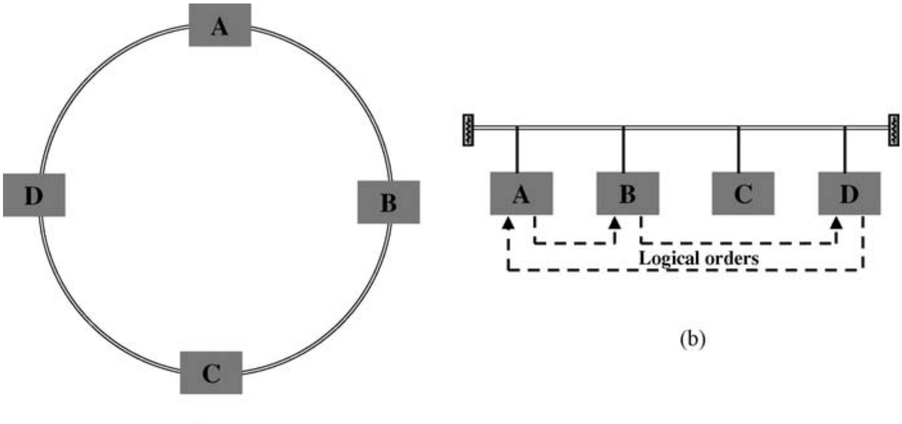


FIGURE 8.14 (*a*) Token ring and (*b*) token bus topologies used in the IEEE 802.5 and IEEE 802.4 in the early days of the LAN industry.

station to oversee the operation and the connections are logical. The end of the cable here is connected to a terminator rather than to the other end of the cable. In this topology, certain terminals, such as a printer, shown as terminal C in Fig. 8.14*b*, can only listen. Similar to IEEE 802.5, IEEE 802.4 also provided for a priority scheme. The priority schemes for IEEE 802.4 and 5 are eliminated upon connection to an Ethernet through a LAN bridge.

8.2.5 Early Enhancements to Legacy Ethernet

As we showed in Fig. 1.9, the legacy 10Base5 Ethernet soon evolved into thin-cable Ethernet, also referred to as "cheapernet" with a coverage of 185 m, which is also referred to as 10Base2 in which the 2 reflects the approximately 200 m coverage. IEEE 802.3a specified the thin-cable medium for 10Base2 in 1985. This adjustment in the medium would facilitate installation and reduce the cost of the cable, resulting in a cheaper solution facilitating rapid growth of LAN installations in educational institutions and mid-sized industrial institutions nationwide. In the mid 1980s, most mid-sized universities started installing 10Base2 to connect computer terminals inside a department and a backbone FDDI at 100 Mb/s to connect the department LANs. The cost of wiring for these networks would exceed several million dollars and the industry quickly exceeded several billion dollars, stimulating further growth of the industry and nurturing new technical innovations to support this growth. Another advantage of cheapernet was that it used BNC connectors, which were more rugged than vampire connectors; and this way, using a power splitter, the BNC cable would hook to the Ethernet. The BNC connector, originally designed to carry RF signals, shown in Fig. 8.15, had a longer life than vampire connectors, reducing the maintenance cost of the network.

After moving from a few giant companies, such as IBM, Intel, DEC, and Xerox, and a few major research universities to mid-sized industry and academic buildings the LAN industry aimed at smaller buildings for small businesses for which installation of new wiring would pose a cost and speed factor. Twisted-pair wiring used for telephone networks attracted attention for these environments. Installation of the twisted-pair wiring was simple and inexpensive, and telephone companies had the expertise and the crews to install them efficiently at a low cost. The first implementation of the Ethernet over

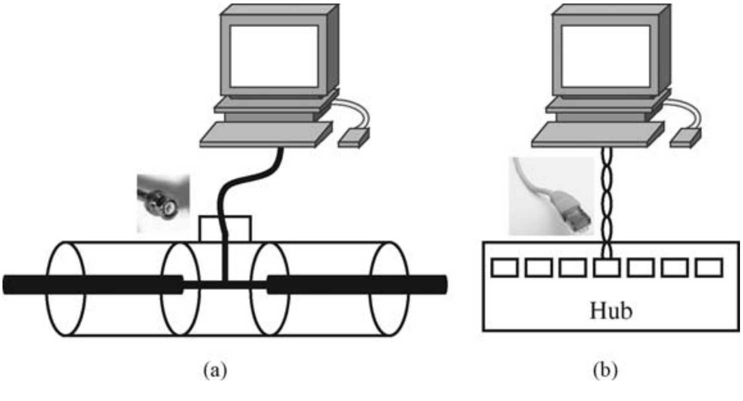


FIGURE 8.15 (*a*) BNC connectors used in the 10Base2 cheapernet; (*b*) RJ-45 connectors used in the 10Base-T hub-and-spoke Ethernet.

twisted-pair telephone wiring was StarLAN. Developed in the mid 1980s as 1Base5 later on formed the basis for the 10Base-T, which was defined by IEEE 802.3i in 1990. The letter "T" in the acronym refers to twisted pair and the maximum length of the line was 100 m. As shown in Fig. 1.9, the topology of this new revolutionary technology of its time was star, which was used in the telephone network for many years; for that reason it needed a hub to connect the computer terminals with one another. In the telephone wiring we simply connect the wires together, but in the Ethernet the medium is also used for collision detection. Therefore, the MAC protocol needed an adjustment. This architecture was also referred to as hub-and-spoke, because a terminal goes to the hub before it communicates with other terminals. This architecture is in contrast with the traditional LBT architecture of the legacy Ethernet. In the hub-and-spoke architecture the hub detects the collision among packets arriving from different ports and transmits the jam signal into all ports. Collision is detected when the hub senses a signal at two different ports at the same time and the detection mechanism does not involve the analog threshold settings used in NIC cards of the legacy Ethernet. Connectors used for 10Base-T were the telephone-like RJ-45 connectors shown in Fig. 8.15*b*.

Hub-and-spoke architecture and RJ-45-like connectors replaced the legacy Ethernet bus and its vampire and BNC connectors, leaving a seal on the evolution of the topology and cable connectors for the popular Ethernet. Since the early 1990s the attention of the Ethernet community has shifted towards higher speeds using more complex transmission technologies and a more diversified transmission medium with some modifications in the use of frame format and CSMA/CD operation. Figure 8.16 summarizes the major steps in the

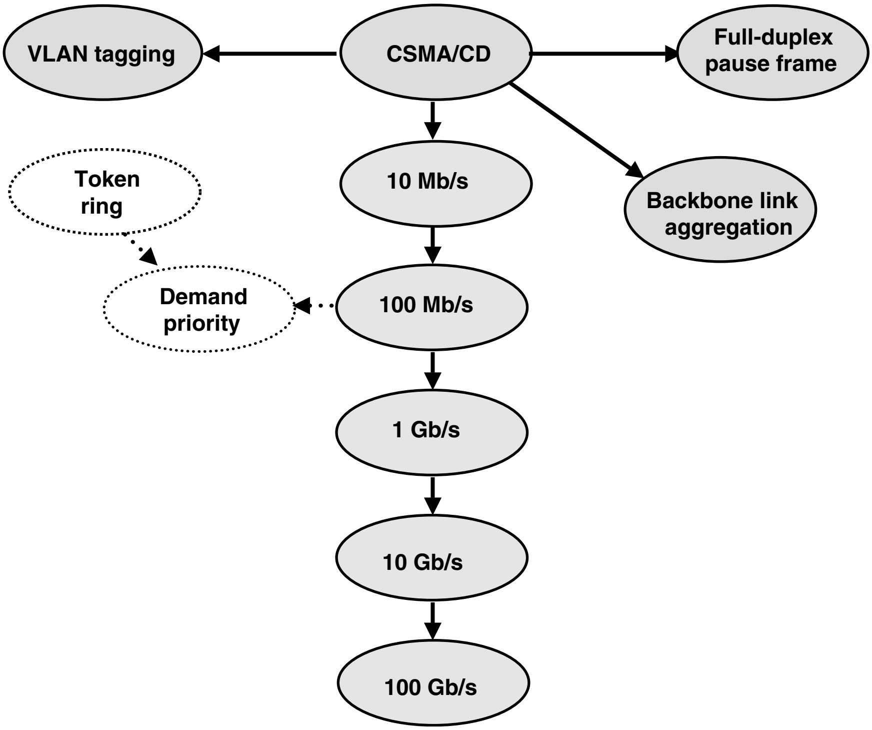


FIGURE 8.16 Major steps in evolution of the MAC and PHY for the Ethernet.

evolution of the Ethernet. In the mid 1990s the data rates moved to 100 Mb/s by introducing the so called "fast Ethernet," allowing Ethernet to assimilate FDDI technology. Later, GIGABIT Ethernet followed by 10 Gb/s and more recently by 100 Gb/s Ethernet started penetrating WAN wirings traditionally dominated by T-carrier and SONET's OC line discussed in Chapter 3. In the frame format and medium access, VLAN was introduced to partition the network, full-duplex operation became possible, eliminating the use of CSMA/CD operation, and link aggregation allowed more flexibility for transmission to find more cost-efficient solutions. Table 8.2 highlights the important IEEE 802.3 standards born out of the evolution of the Ethernet. In the next two sections we discuss more details of the evolution of the physical layer in the Ethernet, followed by the evolution of the frame and the access.

8.3 EVOLUTION OF THE PHYSICAL LAYER

In the early 1990s, 10 Mb/s Ethernet technology and the hub-and-spoke architecture had dominated the market, while FDDI was used for 100 Mb/s operation to connect these LANs, as well as its closest competitor at that time the IEEE 802.5 token ring. The first Ethernet advancement towards higher data rates was the so-called "fast Ethernet," which operated at 100 Mb/s and was a series of standards defined by IEEE 802.3u in 1995 for a variety of fiber and twisted-pair media. This was followed by IEEE 802.3z, defining the gigabit Ethernet for fiber and STP mediums in 1998, followed by IEEE 802.3ab, which extended the gigabit per second Ethernet to the Cat-5 twisted-pair medium. In 2003, IEEE 802.3ae introduced 10 Gb/s Ethernet for long-distance fiber operations, which was extended to wire media by IEEE 802.3ak/an in 2004 and 2006 respectively. The advancements are summarized in Table 8.2. Currently, IEEE 802.3ba is working on the 100 Gb/s Ethernet [IEEE07]. Evolution of these standards reflects the penetration of Ethernet technology into WAN applications.

In principle, increasing the data rate is based on improving the medium, explained in Chapter 3, or by using more sophisticated transmission technologies, as introduced in Chapter 2. To carry higher data rates over a medium, one can improve the bandwidth of the wire or the fiber, increase the number of wires, and shorten the length of the wire. To increase the data rate using transmission techniques, one can use more efficient line coding (rather than Manchester coding), use multilevel and multidimensional transmission techniques, use signal processing techniques (such as equalization, echo cancellation), and use more effective channel coding techniques (such as scramblers, convolutional coding, and TCM). As we will see in the following sections, these techniques have been adopted by a variety of Ethernet standards in one way or another. Therefore, a review of the evolution of physical layer and medium options of Ethernet provides an excellent overview of applied digital communications and information theory.

8.3.1 Fast Ethernet at 100 Mb/s

Figure 8.17 provides an overview of the 100Base-T series of IEEE 802.3T standards for fast Ethernet. There are 100Base-X standards which actually are based on the FDDI technology using token ring, which is adapted to the Ethernet environment using the CSMA/CD medium access protocol. FDDI was originally designed for fiber operation and later was extended to cable operation, which was also sometimes referred to as CDDI, replacing fiber

TABLE 8.2 Overview of the Important IEEE 802.3 Standard Series

1973	The first experiment at 2.94 Mb/s
1980	DEC, Intel, and Xerox released 10Base5
1983	IEEE released the first IEEE 802.3 standard for Ethernet technology
1985	IEEE 802.3a defined “thin” Ethernet, “cheapernet,” or 10Base2
1985	IEEE 802.3b defined 10Broad36 for “broadband” cable systems
1987	IEEE 802.3d defined 2-Fiber Optic Inter-Repeater Link (FOIRL) for up to 1 km
1990	IEEE 802.3i defined 10Base-T
1993	IEEE 802.3j defined 10Base-F (FP, FB, & FL) for up to 2 km
1995	IEEE 802.3u defined 100Base-T (Fast Ethernet)
1997	IEEE 802.3x defined “full-duplex” Ethernet operation (no CSMA/CD and 20 Mb/s/200 Mb/s total)
1997	IEEE 802.3y defined 100Base-T2 for two pairs of Category 3
1998	IEEE 802.3z defined 1000Base-X “gigabit Ethernet” using fiber and STP
1998	IEEE 802.3ac defined VLAN tagging on Ethernet networks
1999	IEEE 802.3ab 1000Base-T standard defined 1 Gb/s operation over four pairs of Category 5 UTP cabling
2000	IEEE 802.3ad, link aggregation allowing several NICs in a computer to connect to one port of a switch
2003	IEEE 802.3ae, 10 Gb/s Ethernet over fiber for connecting routers and switches; although called LAN, this can go up to 80 km
2003	IEEE 802.3af, power over Ethernet for IP phones, webcams, WLAN APs, etc.
2004	IEEE 802.3ah, one mile point-to-multi1000Base technology for home/office access
2004	IEEE 802.3ak, 10GBASE-CX4, works on 4 copper media up to 15 m
2006	IEEE 802.3an, 10GBASE-T, for UTP
2008	IEEE 802.3ba, 100 gigabit Ethernet (?)

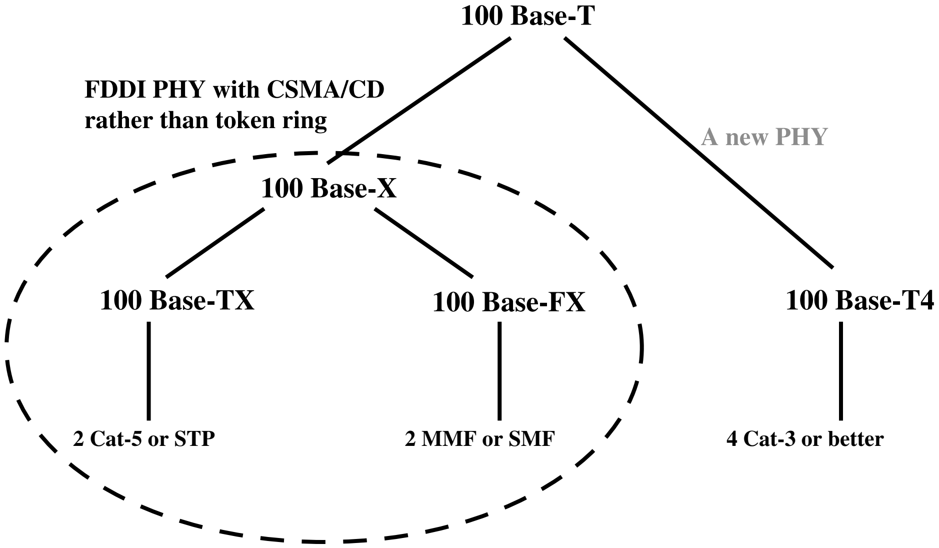


FIGURE 8.17 Overview of 100 Base-T fast Ethernet options.

TABLE 8.3 4B/5B Encoding Technique Originally for FDDI and then Used in 100Base-X

Information	Coded block	Information	Coded block
0000	11110	1011	10111
0001	01001	1100	11010
0010	10100	1101	11011
0011	10101	1110	11100
0100	01010	1111	11101
0101	01011	Idle	11111
0110	01110	Start delimiter	11000
0111	01111	Start delimiter	10001
1000	10010	End delimiter	01101
1001	10011	End delimiter	00111
1010	10110	Transmit error	00100

“F” with copper “C,” and it was very popular for connecting legacy Ethernet LANs. The 100Base-X standard would allow integration of the existing FDDI wirings into a simpler and growing Ethernet environment. The other branch is a newly defined PHY and medium using four pairs of Cat-3 or better wires, 100Base-T4, which turned out to become the most popular Ethernet used everywhere.

100Base-X. Both fiber and copper medium specifications for the 100Base-X series use the so-called 4B/5B line-coding technique. The incoming data stream is formed into 4-bit blocks of information with 16 different possibilities. Each 4-bit block is mapped to one of the 32 possible 5-bit blocks so that the coded block has at least two 1s in the five bits. If the line-coding technique used with the 4B/5B coding codes a “1” into transition in the level of the transmitted signal, then these 1s can support synchronization at the receiver. Table 8.3 shows all 16 possible codes and their related 5-bit codes. The 5-bit output of the coding has 32 combinations. As only 16 of them are used for the incoming data, two of these extra codes are used for starting delimiters, two for ending delimiters, and one for reporting error; the remaining 11 codes are not used [Sta00]. The coding rate of this line coding is $^4/_5$, compared with $^1/_2$ for the Manchester line coding. Figures 8.18 and 8.19 show the details of implementation of 100Base-FX and 100Base-X respectively.

The 100Base-X fast Ethernets also support full-duplex operation. In traditional half-duplex Ethernet the NIC only sends in one direction; in full-duplex, an NIC can transmit and receive at the same time, doubling the aggregated rate of information of the LAN. Full-duplex operation curbs the CSMA/CD protocol because transmission does not need to sense the availability of the medium. The problem with this approach is the receiver buffer may overflow and we need a method to stop transmission if the buffer of the receiver is full. We discuss the details of how it operates later on in this chapter when we address modifications to the legacy Ethernet packet format. Another feature introduced with the fast Ethernet standard is autonegotiation.

The autonegotiation protocol allows two connected devices to choose common transmission modes, such as speed and duplex operation. Using the autonegotiation protocol, devices first share their capabilities and then choose the fastest transmission mode they both support.

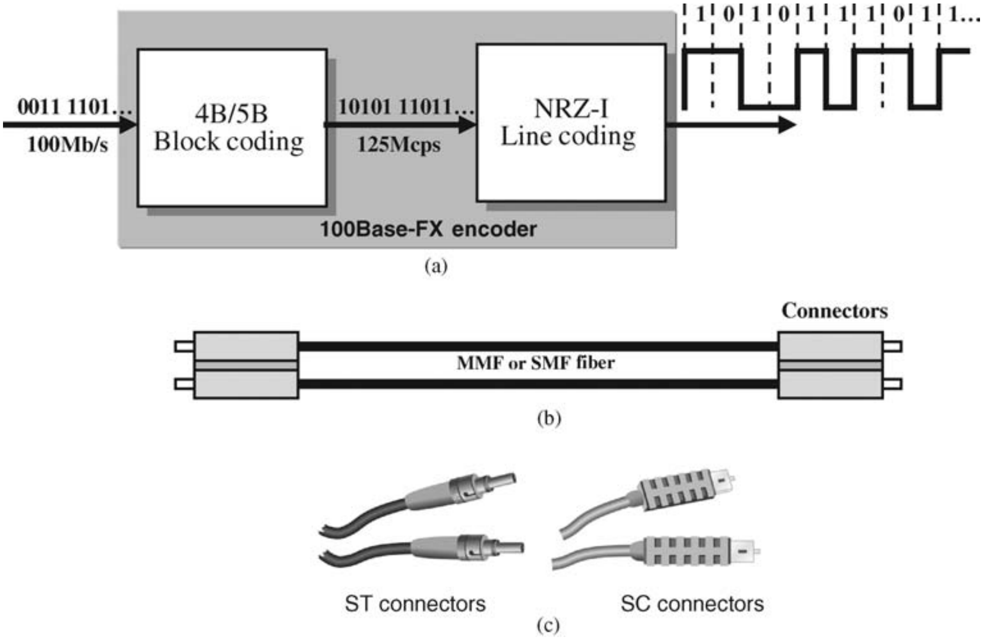


FIGURE 8.18 Overall operation of 100Base-XF: (*a*) 4B/5B and NRZ-I transmission; (*b*) cable and connectors overview; (*c*) two types of connector.

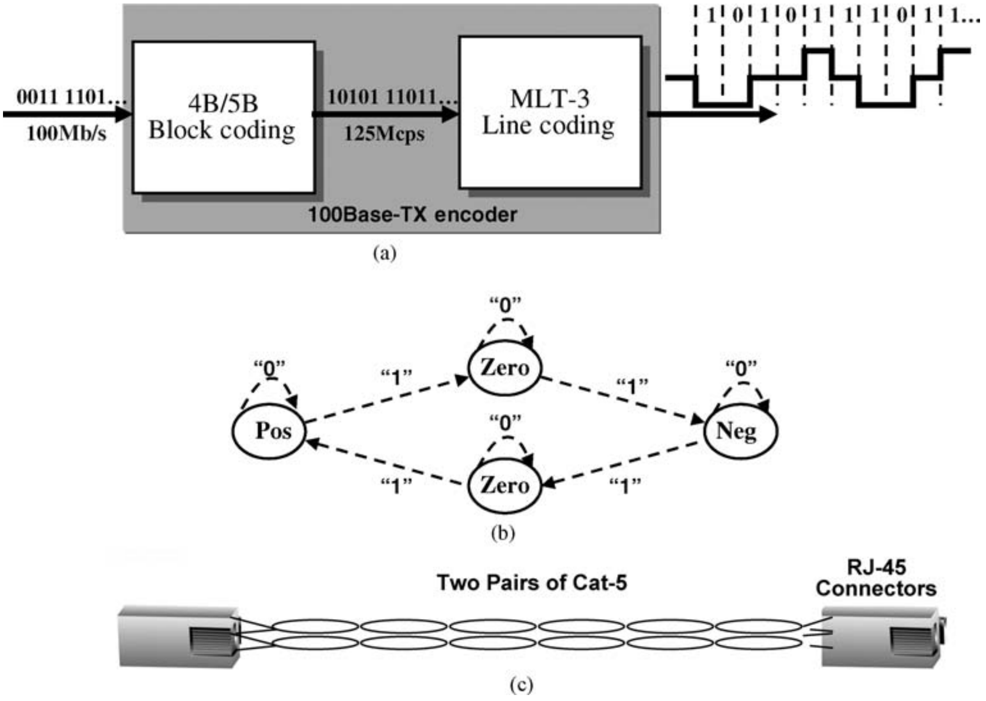


FIGURE 8.19 Overall operation of 100Base-TX: (*a*) 4B/5B and MLT-3 transmission; (*b*) state diagram of the MLT-3 line coding; (*c*) physical connection.

100Base-FX. This uses the resulting 5-bit coded block and then uses NRZ-I line coding technique to create the transmitted waveform for the fiber line. As we explained in Chapter 3, for the NRZ-I line coding, each 1 is coded to a transition at the beginning of the bit. This way, we have at least two transitions every transmitted five, which reflects the fact that 80% of the transmitted bits carry information bits, while in Manchester coding only 50% of the transmitted bits carry information. Figure 8.18 gives an overall perspective of how this coding is implemented. Figure 8.18*a* shows the 4B/5B coding and how it is mapped into NRZ-I line coding to produce the transmitted waveform. Figure 8.18*b*,*c* shows the cables and two types of connector used in 100Base-XF. The maximum length of the cable recommended by the standard using multimode fiber (MMF) with 62.5/125 diameters at 1300 ns is specified at 412 m. Longer lengths are achievable using single-mode fiber (SMF).

100Base-TX. This is designed for 100 ms high-quality Cat-5 or better twisted-pair wiring, which was also used as the medium for the copper or cable version of FDDI. Figure 8.19 gives the overall structure details of 100Base-TX. Figure 8.19*a* shows the coding and transmission; the 100 Mb/s data is first coded into a 125 Mb/s coded stream which is then mapped into a three-level line signal referred to as MLT-3. The MLT-3 line coder shown by the state diagram in Fig. 8.19*b* basically maps each 1 into a transmission in the transmitted waveform. Each time that a 1 arrives, the signal level climbs the three-level signal ladder once. It climbs until there is no more place to go up; then it starts to climb down until there is no more step to switch the direction. The advantage of three-level MLT-3 line coding over the two-level NRZ-I is that it has three points rather than two in the constellation, which results in a longer average distance between the points for the same average transmitted power. As we discussed in Chapter 3, this addition of the distance for the transmitted symbols results in a reduction in the error rate of the transmitted bits. Although MLT-3 has a better performance, it cannot be implemented on fiber because signaling lights in the fiber transmission only take two levels, representing the light in an "on" or "off" position. Figure 8.19*c* shows the two twisted pairs and the RJ-45 connectors used for the physical operation of the 100Base-TX.

Example 8.1: Figure 8.20 shows the signal constellation for transmission of the two-level signal used in 100Base-FX and the three-level signal used in 100Base-TX. In 100Base-TX, the points in the corners of the constellation represent the positive and negative voltage levels and the middle points are associated with the zero-level transmission. Since we have

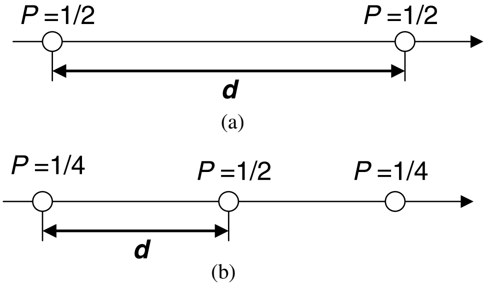


FIGURE 8.20 Signal constellation for transmitted symbols in (*a*) 100Base-FX and (*b*) 100Base-TX.

two zero-level states, the probability of occurrence of a zero-level signal ($^1/_2$) is twice that of the corner points ($^1/_4$). The average energy in the constellation shown in Fig. 8.20*a* is

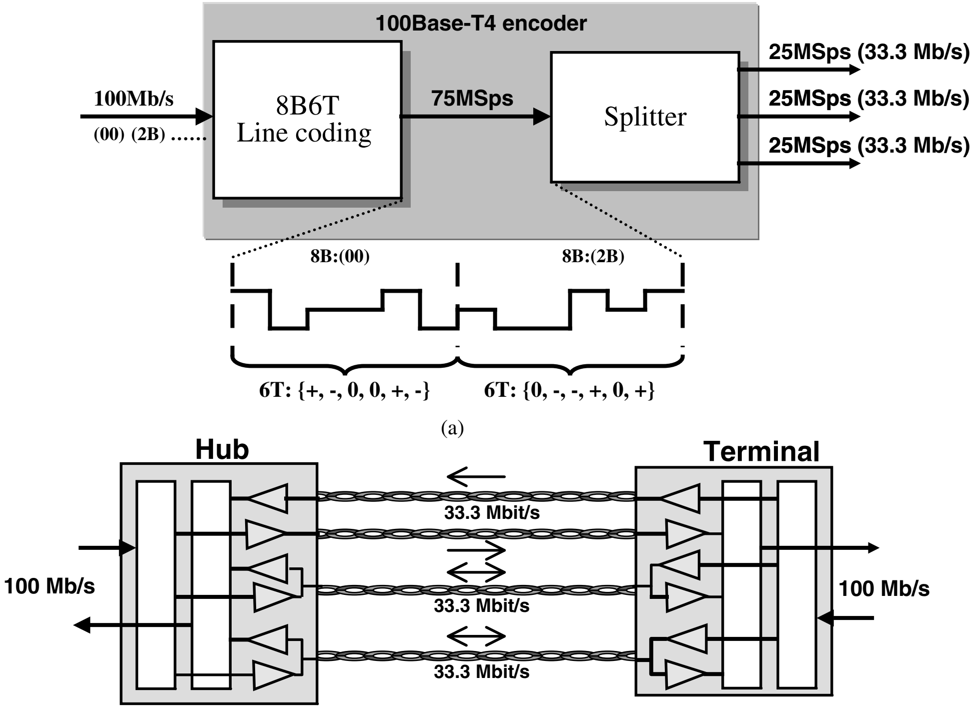
$$\bar{E} = \frac{E_0 + E_1}{2} = \frac{(d/2)^2 + (d/2)^2}{2} = \frac{d^2}{4}$$

The average energy of the constellation shown in Fig. 8.20*b* is

$$\bar{E} = \frac{E_1}{4} + \frac{E_0}{2} + \frac{E_{-1}}{4} = \frac{d^2}{4} + \frac{0}{2} + \frac{d^2}{4} = \frac{d^2}{2}$$

Therefore, the three-level transmission has two times, or 3 dB, advantage over the two-level transmission technique, making it a more desirable technique.

100Base-T4. This supports a 100 Mb/s data rate for 100 m over four pairs of Cat-3 or better twisted-pair cabling and it is the most popular fast Ethernet commonly used to connect computer terminals in offices and homes. The main motivation for creation of this standard was to allow 100 Mb/s Ethernet to be carried over inexpensive telephone-grade Cat-3 cabling as opposed to the Cat-5 cabling required by 100Base-TX. However, in practice, all new installations today use Cat-5 because the cost of labor for installation is much higher than the cost of wire, justifying the difference between the price of Cat-3 and Cat-5. Figure 8.21 shows the details for the overall operation of 100Base-T4. As shown in



(a)

(b)

FIGURE 8.21 Overall operation of 100Base-T4: (*a*) 8B6T transmission; (*b*) connections for four pairs of wires.

TABLE 8.4 Samples of 8B6T Code Words Used in 100Base-T4. The 8-bit Data Block is Represented in Octet Form

Data block		Code word
(00)	→	(+,−,0,0,+,−)
(0A)	→	(−,+,0,+,−,0)
(2B)	→	(0,−,−,+,0,+)
(2E)	→	(−,0,−,0,+,+)
(32)	→	(+,−,0,−,+,0)
(3C)	→	(+,0,−,0,−,+)
(3F)	→	(+,0,−,+,0,−)
(B8)	→	(−,+,0,0,+,0)
(AB)	→	(+,−,+,−,−,+)
(EB)	→	(+,−,+,−,0,+)

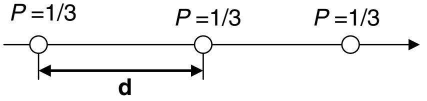


FIGURE 8.22 Signal constellation for transmitted symbols in 100Base-T4.

Fig. 8.21*a*, the input data stream is first encoded using the "8B6T" encoding scheme, in which 8 bits of binary data are converted into six "ternary" signals. For eight input bits we need $2^8 = 256$ symbols, and with six ternary digits we can make $3^6 = 729$ symbols. Therefore, only a portion of the six ternary symbols are used to map the 8-bit data block to the six ternary symbols. The ternary symbols are selected so that the probability of occurrence of all three symbols is the same. Table 8.4 shows samples of the 8B6T codes; the complete set is available in Patel *et al.* [Pat96]. The ternary signals have three values, i.e. $-$, 0, and $+$, resulting in a PAM-3 signal constellation.

Example 8.2: The PHY Layer of 100Base-T4 Figure 8.22 shows the signal constellation for the transmission three-level signal used in 100Base-T4, assuming that all three levels have the same probability of transmission.[2] The average energy in the constellation shown in Fig. 8.22 is

$$\bar{E} = \frac{E_1}{3} + \frac{E_0}{3} + \frac{E_{-1}}{3} = \frac{d^2}{3} + \frac{0}{3} + \frac{d^2}{3} = \frac{2d^2}{3}$$

Compared with the binary constellation used in 100Base-TX shown in Fig. 8.20*b*, the constellation in 100Base-T4 has 1.33 times, or 1.25 dB, advantage.

Since each 8 bits is transmitted with six ternary symbols, the symbol transmission rate of the ternary signal is

$$100\,\text{Mb/s} \times \frac{6\ (\text{ternary symbols})}{8\ (\text{bits})} = 75\,\text{MS/s}$$

[2] This is a valid assumption because the number of three levels used in the code word table is approximately the same.

The 75 MS/s stream of ternary symbols is then splinted into three streams, shown in Fig. 8.21*a*, each carrying 25 MS/s of ternary data. The actual binary data stream carried with each of these three lines is

$$25\,\mathrm{MS/s} \times \frac{8\ (\text{bits})}{6\ (\text{ternary symbols})} = 33.3\,\mathrm{Mb/s}$$

This scheme effectively splits the 100 Mb/s data rate over three streams of 33.3 Mb/s so that it can carry each of these streams over Cat-3 twisted-pair wiring, which has a bandwidth capable of carrying 25 MS/s. Cat-3 wiring carries 20 MS/s for the 10Base-T and its maximum capacity is 32 MS/s. If binary signaling was used instead of ternary, then each line should have carried 33.3 MS/s, which was beyond the capability of the Cat-3 wiring.

The designers of the 100Base-T4 map these three streams of data to the four pairs of twisted-pair wires inside a bundle using the format shown in Fig. 8.21*b*. Two pairs are used for full-duplex communication mode, carrying data only in one direction, and the other two for half-duplex communication mode, carrying data in both directions. This scheme provides an efficient method for utilizing the transmission capabilities of the four pairs of wires inside a bundle to support both full-duplex and half-duplex data transmission. In practice, however, 100Base-T4 is always used in half-duplex mode and it cannot support full-duplex operation. Similar to 100Base-TX, 100Base-T4 also uses telephone-jack-like click-on RJ-45 connectors.

8.3.2 Alternative for Fast Ethernet

There are four other LAN technologies with data rates around 100 Mb/s: FDDI, 100VG-ANYLAN, ATM LANE, and 100Base-T2. Today, these technologies are dominated by the fast Ethernet, and in particular the 100Base-T4 technology. A brief overview of these technologies provides the reader with an insight into the evolutionary steps of wired LANs and networking industry.

FDDI. This was written by ANSI, rather than IEEE 802, and was approved by ISO/IEC in 1990. The original goal was to provide a 100 Mb/s fiber to every desktop. High cost and installation difficulties prevented this market developing; later, around the 1990s, because of high bandwidth and long coverage, it was used as the backbone to connect 10 Mb/s Ethernets. The UTP physical layer of the FDDI was adopted by ANSI in 1995 and it is sometimes referred to as CDDI. With the popularity of Ethernet technologies, FDDI and CDDI wiring are now used by different Ethernet technologies such as 100BASE-T or gigabit Ethernet, and this technology has become redundant. FDDI uses token ring topology, but handling of the token is different from IEEE 802.5. In FDDI the token is released after the packet is transmitted, while in IEEE 802.5 the token is released after the packet has returned to the terminal. The coverage of FDDI can range up to 200 km, and thousands of terminals can connect to the networks.

100VG-ANYLAN. Recommended by IEEE 802.12, this is another 100 Mb/s Ethernet standard over four pairs of voice-grade (VG) Cat-3 wiring each carrying 25 Mb/s using 5B6B coding. The incoming 100 Mb/s data stream is divided into four streams each carrying

25 Mb/s. Each of the 25 Mb/s streams is 5B6B coded to ensure adequate transitions for synchronization. The resulting 25 Mb/s × 6 (bits)/5 (bits) = 30 Mb/s coded streams are carried with one of the four pairs of Cat-3 twisted pairs in the bundle. The 30 MS/s transmission is still less than the 32 MS/s capacity of Cat-5. Comparing the physical layer of 100VG-ANYLAN with the physical layer of 100Base-T4, this is a more balanced and simpler solution which does not attempt to have a useless partial full-duplex operation. This LAN technology has a new MAC with polling architecture which is called demand priority access method (DPAM) and supports either 802.3 or 802.5 (any-LAN), but not both of them at the same time. The four wires in the DPAM network connecting the data terminal node to the hub carry an idle signal: two wires from the node to the hub and two from the hub to the node. When the node or a hub has a packet to send, a request signal is sent to the other end using its two pairs of wires, and after a short silence time the end with the data packet uses all four pairs of wire to send its frame. After completion of the frame, the transmission lines go to their original idle states. This standard was ratified by the ISO in 1995, but it became practically extinct in 1998.

LANE. This was a standard defined by the ATM Forum in the early 1990s to emulate the operation of traditional LANs such as Ethernet or token ring over an ATM network. In that time frame, several major telecom companies perceived that ATM would be the end-to-end solution for all networks and that LANs had to comply with that; for that reason, LANE technology had attracted attention. They assumed ATM-LANs, which could support star topology with 155 Mb/s (over fiber) and 51 or 25 Mb/s (over UTP), would be the next generation of legacy 10 Mb/s LANs which could easily integrate with the ATM WANs. ATM WANs specified 155 Mb/s, 622 Mb/s, 1.2 Gb/s, and 2.4 Gb/s as the backbone, and with those data rates at the backbone those telecom companies started to look into a temporary solution to integrate existing Ethernet and token ring LANs into the ATM world. LANE technology was expected to implement this environment. The LANE protocol defines a service interface for network layer protocols that is identical to the one in traditional LANs, so that the data sent across the ATM network is packed in the appropriate LAN MAC packet format. LANE did not attempt to emulate the actual MAC protocol of the LAN, and it did not require any modifications to higher ayer protocols to enable their operation over an ATM network.

In order to develop an emulated LAN which satisfies the above objectives, it was necessary to settle the differences between the Internet/LAN protocols and the ATM networks. LANs are connectionless datagram services, while ATM is a virtual connection-oriented network demanding the establishment of a connection before data transfer. In addition, LANs are broadcast and multicast environments, allowing every frame sent on the medium to be received by every terminal. LAN MAC addresses are based on manufacturing serial numbers, which are burned into the LAN adapter card during manufacturing, which makes the addresses independent of the network's geographic location. The ATM addresses are associated with the location of connection; and if a terminal is moved from one point to another, then the address is changed. Any LANE system will need to use the real LAN addresses for some functions; therefore, there is a need for a database that allows mapping from LAN addresses to ATM addresses. To settle these differences, the LANE standard defined an architecture with three proxy servers: one to keep the address of local entities and two local proxies, one for call establishment, and the other for handling broadcast and multicast in an ATM network. LANs will connect to bridges with

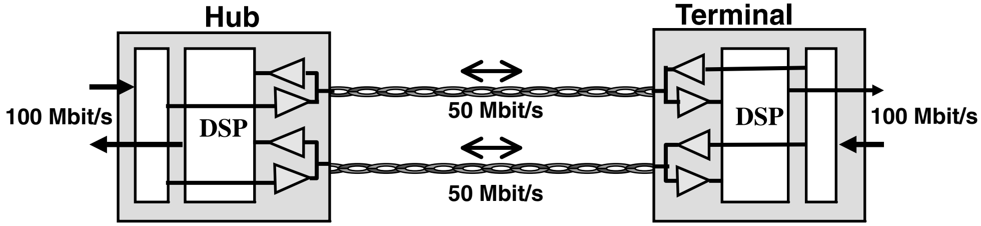


FIGURE 8.23 Overall operation of 100Base-T2.

LANE software to convert the LAN packets to ATM packets. Computers use LANE to map the LLC data to the ATM stack protocol. More details of ATM are discussed in Chapter 6, where we explain switching techniques.

100Base-T2. This standard was prepared by IEEE 802.3y in 1997 and it supports a 100 Mb/s transmission rate over two pairs of VG Cat-3 twisted pairs. Figure 8.23 shows the overview of operation of this technology. Basic wiring is similar to the two lower wires of the 100Base-T4 shown in Fig. 8.21c; however, in 100Base-T2 each wire carries 50 Mb/s rather than 33.3 Mb/s. This technology uses PAM 5 × 5 constellation to achieve 50 Mb/s per wire. As we discussed in Chapter 2, implementation of this two-dimensional constellation needs two pair of wires. In Example mprb3.8, using Fig. 3.7, we showed that the average energy in this constellation is given by $\bar{E} = 3d^2$, which has $6/(2/3) = 9$ or 9.5 dB better performance than the PAM-3 used in 100Base-T4. The average number of bits per constellation for the PAM 5 × 5 is

$$m = 4 \times \frac{1}{64}\log_2 64 + 12 \times \frac{1}{32}\log_2 32 + 9 \times \frac{1}{16}\log_2 16 = 4.5\ \text{bits/S}$$

Using this average number of bits per constellation, the required bandwidth for support of $100\,\text{Mb/s}$ is $100\,\text{Mb/s} \div 4.5\,\text{bits/S} = 22.2\,\text{MS/s}$, which is well below the 32 MS/s capacity of the Cat-3 wirings. Since it supports 100 Mb/s by using only two pairs of wires, if we use two pairs of wires for each direction then this technology can support full-duplex 100 Mb/s over Cat-3 wiring. Implementation of this complex signal constellation needs more sophisticated signal processing algorithms, such as NEXT echo cancellation and adaptive equalization used in voice-band, cable, and DSL modems, which increases the cost. As a result, though introduced in 1997, 100Base-T2 was not a commercial success, but it played a very important role in evolution of the Ethernet because it laid the foundation for gigabit per second Ethernet.

8.3.3 Gigabit Ethernet

By the late 1990s Ethernet had evolved into the most widely implemented physical and link layer protocol. Fast Ethernet increased the speed from 10 to 100 Mb/s, replacing the existing backbone FDDI installations, and 100Base-T4 in particular replaced 10 Mb/s legacy Ethernet as the main technology to access the end user. The next natural step was gigabit Ethernet to increase the speed to 1000 Mb/s and extend the Ethernet technology to wide-area networking. The initial standard for gigabit Ethernet was IEEE 802.3z, ratified

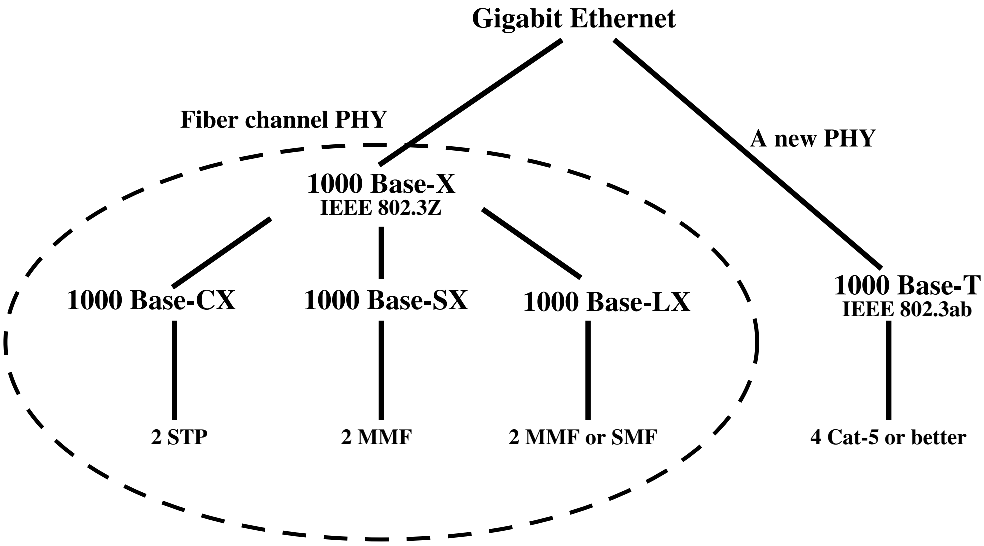


FIGURE 8.24 Overview of gigabit Ethernet options.

in 1998, for two pairs of fibers and STPs. The physical layer of this standard follows ANSI's fiber channel standard for gigabit storage networking. This is of a similar manner to the adoption of ANSI's FDDI physical layer by the 100Base-X standard and reflects the growth of Ethernet technology to assimilate other standards. In 1999, IEEE 802.3ab defined gigabit Ethernet for four pairs of unshielded Cat-5 or better media. This standard brought gigabit Ethernet to desktop terminals, because this standard is defined for the Cat-5 wiring already in place in many offices. As a result, some computer manufacturers started to use gigabit Ethernet connections in their products, and the gigabit Ethernet originally designed for the backbone became available to end users. Figures 8.24 and 8.25

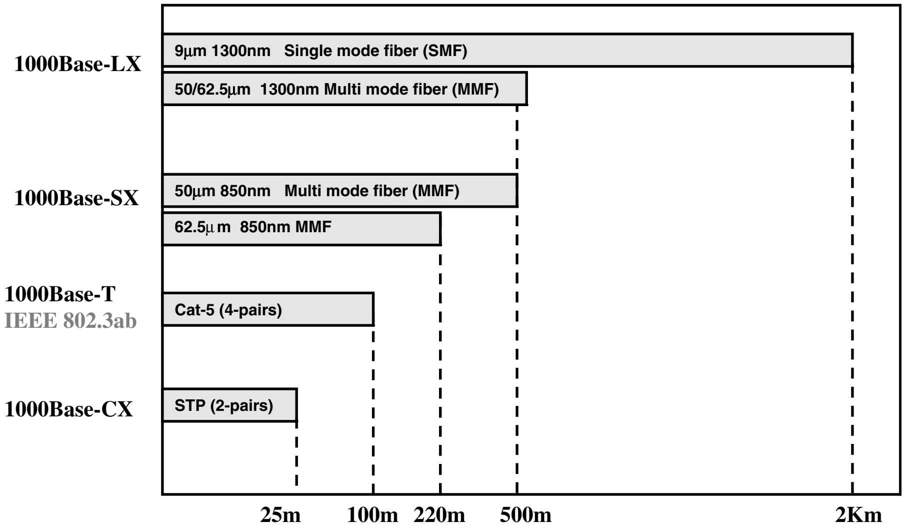


FIGURE 8.25 IEEE 802.3z and IEEE 802.3ab media options for gigabit per second Ethernet.

provide a summary of the gigabit Ethernet standards. The different media support a variety of lengths designed to address different applications. The 25 m distance for 1000Base-CX supports short distances such as connecting mass storage to the main computer. The 100 m distance of the 1000Base-T supports desktop access, and longer distances support a variety of backbone applications as Ethernet starts to penetrate metropolitan-area applications. This pattern of supporting higher data rates for longer distances continues into higher rate Ethernets, allowing this technology to penetrate wide-area networking.

1000Base-X operates similar to 100Base-FX shown in Fig. 8.18. In the 1000Base-X series, instead of 4B/5B we use 8B/10B coding and NRZ line coding (which was described in Chapter 3) is used instead of NRZ-I line coding. In 8B/10B coding, every 8 bits of user data is mapped into a 10-bit symbol prior to transmission over the media using NRZ line coding. The efficiency of the code is 80%, demanding a 1.25 Gb/s line for a data rate of 1 Gb/s. Since the 8-bit data takes $2^8 = 256$ possibilities, while there are $2^{10} = 1024$ choices for the coded sequence, a number of coded words are not used for mapping the data. Similar to 4B/5B and 8B/6T coding techniques, the coded words, originally selected by the ANSI group working on the fiber channel, have been selected so that the coded words are "DC balanced" so that valid coded symbols have five "1"s and five "0"s. This allows the receiver to align the symbols easily and ensures that incoming bits have frequent transitions to secure synchronization. In addition, some of the extra symbols are used to define transfer control signals, such as starting and ending delimiters and idle signal. The details of the coding table are beyond the scope of this book. Similar to the 100Base-X series, 1000Base-X supports full-duplex and autonegotiation operation.

1000BASE-T. Defined by IEEE 802.3ab, this is shown in Fig. 8.26. Similar to 100Base-T4, this standard also uses four pairs of wires to cover 100 m distance, but the minimum grade of the wires for the 1000Base-T is Cat-5, while the minimum grade of wiring for 100Base-T4 was Cat-3. The format of wiring for 1000Base-T shown in Fig. 8.25 is different from 100Base-T4 shown in Fig. 8.21. Similar to 100VG-ANYLAN, 1000Base-T distributes the load over all four pairs evenly. Similar to 100Base-T2, 1000Base-T uses multidimensional transmission using five levels. 100Base-T2 was using two-dimensional five-level PAM 5 × 5 transmission the 1000Base-T uses a four-dimensional five PAM-encoding technique

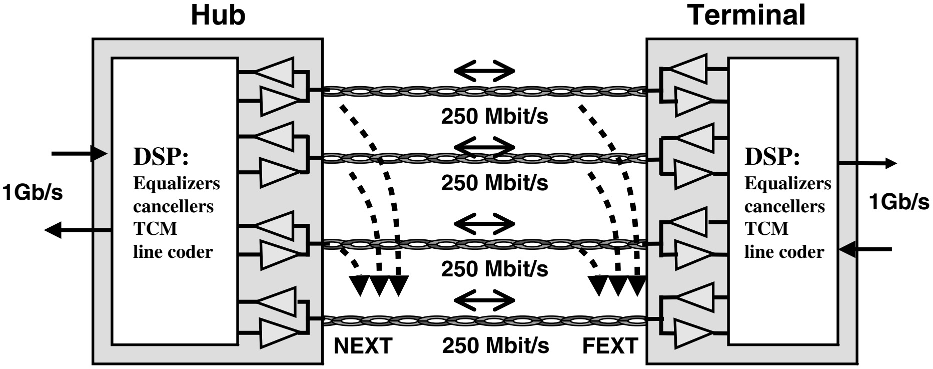
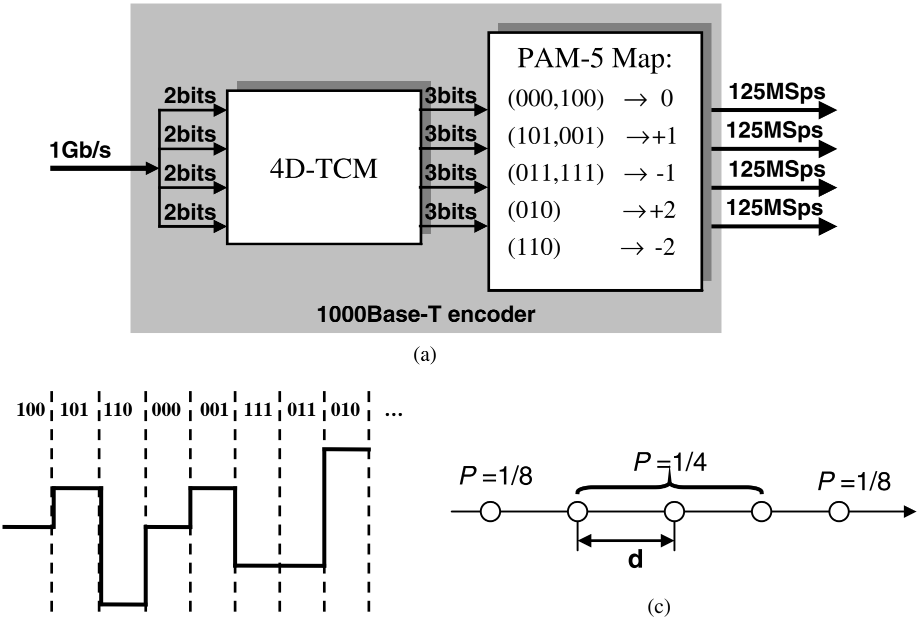


FIGURE 8.26 Overall operation of 1000Base-T.

which involves TCM, briefly introduced in Chapter 3. The TCM coding uses a trellis to code the transmitted symbols and it gives a 6 dB gain across the four wires. Each Cat-5, or better, line carries 125 MS/s using five levels (PAM-5); this is in contrast to 100Base-TX, which carries a binary NRZ-I signal at 125 MS/s. Since 1000BASE-T uses all four pairs of wires in one bundle for simultaneous transmission in both directions, as shown in Fig. 8.25, we will have NEXT and FEXT, which needs to be controlled using echo cancellation techniques. To implement a five-level PAM (PAM-5) we need to improve the channel using an equalizer.

Figure 8.27 shows the details of coding involved in 1000Base-T. As shown in Fig. 8.27*a*, the incoming data stream is divided into blocks of 8 bits which are further divided into four 2-bit blocks. The four 2-bit blocks are TCM coded to four 3-bit blocks. As we discussed in Chapter 3, the TCM adds one bit to each symbol to provide the redundancy that can be exploited using a trellis decoder at the receiver to provide for an overall 6 dB gain over uncoded transmission. The 3-bit TCM-coded data in each of the four lines is then mapped into the PAM-5 constellation. Figure 8.27*b* shows a sample of transmitted symbols in one of the four lines at the rate of 125 MS/s. The mapping rule of the PAM-5 constellation is so that the three middle points are transmitted twice as much as the corner points, which results in the signal constellation shown in Fig. 8.27*c*. As we showed in Example 3.4 of Chapter 3, for this constellation the average energy of the constellation is $\bar{E} = 3d^2/2$. Comparing with the binary constellation of Fig. 8.21*a*, with $\bar{E} = d^2/4$ the PAM-5 constellation provides another ${}^3/_2 \div {}^1/_4 = 6$ (7.8 dB) edge. The additional edges provided by TCM and PAM-5 coding and the use of sophisticated signal processing algorithms for



(a)

(b)

(c)

FIGURE 8.27 Details of 1000Base-T coding: (*a*) multidimensional TCM and PAM-5 mapping; (*b*) sample output of each of the four output lines; (*c*) signal constellation.

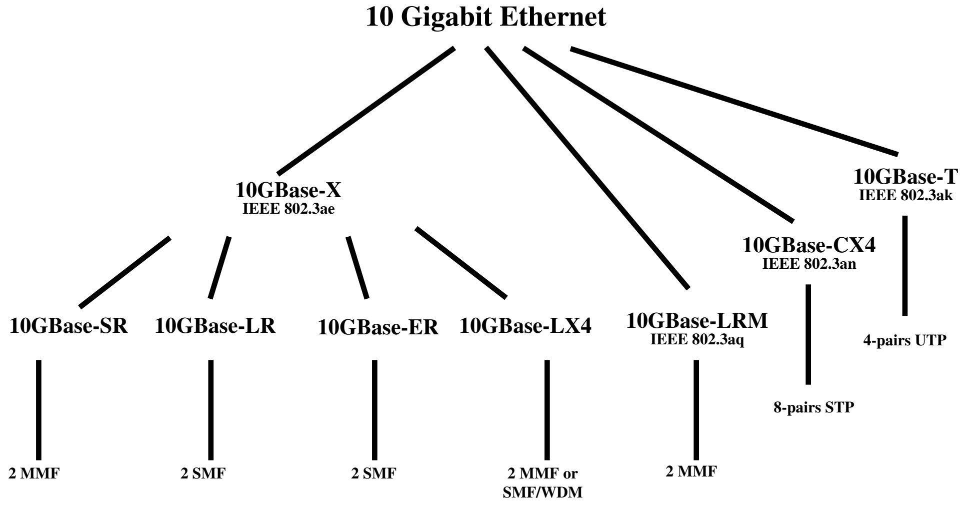


FIGURE 8.28 Overview of 10 gigabit Ethernet options.

echo cancellation and equalization enabled 1000Base-T to meet an acceptable transmission bit error rate.[3] The receiver for 1000Base-T is a complex receiver because each line needs an adaptive equalizer, an echo canceller, three NEXT cancellers and one trellis decoder. Therefore, we have 20 adaptive filters and four trellis decoders for a node, which results in a complex and relatively expensive network adaptor card.

8.3.4 10 Gb/s Ethernet and Beyond

In the early 2000s, different gigabit per second Ethernets became a popular product in the backbone of MANs, and several computer manufacturers included it as one of the connections in their desktop computer products. This success encouraged standardization of 10 Gb/s Ethernet, also referred to as 10GbE, for desktops and further penetration into WANs. Several standardization groups within the IEEE 802.3 community defined a number of 10 Gb/s Ethernet standards for different media options to support a variety of applications. At the time of writing, these standards are awaiting commercial acceptance to see which one of them will gain a sizable market. As shown in Fig. 8.28, there are four major standards: IEEE 802.3ae and IEEE 802.aq, which define 10 Gb/s for fiber media, and IEEE 802.3ak and IEEE 802.11an, which address 10 Gb/s over copper cables. The 10 Gb/s Ethernet is defined only for full-duplex operation, which does not use CSMA/CD. Therefore, the MAC, which was the first differentiating icon of the IEEE 802.3 Ethernet standard, is no longer used, and specification of the physical layer to specify the transmission technique and media options became the differentiation of the different subgroups working on 10 Gb/s Ethernet.

IEEE 802.3ae, ratified in 2003, defines four physical layer specifications to support transmission of Ethernet frames over MMF up to 300 m and SMF up to 40 km. In addition to

[3] For LAN applications this error rate is around 10^{-10}.

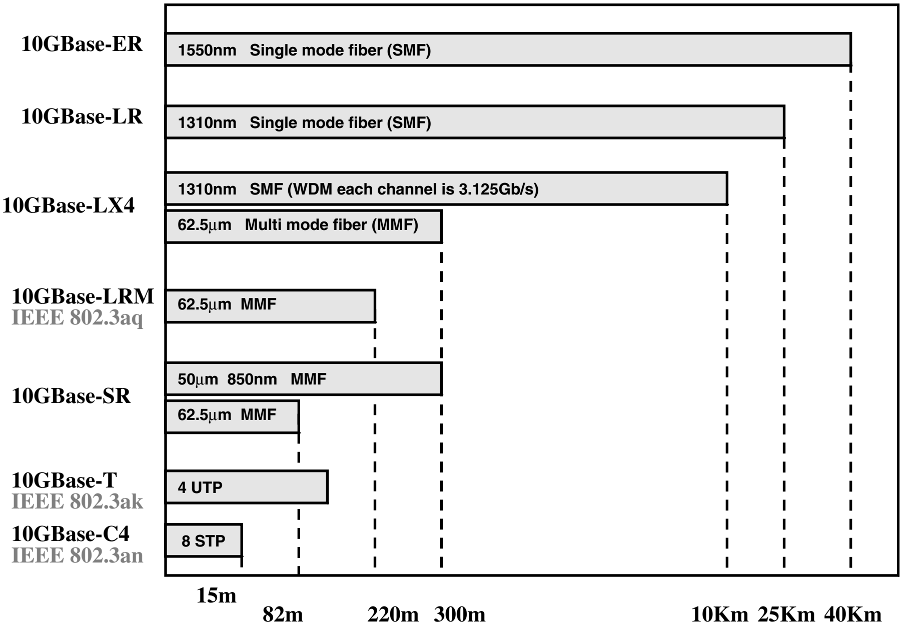


FIGURE 8.29 Overview of the IEEE 802.3ae and other 10 gigabit Ethernet options.

support for 10 Gb/s LANs, the IEEE 802.3ae also supports OC-192/STM-64 (SONET/SDH) at 9.958 Gb/s for wide-area networking and defines interoperation between them. The first of the IEEE 802.3ae specifications is the "short range" *10GBase-SR* defined for two MMF lines operating at 850 nm wavelength. As shown in Fig. 8.29, for existing 62.5 µm fibers it can cover up to 82 m and for new 50 µm fibers it can cover up to 300 m. The second of the IEEE 802.3ae specifications is the "long range" *10GBase-LR* supporting up to 25 km on 1310 nm SMF. The third of the IEEE 802.3ae standards is the "extended range" *10GBase-ER* defined for two SMF lines operating at 1550 nm wavelength supporting distances up to 40 km. The fourth of this standard series is the long-range *10GBase-LX4* using four lasers with WDM to support up to 300 m over deployed MMF. Each line operates at 3.125 Gb/s on a unique wavelength. This standard also supports 10 km over 1310 nm SMF. These three standards use 8B/10B line coding, which was also used in fiber channel and 1 Gb/s Ethernet.

The main technical differences between these technologies are the length of the cable and the line coding technique. In Chapter 2 we explained that the length of the cable depends on the LED type and its link budget, as well as the type of fiber and its path-loss characteristics. The coding technique supports DC offset by keeping the number of "1"s and "0"s in the blocks of data the same and also assures a minimum number of transmissions needed for adequate synchronization between the transmitter and the receiver. 10Base-LX4 uses the popular 8B/10B coding used in gigabit Ethernet, fiber channel, and many other applications, which is a look-up table coding similar to the 4B/5B or 8B/6T coding we described before. With the 8B/10B coding technique the effective data transmission rate is 4×3.125 Gbps $= 12.5$ Gbps, which supports an effective data transmission rate of 12.5 Gbps $\times$ $(8B/10B) = 10$ Gbps. The rest of the IEEE 802.3ae standard specification uses another line coding technique referred to as 64B/66B coding, which was also used in the fiber channel for storage area networking. We describe the details of this coding in more

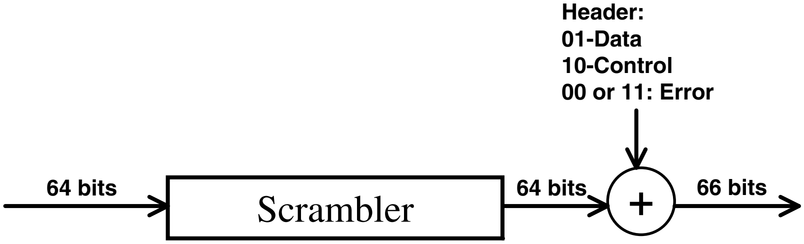


FIGURE 8.30 Overview of the 64B/66B coding.

detail because it provides a new and more efficient coding technique using scrambling techniques described in Chapter 4. With 64B/66B coding, the effective transmission rate of the coded symbols is 10 Mbps $\div$ $(64\text{B}/66\text{B}) = 10.3125$ Gbps rather than the coding rate of 12.5 Gb/s needed with the 6B/8B coding technique.

The principles of operation of the 64B/66B coding are explained in Fig. 8.30. In this coding technique the 64 bits of data is passed through a scrambler to achieve DC balancing and support an adequate number of transmissions. This means that, statistically, there are as many 1s as 0s in a string and that there are not too many 1s or 0s in a row. Another feature of look-up table codes was that we had a number of extra symbols which are not used for coded symbol transmission and we were using a few of them to construct starting delimiter, ending delimiter, idle and possibly other control signals. In 64B/66B coding, two preamble bits are added to the scrambled data to identify these situations. If the 2-bit preamble is "01" then the 64 bits are entirely data, if it is "10" then it contains control information, and if it is "00" or "11" then it indicates than an error has occurred. In addition, using 01 and 10 assures that, in the unlikely event of all "0" or all "1," after scrambling we will still have at least one transition at the start of the data. As we discussed in Chapter 4, scramblers use the LFSR structures which scrambles the received strings of 1s or 0s and they have been in common use in voice-band, DSL, cable, and OFDM wireless modems.

WAN technologies defined by IEEE 802.3ae specify physical transport for 10 Gb/s Ethernet across telecom OC-192/STM-64 SONET/SDH systems without having to directly map the Ethernet frames into SDH/SONET at 9.953 Gb/s. The main three standards for different wavelength ranges are mapped to their corresponding SDH/SONET. In this series, 10GBASE-SW corresponds to 10GBASE-SR, 10GBASE-LW corresponds to 10GBASE-LR, and 10GBASE-EW corresponds to 10GBASE-ER, each supporting the same type of fiber and its associated distance. There is no wide-area physical layer standard corresponding to 10GBASE-LX4 because the original SONET/SDH standard requires a serial implementation.

Another related standard for 10 Gb/s Ethernet over fiber lines is 10GBASE-LRM, ratified by IEEE 802.3aq in 2006. This standard supports up to 220 m over installed FDDI-grade 62.5 μm MMF operating at 1310 nm. The difference between this standard and 10GBase-LX4 is that the latter uses WDM while 10GBase-LRM has a single wavelength.

10 Gb/s Ethernet over four pairs of copper has two standards options specified by *IEEE 802.3ak* and *IEEE 802.3an*, compared with the other five fiber solutions shown in Figs. 8.28 and 8.29. The IEEE 802.3ak's standard 10GBASE-CX4, ratified in 2004, is the cheapest solution with no complex signal processing for distances up to 15 m. There are four pairs in each direction to support full-duplex operation. Similar to 10GBase-LX4, this standard uses

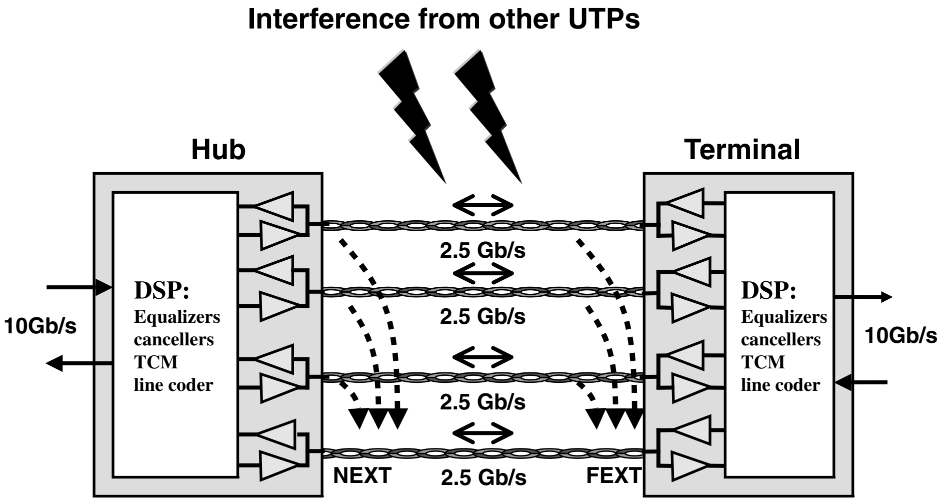


FIGURE 8.31 Overall operation of 1000Base-T.

four lanes, each carrying 3.125 Gb/s 8B/10B coded data, giving an effective data rate of 3.125 × 8/10 = 2.5 Gb/s which is NRZ coded for transmission. The 10GBase-LX4 uses 62.5 μm fiber SMF and MMF to cover long distances up to 10 km for wide-area applications and 10GBase-C4 uses four pairs of popular Cat-5 wiring in each direction to cover short distances up to 15 m for stacking and adjacent intra-rack connection and other short-distance applications. In the application of this technology, shielding interference between two sets of wires needs expensive cable assembly.

10GBase-T, specified by IEEE 802.3an in 2006, uses four twisted pairs of high-quality Cat-6 and Cat-7 wiring with a minimum symbol transmission rate of 800 MS/s, 55 dB echo, 40 dB NEXT, and 25 dB FEXT in a wiring structure, shown in Fig. 8.31, which is similar to the 1000GBase-T wiring shown in Fig. 8.26. The transmission technique for this standard is PAM-16 encoded in a two-dimensional double square DSQ-128 constellation, shown in Fig. 8.32. With unscreened Cat-6 this standard covers up to 56m, with screened Cat-6 or Cat-7 it covers up to 100 m. It uses the popular RJ-45 connectors at the two ends of the wiring.

Figure 8.33 shows the complex overall transmission technique used in 10GBase-T. The incoming data stream at 10 Gb/s is first 64B/65B encoded, which scrambles the 64 bits and adds a 1 bit preamble to differentiate data from control bits. In the next stage of coding, 9-bit CRC-8 parity codes, described in Chapter 2, are added to 50 consecutive blocks of 65 bits. The resulting 3259 bits are divided into one block of 1723 bits and one block of 1536 (3 × 512) bits. The 1723-bit block is passed through another (2048, 1723) block code to generate 2048 (4 × 512) bits. The resulting 512 blocks of 7-bit coded data are delivered to the DSQ-128 line-coding module to get mapped to the signal constellation shown in Fig. 8.32. Each two pairs of wires map one 7 bits of data to one of the symbols of the DSQ-128, resulting in $512 \div 2 = 256$ blocks of four-wire symbols. Each symbol is identified by two coordinates (a_n, b_n) and each of these coordinates can take 16 different values, forming a PAM-16 signal to appear on the associated line. Therefore, as shown in Fig. 8.33, each of the four output lines carries $7 \div 2 = 3.5$ bits with a symbol transmission rate of 800 MS/s, resulting in an equivalent 800 MSps × 3.5 bpS = 2.8 Gbps.

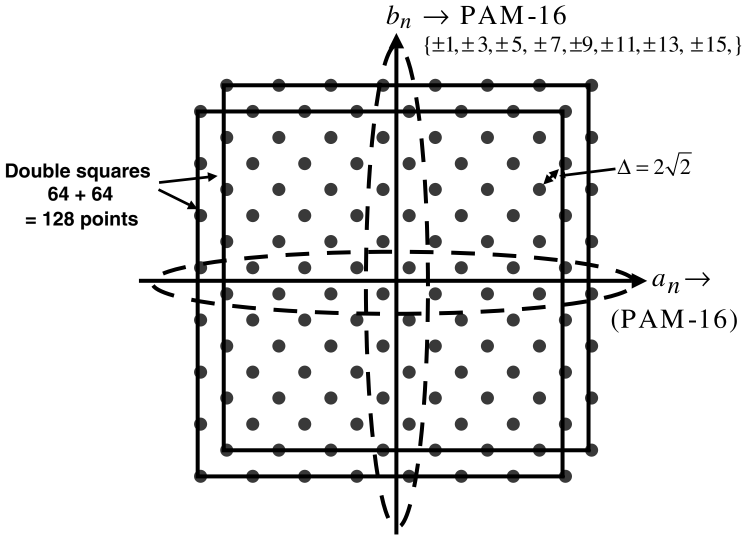


FIGURE 8.32 The double square 128-point (DSQ-128) constellation recommended by IEEE 802.3an for 10GBase-T.

Example 8.3: Input and Output Data Rates in 10GBase-T To check the details of the data transfer, consider the overall block coding technique used in 10GBase-T. Each $50 \times 64 = 3200$ bits of information data is coded into $7 \times 512 = 3584$ bits of coded data, resulting in an overall (3584, 3200) coding. The out put data rate and the input data rate are related by

$$4 \times 2.8\,\text{Gbps} \times \frac{3200}{3584} = 10\,\text{Gbps}$$

Similar to 1000Base-T, implementation of 10GBase-T requires complex signal processing to implement adaptive equalization, echo cancellation and constellation mapping at very high sampling rate and 10 bits per sample precession. As a result, at the time of writing, the cost of 10GBase-T may exceed $500 per port.

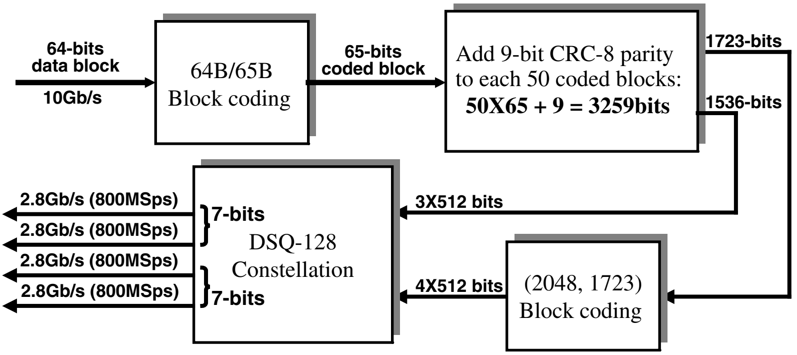


FIGURE 8.33 10GBase-T transmitter.

100 Gigabit Ethernet, or *100GbE*, is an Ethernet standard presently under early development by IEEE 802.3ba, with other data rates such as 40 Gb/s as options [IEEE07]. This standard will include 100GbE optical fiber Ethernet standards to support from 100 m up to 40 km with full-duplex operation using the Ethernet frame format. It is worth mentioning that the fundamental science behind complex technologies such as adaptive equalization, echo cancellation, TCM, and multidimensional modulations was originally designed for voice-band data communications at 2400 S/s and it took over half a century to develop [Pah88]. With the advancement in computing speeds and micro- and nano-technologies, these fundamental sciences are becoming available to high-speed data communications over wired and wireless channels. Therefore, the physical layer design of networks, which has been at the center of recent advancements in this field, owes itself to two roots: applied information theory for the design of voice-band modems and the advancement in computation power and memory size of recent years. Our emphasis in this book is on the fundamental changes in modern network technologies which are affected by applied transmission techniques and understanding and design of new media for data transmission.

8.4 EMERGENCE OF ADDITIONAL FEATURES FOR ETHERNET

With the emergence of Ethernet as the dominant LAN technology and its penetration into metropolitan- and wide-area networking, the IEEE 802.3 committee added more features to it to support a wider variety of applications. The Ethernet frame format defines only one frame type, while other competing legacy LAN technologies such as token ring or FDDI defined several frame formats to accommodate control signaling. The Ethernet frame format and MAC also does not support any priorities in the packet format. As Ethernet became more popular and dominated the LAN market, there was a need to make some modification to the legacy Ethernet frame format to accommodate enhancements to the operation of the network. The first change in the frame format resulted in the VLAN, and a second change introduced the PAUSE frame. The full-duplex operation uses PAUSE frame to avoid congestion. Another additional feature is link aggregation, which was introduced to allow parallel links between a switch and a terminal. In this section we address these four newly introduced features of the Ethernet.

8.4.1 Frame Format for the Virtual Local-Area Network

The basic concept behind a VLAN is shown in Fig. 8.34. A VLAN partitions the broadcast domain, allowing different groups to operate as virtual Ethernet LANs. In other words, a VLAN breaks a single broadcast domain into multiple domains to allow having multiple logical Ethernet switches on a single physical switch. The major benefits of a VLAN are easing network administration, allowing formation of work groups, enhancing network security, and providing a means of limiting the broadcast domain of the Ethernet. More details on VLANs are discussed in the Chapter 6 where we address the details of bridges. Here, we focus on implementation of VLAN tags on the legacy Ethernet frames. In order to implement a VLAN we need to change the frame format so that the bridges can recognize VLAN tags from common Ethernet frames. The change in the packet format should be done in the IEEE 802.3 community, and the IEEE 802.1 working on issues related to bridging needs to approve the changes so that the bridge manufacturers following IEEE 802.1 standards can implement the changes in the bridges. For the implementation of a VLAN,

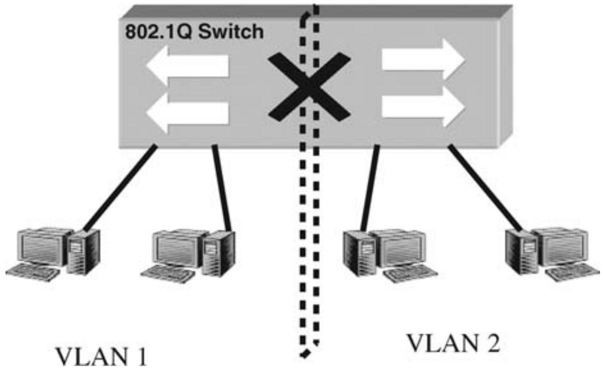
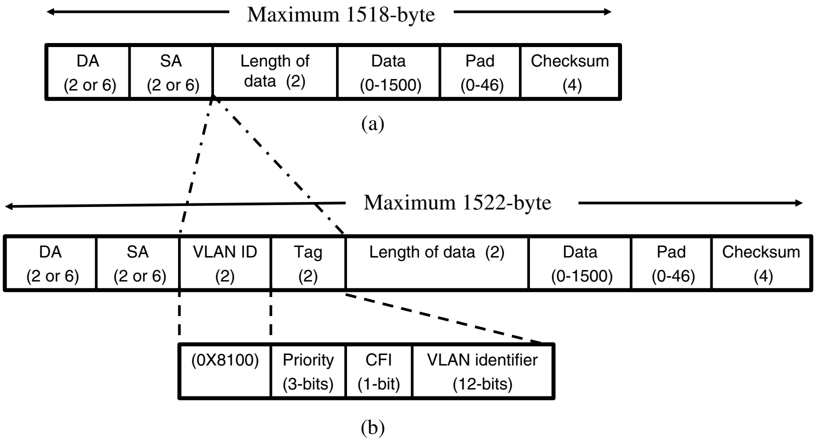


FIGURE 8.34 Basic concept behind VLAN partitioning.

IEEE 802.3ac defines the protocol and changes in the frame format to support VLAN tagging on Ethernet networks and gets its frame format changes approved by the IEEE 802.1Q. The VLAN protocol permits insertion of an identifier, or “tag,” into the Ethernet frame format to identify the VLAN to which the frame belongs. It allows frames from stations to be assigned to logical groups.

As we discussed in Section 8.2.1, the legacy Ethernet standard defined the minimum frame size as 64 bytes and the maximum frame length as 1518 bytes, 18 bytes of which is for the overhead. If we consider the common practice of 6-byte addressing, then these 18 bytes are used for MAC addressing, length of data, and the frame check sequence. For the minimum addressing length of 2 bytes it also includes the preamble and start of frame delimiter fields. In 1998 the IEEE 802.3ac standard released the VLAN tag specification, extending the maximum allowable frame size to 1522 bytes.

Figure 8.35 shows the frame format of the details of the IEEE 802.3 and IEEE 802.1Q frame formats. The 4-byte VLAN tag is inserted into the 802.3 frame between the source MAC address field and the length of the data field. The first 2 bytes of the VLAN tag is the fixed number 0X8100 which is greater than 1500 bytes, the maximum length of the data.



DA (2 or 6)	SA (2 or 6)	Length of data (2)	Data (0-1500)	Pad (0-46)	Checksum (4)

(a)

DA (2 or 6)	SA (2 or 6)	VLAN ID (2)	Tag (2)	Length of data (2)	Data (0-1500)	Pad (0-46)	Checksum (4)

(0X8100)	Priority (3-bits)	CFI (1-bit)	VLAN identifier (12-bits)

(b)

FIGURE 8.35 Frame format of (*a*) IEEE 802.3 Ethernet and (*b*) IEEE 802.1Q with VLAN tag.

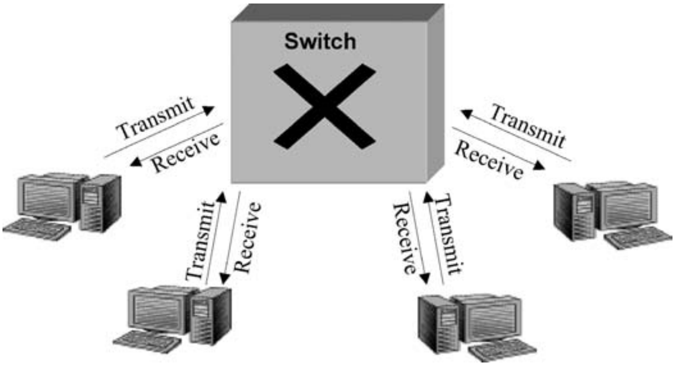


FIGURE 8.36 Basic concept behind full-duplex operation.

The 0X8100 value is actually reserved by the IEEE 802.1Q to indicate the presence of the VLAN tag and it indicates that the traditional length of the data field can be found at an offset of 4 bytes further into the frame. The last 2 bytes of the VLAN tag contain 3 bits as the user priority field to assign priority to Ethernet packets, 1 bit as canonical format indicator (CFI) to indicate the presence of routing information field,[4] and 12 bits are the VLAN identifier uniquely identifying the VLAN to which the Ethernet frame belongs. The additional 4 bytes increases the maximum length of the packet from 1518 to 1522 bytes. A LAN switch or bridge complying with the IEEE 802.1Q standard recognizes these tags from the normal frames when it reads the 0X8100 as the length of the packet, in which case it looks for the appropriate VLAN address to direct the packet.

8.4.2 Full-Duplex Operation

In 1997, IEEE 802.3x released the standard for full-duplex Ethernet packet transmission which bypasses the CSMA/CD protocol. Figure 8.36 illustrates the basic concept behind the full-duplex operation. Each terminal is connected to the hub by two separate lines used for transmission and the reception. Therefore, each terminal has a separate collision domain and there is no need for CSMA/CD. One pair of twisted-pair wire or one fiber strand is used exclusively for transmission and another pair or strand for reception. This wiring with full-duplex connection resembles telephone network wiring and switches, except that here we have packet switching rather than circuit switching. The CSMA/CD was a half-duplex protocol in which each terminal is allowed either to transmit or receive data, but never both at the same time. The half-duplex nature is inherited in the CSMA/CD protocol, for which simultaneous transmission and reception of data is an indicator of collision between the two packets. Full-duplex mode allows two stations to exchange data simultaneously over a point-to-point link with two independent physical transmit and receive wiring paths. In CSMA/CD, the embedded half-duplex operation for collision detection purposes would not allow full utilization of the transmission medium even when we have two separate pairs of wires in a point-to-point connection. The half-duplex in the presence of two transmission mediums connecting the terminals reduces the utilization of the transmission medium capacity to 50%. In full-duplex transmission over two pairs of

[4] The routing information field is practically only used for IEEE 802.5 frames.

transmission medium resources, each station can simultaneously transmit and receive data and the aggregate throughput of the link is effectively double the half-duplex operation. This means that a full-duplex 100 Mb/s station provides for 200 Mb/s of bandwidth if the physical medium is capable of supporting full-duplex with simultaneous transmission and reception without interference from other terminals. A star topology and two separate links are needed for full-duplex operation. Among the IEEE 802.3 media specifications we discuss in this chapter, the ones which follow this specification are 10Base-T, 100Base-X, 100Base-T2, 1000Base-X, and all 10GbE options. The bus topology of the 10Base5 and 10Base2 and the uneven wiring of 100Base-T4 do not allow full-duplex operation. Full-duplex operation is restricted to point-to-point links connecting only two stations when there is no contention for the shared medium and, consequently, no need for collision detection. Both stations must be configured for full-duplex operation and frames may be transmitted at will, limited only by the required separation of the minimum interframe gap of 96 bits.

In addition to doubling the maximum capacity, the efficiency of the link is improved by eliminating the potential for collisions. Elimination of the collision detection requirement lifts the segment length restriction as well. Segment lengths are no longer limited by the timing requirements of half-duplex Ethernet to ensure collisions are propagated to all stations within the required 512-bit times. This change allows support of longer transmission lengths. For example, 100Base-FX is limited to a 412 m segment length in half-duplex CSMA/CD operation mode, but with full-duplex operation mode it can support segment lengths as long as 2 km. The adaptor card and the hub must have full-duplex connectors. Therefore, full-duplex is used only with switches, and traditional repeaters cannot support full-duplex operation.

8.4.3 PAUSE Frames

The full-duplex mode defined by the IEEE802.3x includes an optional flow control operation implemented by a new frame called “PAUSE” frame. PAUSE frame in a full-duplex link between two end stations allows one of the stations to stop transmission of traffic temporarily from the other end station. Figure 8.37 shows the application of PAUSE frame in the full-duplex operation. Packets arriving from station A are at a rate that causes congestion in station B so that there is no input buffer space to receive additional frames. Station B transmits a PAUSE frame to station A to stop transmission for a specific period of time defined in the packet. When the PAUSE frame arrives in station A, this station suspends frame transmission for the specified time unless another PAUSE frame arrives and changes the suspension period for station A. After completion of the suspension period, station A resumes its normal packet transmission. This operation allows station B to recover from congestion during the specified PAUSE time requested from station A. If the congestion problem is completed sooner than the specified period, then station B has a choice to send another PAUSE frame with zero waiting time to activate terminal A. During the PAUSE state the station is allowed to send only PAUSE frames, allowing two stations to PAUSE each other at the same time. The PAUSE frame can be sent by a device not supporting PAUSE operation. Stations use the autonegotiation protocol which was defined as a part of the fast Ethernet standard to learn PAUSE capability of the station at the other end of the link.

PAUSE frame is a control signaling frame and to implement it one needs to find a way to differentiate among different Ethernet frames which were originally defined as a universal

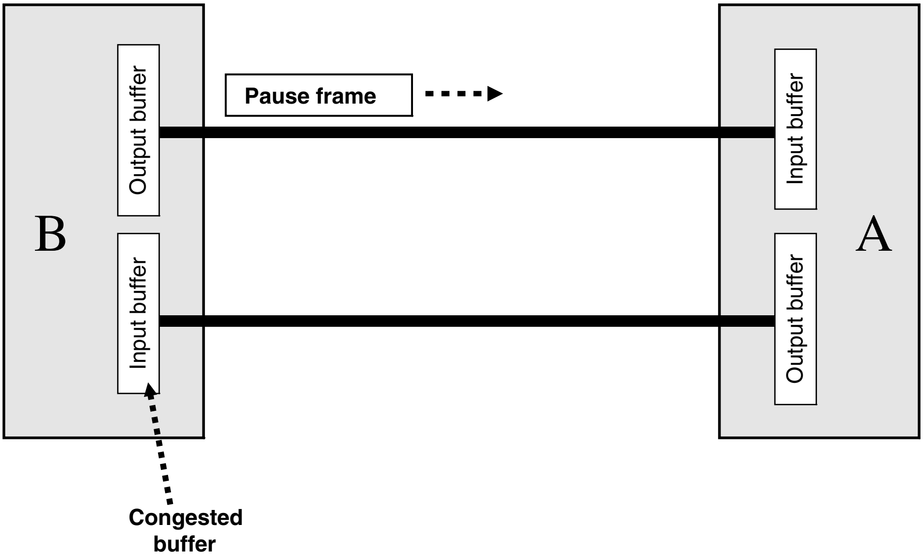
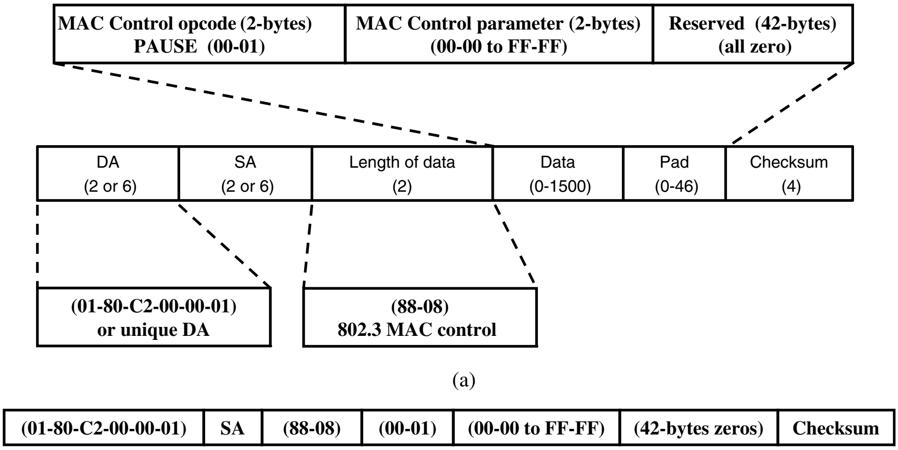


FIGURE 8.37 PAUSE frame application in full-duplex operation.

single formatted frame. Figure 8.38 shows the details for implementation of the PAUSE frame. In a manner similar to VLAN frames, IEEE 802.3x uses the length of the address field to differentiate between normal and control frames. This time the standardization committee assigns 88–08 (34 814 > 1500) to indicate that the packet is a MAC control packet. The next 2 bytes (00–01) defines the opcode for the PAUSE frame. The MAC control opcode field of 00–01 indicates the type of MAC control frame being used is a PAUSE frame. The PAUSE frame is the only type of MAC control frame currently defined; other combinations



(a)

(b)

FIGURE 8.38 Details of the PAUSE frame: (*a*) modifications in fields; (*b*) overall look of the packet.

of opcodes are available for future control signals. The PAUSE opcode is followed by another 2 bytes (00–00 to FF–FF) to specify the duration of the PAUSE. These 2-byte MAC control parameter fields specify the duration of the PAUSE event in units of 512-bit times. If an additional PAUSE frame arrives before the current PAUSE time has expired, then its parameter replaces the current PAUSE time, so a PAUSE frame with parameter zero allows traffic to resume immediately. The next 42 bytes are padded with 0s to keep the minimum length of 46 bytes for the data for Ethernet frames.

The destination address of the PAUSE frame may be set to either the unique destination address of the station to be paused or to the globally assigned multicast address (01–80–C2–00–00–01). This multicast address has been reserved by the IEEE 802.3 standard for use in MAC control PAUSE frames. It is also reserved in the IEEE 802.1D bridging standard as an address that will not be forwarded by bridges. This ensures the frame will not propagate beyond the local link segment.

8.4.4 Link Aggregation

In 2000, IEEE 802.3ad started working on the backbone link aggregation, which allows several NICs in a computer to connect to different ports of a switch. Figure 8.39 shows the basic concept behind link aggregation, which applies only to full-duplex operation. Link aggregation between two Ethernet stations increases the link bandwidth by combining multiple physical links into a single logical link. For example, it allows several less-expensive 1 Gb/s Ethernet solutions for long distances to support a higher data rate of up to 8 Gb/s to avoid using more expensive 10 Gb/s Ethernets. At first glance it seems that having multiple links between two stations should be possible anyway. The difficulty in having multiple links was that, as we discuss in Chapter 6, LAN switches use an STA to avoid loops, which eliminates parallel paths between two stations. Therefore, without link aggregation standardization, the only way to have multiple links between two stations in a single network is to use each link in a different VLAN. The link aggregation standard allows multiple links between servers, switches, and end-user stations to resolve this shortcoming of Ethernet networks. Each of the Ethernet ports involved in the link aggregation at the switch and at the terminal has its own unique MAC address. Link aggregation is implemented by a new layer of function between the MAC and the higher layer protocols. As frames pass this

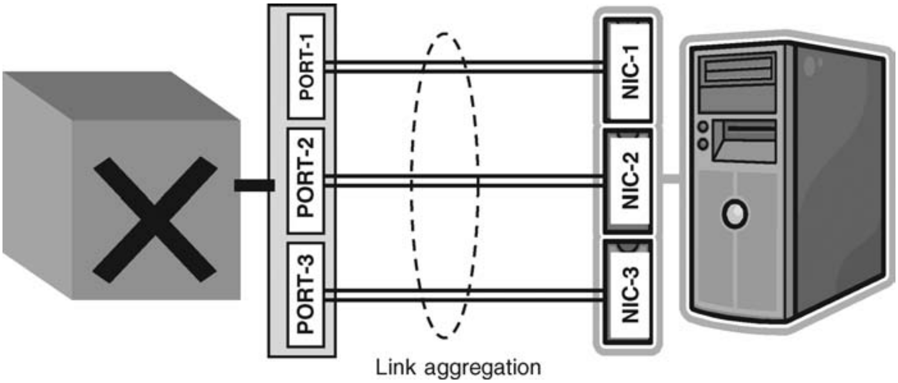


FIGURE 8.39 Basic concept behind link aggregation.

additional layer, MAC addresses are changed so that the aggregated ports appear as a single link with a unique MAC address. This method of operation masks the link aggregation function to all higher layer protocols and functions, such as the STA, or VLAN. The link aggregation protocol ensures the arrival of the frame sequence numbers distributed among the multiple links.

QUESTIONS

1. What are the main objectives of the IEEE 802.1 standardization activity and how do they relate to standards for LANs?
2. What is the purpose of the PAD field in the MAC frame of IEEE 802.3 and why it does not exist in other LANs such as IEEE 802.5?
3. What is the length of the contention slot (window) in IEEE 802.3? Why is this length selected?
4. What is the binary exponential back-off algorithm, which standard uses that, and what is the purpose of using it?
5. What are the options for the length of the MAC address field in IEEE 802 standards? Which length is commonly used in practice?
6. What is the difference between 10BASE-5, 10BASE-2, and 100BASE-T4 Ethernet in terms of data rate, coverage, type of cable, and type of connector?
7. What are the advantages of 4B/5B encoding over Manchester coding?
8. What is the difference between 100BASE-TX and 100BASE-FX?
9. What is the advantage of ternary signals over binary signals?
10. What is PAM 5X5, which standard uses it, and how does it compare with differential Manchester coding used in the legacy Ethernet?
11. Sketch the signal constellation used in 100BASE-T2 and identify the probability of transmission of each symbol. What is the symbol transmission rate and how does it relate to the overall data rate and the number of wirings of the LAN?
12. What category of TP wiring is used in 100BASE-T2 and 100BASE-T4? How many pairs of wires are used in each standard and how are these wires connected? Give a diagram for wiring connections.
13. What are NEXT and FEXT and how are they related to the selection of maximum length of wire in LANs?
14. What is the Pause Frame in Ethernet and how does it get implemented without changing existing Ethernet frames?
15. What is the purpose of the PAUSE frame and when is it used in a Full-Duplex operation?
16. What is a VLAN and how does it differ from traditional Ethernet in terms of frame format and services?
17. How many levels of priority and how many different VLANs are supported by the 802.1Q tag? How many bytes does this tag add to the 802.3 traditional MAC frames?
18. What is link aggregation and how is it useful to local network designers?
19. How does 1000Base-T provide 1Gbps over four pairs of Cat-5 TPs and what is the relation between PAM 5 modulation and 1000Base-T transmission techniques?
20. What is the maximum coverage of a Gigabit Ethernet and which medium supports this maximum coverage?

21. What is the difference between the MAC and PHY layers of 100BASE-X and FDDI?
22. Compare the complexity of the DSP in the 1000BASE-T and 1000BASE-F transceivers.
23. Draw the wiring format of 100BASE-T4 and 1000BASE-T and explain their differences.
24. Explain the differences among 10GBASE-XR and 10GBASE-XW standard specifications.
25. What is the difference between the 10GBASE-LR and 10GBASE-LRM standards for Ethernet?
26. What are the similarities and differences among 10GBASE-C4 and 10GBASE-LX4 technologies
27. Explain how DSQ128 coding used in 10GBASE-T works. Compare this technology with 4D-PAM5 coding.

PROBLEMS

Problem 1:

(a) In legacy 10Base5 Ethernet, what is the maximum length of the cable?

(b) What is the round trip delay of signal transmission over that length? Assume that the speed of propagation in cable is 200,000 Km/s.

(c) How does this length compare with the duration of transmission of 512 bits of data in 10Base5 Ethernet? Explain the reason for relation between the two values.

Problem 2:

(a) Sketch and label the waveform for Differential Manchester coded signal used in the IEEE 802.3 10BASE for the following 8-bit data stream

Given data stream: | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

(b) What is the symbol transmission rate of this system?

(c) If a 4B/5B coding with the same symbol transmission rate as in (b) was used to code the information bits

c-1) Give the transmitted waveform.
c-2) Give the information transmission rate.

(d) Repeat (c) if 8B/6T (rather than 4B/5B) was used.

Problem 3:

We want to install a LAN in a 5 story office building with identical floor plans. Each floor of the building is a 80 m × 80 m square with a height of 4 m. There are 40 terminals in each floor of the building and the external wiring comes to the first floor.

(a) What is the total cost of wiring, equipment and installation of the entire network if an IEEE 802.3 star network with one 240 port switch in the first floor connects all terminals to each other and the external connection? Assume a charge of \$150 per run of wiring between two locations, a \$3,000 cost for the switch, and a \$20 cost for the network interface card per terminal.

(b) To avoid wiring costs, assume that we use four access points per floor to support the terminals using a wireless solution. What is the total cost of the wireless LAN solution if each access point costs \$400 and each network interface card is sold for \$50? Note that we still need a switch and wiring from the switch to each access point.

(c) Compare the advantages and disadvantages of the two solutions

Problem 4:

(a) Considering Figure P8.1 for a token ring MAC protocol in which propagation time over the ring, *T*, is smaller than packet length, *P*, and the length of the token is assumed to be negligible. In this network the normalized propagation time $a = T/P < 1$. In this MAC protocol, after completion of transmission of a packet and its return to the original station the Token is passed to next station, $\frac{T}{N}$ away, in which N is the number of users connected to the ring. After $P + \frac{T}{N}$ period next station can transmit its packet. Show that the channel utilization or throughput of this MAC protocol is given by:

$$U = \frac{Message\ Time}{Average\ Cycle\ Time} = \frac{1}{1 + a/N}$$

(b) Draw a figure similar to Figure P8.1 for the case where $a = T/P > 1$ and use the figure to show that the utilization is now given by

$$U = \frac{Message\ Time}{Average\ Cycle\ Time} = \frac{1}{a + a/N}$$

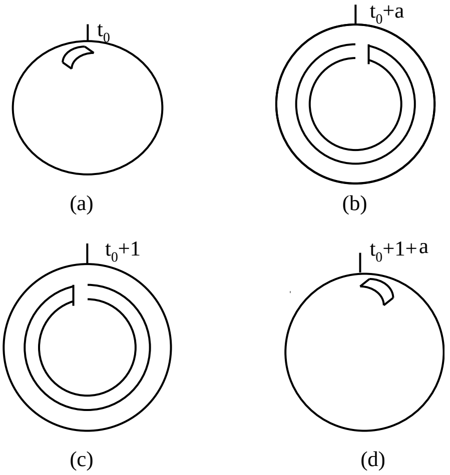


FIGURE P8.1: Packet transmission using Token Ring MAC (a) packet leaves a terminal (b) first bit is received after propagation time (c) last bit of the packet is transmitted (d) last bit is received.

(c) Using derivation of throughput presented in Section 8.2.3 give the utilization of the CSMA/CD in terms of normalized propagation delay.

(d) Plot the utilization of the of CSMA/CD and token ring MAC protocols as a function of N from 1 to 30 users for $a = 0.1$ and $a = 1$. Discuss the relative performance of the two protocols for small and large number of users.

(e) Use the equation for utilization of CSMA/CD and two cases of token ring to calculate the limits of utilization as N grows to infinity. Verify your results by checking the plots found in part (d).

Problem 5:

Consider a 10 Mbps, 1 Km LAN with N stations. Use results of Problem 4 to determine the throughput and delay if CSMA/CD or token ring MAC protocol were used in the LAN for the following conditions:

(a) Packet length = 1000 bits and N = 10

(b) Packet length = 1000 and N = 100

(c) Packet length = 10,000 and N = 10

(d) Packet length = 10,000 and N = 100

Problem 6:

Assume we have an environment with 5 users connecting to a 100BASE-T LAN all running the same application that produces 1500 byte length paths at a rate of 5 packets per minute.

(a) Determine the data transmission time T and normalized propagation parameter "a" assuming that the propagation speed is 200,000 Km/s and the terminals are located at the maximum length allowed by the standard for the network span.

(b) Determine the maximum utilization (throughput) S of the network assuming that the retransmission probability is p = 1/N.

(c) Determine the throughput if a 100VG-ANYLAN using token bus protocol.

(d) Compare the ratio of the throughput of the two LANs with the ratio of their throughputs if infinite number of terminals were used.

Problem 7:

In the 100BASE-T2 standard, rather than PAM5X5, if a 16-QAM constellation was used,

(a) What would be the difference in the number of signal levels and the symbol transmission rate in each of the 2-pairs of TP wires?

(b) What would be the difference in received signal to noise ratio requirement in dB?

(c) If we assume a signal loss of 10dB per 100 meters for the twisted pair wiring, what would be the maximu length of the cable if we were using 16-QAM instead of PAM5X5?

9

IEEE WIRELESS LOCAL-AREA NETWORK STANDARDS

9.1 INTRODUCTION

In Chapter 8 we considered IEEE 802.3 or Ethernet, the primary wired technology that is used as the access network to the Internet. Ethernet also started as a local-area networking technology that connected hosts that belonged to the same organization. In a similar manner, IEEE 802.11, or WiFi, is the primary WLAN technology. For wider areas, IEEE 802.16 or WiMAX is emerging as the wireless access networking technology. The goal of this chapter is to provide an overview of both of these technologies. Obviously, WiFi is the more mature technology, with widespread deployment in organizations, campuses, hotspots (in coffee shops, airports, and hotels) and residences. Consequently, our treatment of WiFi is more detailed than that of WiMAX. In this section, we provide a history of wireless local-area networking. Section 9.2 discusses the IEEE 802.11 standard from a top-down view. Section 9.3 briefly discusses WiMAX.

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

During the past two decades, as the vision of the WLAN industry evolved, WLANs were implemented based on a variety of innovative technologies and raised a lot of hopes for development of a sizable market several times. Today, the major differentiation of WLANs from wide-area cellular services is the method of delivery of data to users, data rate limitations, and frequency-band regulation. Cellular data services are delivered by operating companies as services, while the WLAN users belong to the organization that owns the network. At a time when the 3G cellular industries are striving for 2 Mb/s packet data services, WLAN standards are working on more than 100 Mb/s services. Another differentiation with other radio networks is that today almost all WLANs operate in the unlicensed bands, where frequency regulations are loose and there is no charge or waiting time to obtain the band. To obtain a deeper understanding of all these issues it is very educational to go over the history of the WLAN industry to see how all of these unique issues evolved.

9.1.1 Early Experiences

Gfeller at the IBM Rüschlikon Laboratories in Switzerland first introduced the idea of a WLAN in the late 1970s [Gfe80]. The number of terminals in manufacturing floors was growing and wiring them in that environment was difficult. In offices, wires are normally snaked within suspended ceilings and through the interior partitions and walls, but these options are not available in manufacturing floors. In office environments, in extreme cases, it is possible to install wiring under the floor, using conduits, or even simply left over the floor with some cover. In manufacturing floors, the environment is rugged, underfloor wiring is more expensive, and simply leaving wires on the floor can be dangerous because heavy machinery may roll over them. DFIR technology was selected at IBM laboratories for the implementation of a WLAN to avoid interference with the electromagnetic signals radiating from machinery and to avoid dealing with long-lasting administrative procedures with frequency administration agencies. The principal researcher of this project abandoned the project because the goal of 1 Mb/s with reasonable coverage did not materialize.

Ferrert, at HP's Pal Alto Research Laboratories in California, performed the second project on WLANs around the same time [Fer80]. In this project, a 100 kb/s DSSS WLAN operating around 900 MHz was developed for office areas that used CSMA as the method of access. This project was conducted under an experimental license agreement from the FCC. The principal of this project failed to obtain the necessary frequency bands from the FCC and, discouraged with the administrative complexity, he also abandoned his project. A couple of years later, Codex, Motorola attempted to implement a WLAN at 1.73 GHz, and that project was also abandoned after negotiations with the FCC.

Although all the pioneering WLAN projects were abandoned, WLANs continued to attract attention and negotiations continued with the FCC to secure frequency bands for this purpose [Pah85]. These projects revealed several important challenges facing the WLAN industry that remain to this day:

1. *Complexity and cost.* The alternatives for implementing WLANs, such as IR, spread spectrum, or traditional radios, are far more complex and diversified than wired LANs.
2. *Bandwidth.* Data rate limitations of the wireless medium are more serious than those of wired media.
3. *Coverage.* The coverage of a WLAN operating within a building is less than that of a single cable (bus or ring) or even twisted-pair-based LANs.

4. *Interference.* WLANs are subject to interference from other overlaid WLANs or other users operating in the same frequency bands.
5. *Frequency administration.* Radio-based WLANs are subject to expensive and untimely frequency regulations.

9.1.2 Emergence of Unlicensed Bands

WLANs need a bandwidth of at least several tens of megahertz, while they have not yet shown a market compatible in strength with the cellular voice industry that originally started with two pieces of 25 MHz bands that produced a huge market. Comparable sizes of bands for PCS applications were auctioned in the USA for tens of billions of dollars, while the market for WLANs had not yet passed a billion dollars per year. The dilemma for the frequency administration agencies was to justify a frequency allocation for a product with a weak market.

In the mid 1980s, the FCC found two solutions for this problem. The first and the simplest solution was to avoid the 1–2 GHz bands used for the cellular telephone and PCS applications and approve higher frequencies at several tens of gigahertz where plenty of unused bands were available. This solution was first negotiated between Motorola and the FCC and resulted in Motorola's Altair, the first WLAN product operating in the licensed 18–19 GHz bands. Motorola had actually established a headquarters to facilitate user negotiation with the FCC for the usage of WLANs in different areas. A user who changed the location of operation of their WLAN substantially (from one town to another) contacted Motorola and they would manage the necessary frequency administration issues with the FCC.

The second and more innovative approach was resorting to unlicensed frequency bands as the solution. In response to the applications for bands for WLAN projects mentioned in the previous section, and motivated by studies for various implementations of wireless LANs [Pah85], Mike Marcus of the FCC initiated the release of the unlicensed ISM bands in May 1985 [Mar85]. The ISM bands were the first unlicensed bands for consumer product development and played a major role in the development of the WLAN industry. In simple words, licensed and unlicensed bands can be compared to private backyards and public gardens. If one can afford it, one can own a private backyard (licensed band) and arrange a barbeque dinner (a wireless product). If one cannot afford to buy a house with a backyard, then one simply moves the barbeque party to the public park (unlicensed band) where one should observe certain rules or *etiquette* that allow others to share the public resource as well. The rules enforced on ISM bands restricted the transmit power to 1 W and enforced the modems radiating more than 1 mW to employ spread-spectrum technology. It was believed that spread-spectrum communications would restrict interference and allow the coexistence of several wireless applications in the same band. Table 9.1 provides a summary of the important features of the ISM bands.

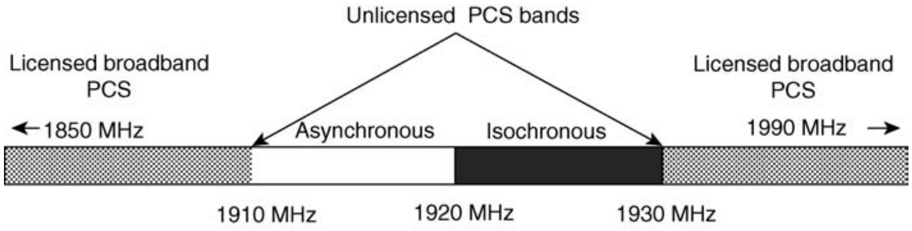
TABLE 9.1 Summary of ISM Bands

Properties of the ISM bands
Frequencies of operation: 902–928 MHz; 2.4–2.4835 GHz; 5.725–5.875 GHz
Transmit power limitation of 1 W for DSSS and FHSS
Low power with any modulation

9.1.3 Products, Bands, and Standards

Encouraged by the FCC ruling and some visionary publications in wireless office information networks summarizing previous works and addressing the future directions in this field [Pah85, Pah88, Kav87], a number of WLAN product development projects mushroomed almost exclusively over the North American continent. By the late 1980s the first generation of WLAN products using three different technologies, licensed bands at 18–19 GHz, spread spectrum in the ISM bands around 900 MHz, and IR appeared in the market. At around the same time a standardization activity for WLANs under IEEE 802.4L was initiated that was soon converted into an independent unit – IEEE 802.11, which was finalized in 1997! The first-generation products consisted of shoebox-sized APs and receiver boxes or PC-installed cards that could connect workstations to LANs wherever wiring difficulties for the LANs justified using a more expensive WLAN connection. Today, we call this application LAN-extension [Pah95]. Market predictions at that time were estimating a shift of around 15% of the LAN market to WLANs that would generate a few billion dollars of sales per year by the first few years in the 1990s. In May 1991, to create a scientific forum for the exchange of knowledge on WLANs, the first IEEE-sponsored WLAN workshop was organized concurrent with the 802.11 meeting, in Worcester, Massachusetts [Wor91].

In 1992, as a follow up to the initial momentum for WLAN developments, led by Apple, an industrial alliance called WINForum was formed aimed at obtaining more unlicensed bands from the FCC for so-called Data-PCS activities. WINForum finally succeeded in securing 20 MHz of bandwidth in the PCS bands that was divided into two 10 MHz bands – one for isochronous (voice-like) and one for asynchronous (data-type) applications. The original aim of WINForum was to secure 40 MHz for asynchronous applications. WINForum also defined a set of rules or etiquettes for these bands that would allow the coexistence. Figure 9.1 shows the unlicensed PCS bands and the spectrum etiquette associated with them. The WINForum etiquette is based on CSMA rather than CDMA and spread-spectrum communications used in ISM bands. This was a better choice, because implementation of CDMA needed power control and a larger bandwidth that was not feasible in uncoordinated, multiuser, multivendor WLANs and spread spectrum without CDMA offers a lesser bandwidth-efficient solution.



Three basic rules

1. Listen before talk (or transmit) *LBT protocol*
2. Low transmitter power
3. Restricted duration of transmissions

FIGURE 9.1 Unlicensed PCS bands and their spectrum etiquette.

TABLE 9.2 Properties of the U-NII Bands

Band of operation (GHz)	Maximum Tx power (mW)	Max. power with antenna gain of 6 dBi (mW)	Maximum PSD (mW/MHz)	Applications: suggested and/or mandated	Other remarks
5.15–5.25	50	200	2.5	Restricted to indoor applications	Antenna must be an integral part of the device
5.25–5.35	250	1000	12.5	Campus LANs	Compatible with HIPERLAN
5.725–5.825	1000	4000	50	Community networks	Longer range in low-interference (rural) environs

Another standardization activity started in 1992 was the HIPERLAN. This ETSI-based standard aimed at high-performance LANs with data rates of up to 23 Mb/s, which was an order of magnitude higher than the original 802.11 data rates of 2 Mb/s. To support these data rates, the HIPERLAN community was able to secure two 200 MHz bands: 5.15–5.35 GHz and 17.1–17.3 GHz for WLAN operation. This encouraged the FCC to release the so-called U-NII bands in 1997 when the original HIPERLAN standard (now called HIPERLAN/1) was completed. Table 9.2 summarizes the U-NII bands and their restrictions. The WINForum etiquette was evaluated for the U-NII bands, but it was found unsuitable because research activities around that time favored wireless ATM that could not operate on an LBT etiquette. Today, U-NII bands are used by IEEE 802.11a/n and HIPERLAN/2 projects for the implementation of > 54 Mb/s OFDM-based WLANs. In the early 2000s, the FCC allocated several gigahertz of bandwidth around 60 GHz for wireless applications, and the IEEE 802.11 and 802.15 working groups are considering extremely high data-rate WLAN- and WPAN-related standards in these frequency bands, especially in light of the plentiful bandwidth available. We discuss WLAN standards in more detail in Section 9.2.6.

9.1.4 Shift in Marketing Strategy

In the first half of the 1990s, WLAN products were expecting a sizable market of around a few billion dollars per year for shoebox-sized products used for LAN-extension in indoor areas, but this did not materialize. Under this situation, two new directions for product development emerged. The first and simplest approach was to take the existing shoebox-type WLANs, boost up their transmitted power to the maximum allowed under regulations, and equip them with directional antennas for outdoor interbuilding LAN interconnects. These technically simple solutions would allow coverage of up to a few tens of kilometers with suitable rooftop antennas. The new inter-LAN wireless bridges could connect corporate LANs that were within range. The cost of the inter-LAN wireless solution was much cheaper than the wired alternative, T1-carrier lines, leased from the PSTN service providers. The second alternative was to reduce the size of the design to a PCMCIA WLAN card to be used with laptops, which were enjoying a sizable growth and demanded mobility for LAN connectivity. However, this approach was not available for all existing products, and it was more suitable for the spread-spectrum products operating in lower frequencies. Figure 9.2

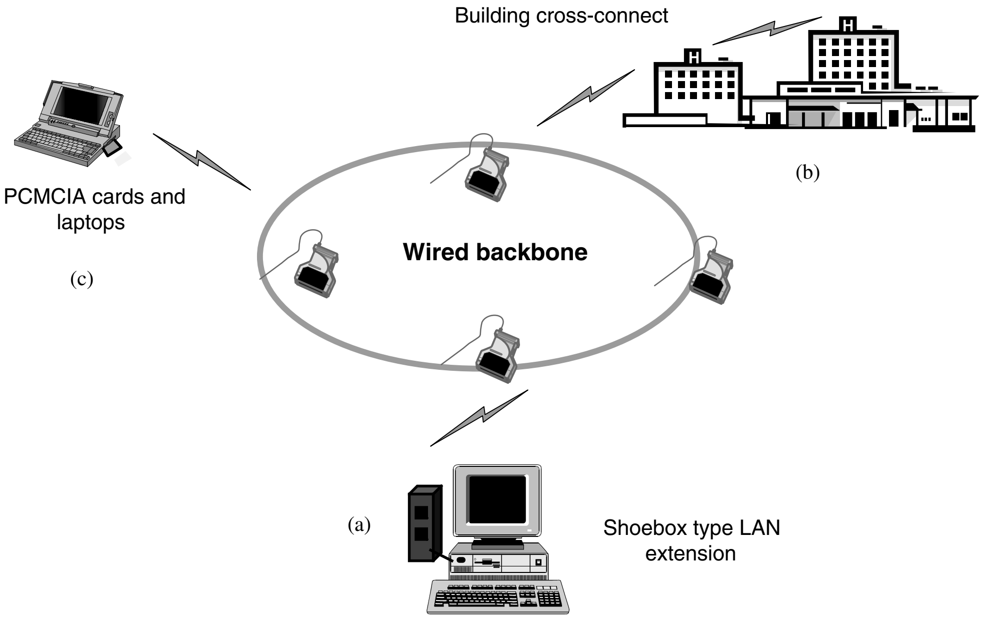


FIGURE 9.2 Different forms of WLAN products: (*a*) LAN-extension; (*b*) Inter-LAN bridge; (*c*) PCMCIA cards for laptops.

illustrates these three applications for WLANs. Moreover, there are low-cost products for LAN extension that can convert a serial port or Ethernet connector to a WLAN interface for desktop PCs and workstations.

The original marketing strategy for LAN-extension application was indeed a horizontal one aiming at selling individual WLAN components directly to the customers. Another major shift in the marketing strategy of a few successful companies in the mid 1990s was the move toward vertical markets, where a wireless network was sold as a complete solution to an application. The major vertical markets approached by the WLAN industry were "*barcode*" *industries* providing wireless inventory checking and tracking in warehouses and manufacturing floors, *financial services* providing wireless financial updates in large stock exchanges, *healthcare* networks providing wireless mobile services inside hospitals, and *wireless campus-area networks* (WCANs) providing for wireless classrooms and offices. All these efforts boosted the market for WLANs to above half a billion dollars per year over the last few years of the 1990s.

Example 9.1: A WCAN in WPI Figure 9.3 illustrates the schematic of an experimental NSF-sponsored WCAN that was design as a testbed for performance monitoring of WLAN products at the Center for Wireless Information Network Studies (CWINS), WPI in 1996. The testbed connects five buildings with inter-LAN bridges using different technologies. Inside each building, APs provide coverage to the laptops that are carried by the students. The professor broadcasts his image and writing on the electronic board to allow students to participate in the wireless classroom from different buildings in the campus. The entire wireless network is connected to the backbone through a router to isolate the traffic for traffic monitoring experimentations.

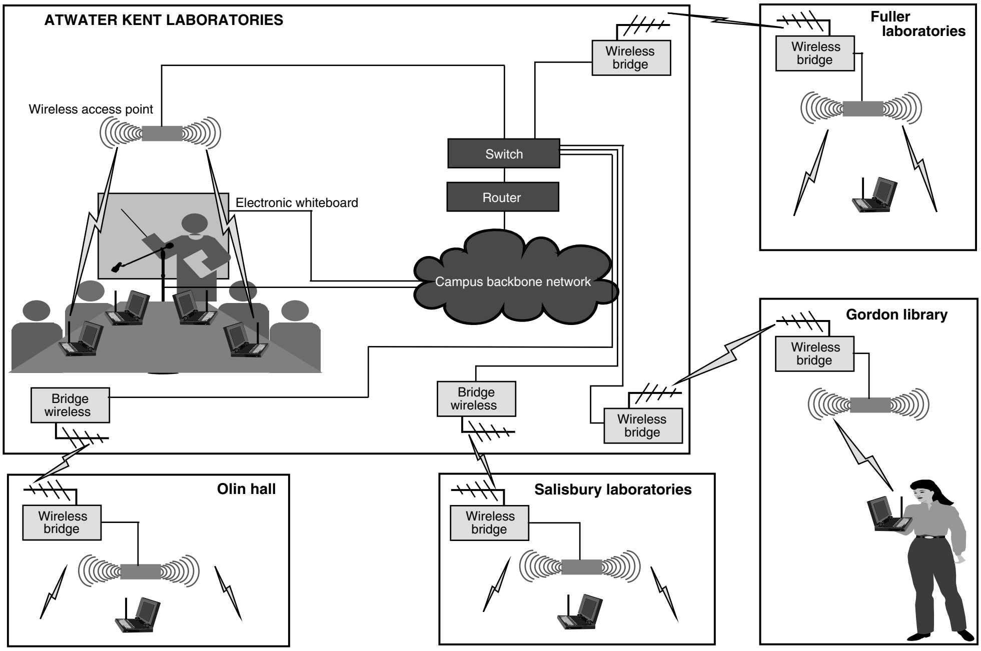


FIGURE 9.3 The experimental NSF-sponsored WCAN at WPI.

Today, the horizontal market for the WLAN industry is mainly focused on WLANs as an alternative to wiring additional LAN segments wherever the cost of the WLAN is justifiable. One example of this situation is installation with frequent relocations where the additional cost of the WLAN solution is justified by the relocation costs of the wired solution. Temporary networking situations, such as registration sites in conferences or fairs (jobs, food, etc.), is another example where a wireless solution is preferred to the expensive but more reliable wired alternative. Buildings with difficult- or impossible-to-wire situations, such as marble buildings or historical monuments where drilling for wiring is not favored, provide another example of situations where WLANs are justified. The most prominent incentive for WLANs is their general use in laptops and smart phones in the home and offices.

9.2 IEEE 802.11 AND WLANs

IEEE 802.11 is the first WLAN standard and, so far, the only one that has secured a market. Within the IEEE, there are several standards activities carried on by different groups. The IEEE 802 LAN/MAN Standards Committee is responsible for LAN standards and MAN standards. Individual working groups are in charge of a variety of LAN/MAN standards, of which the 802.11 working group is responsible for *wireless local-area networking standards.* The IEEE 802.11 standardization activity originally started in 1987 as a part of the IEEE 802.4 token bus standard under the group number IEEE 802.4L. The IEEE 802.4, a counterpart of the IEEE 802.3 and 802.5, pays special attention in supporting factory environments. One of the early motives for using WLANs was in factories for control of and communication between equipment. For this reason, car manufacturers such as GM were actively participating in the IEEE 802.4L activities in the early days of this industry. In 1990 the 802.4L WLAN group was renamed as IEEE 802.11, an independent 802 standard, to define the PHY and MAC layers for WLANs.

The IEEE 802.11 standard was the first WLAN standard facing the challenge of organizing a systematic approach for defining a standard for wireless wideband local access. Compared with wired LANs, WLANs operate in a difficult medium for communication and they need to support mobility and security. The wireless medium has serious bandwidth limitations and frequency regulations. It suffers from time- and location-dependent multipath fading. It is subject to interference from other WLANs, as well as from other radio and nonradio devices operating in the vicinity of a WLAN. Wireless standards need to have provisions to *support mobility* unlike other LAN standards. The IEEE 802.11 body had to examine connection management, link reliability management, and power management, none of which were concerns for other IEEE 802 standards. In addition, WLANs have no physical boundaries and they overlap with each other; therefore, the standardization organization needed to define provisions for the *security* of the links. For all these reasons, and because of several competing proposals, it took 10 years for the development of IEEE 802.11, which was far longer than other 802 standards designed for wired mediums. Once the overall picture and the approach became clear, it only took a reasonable time to develop the IEEE 802.11b and IEEE 802.11a enhancements. The IEEE 802.11n standard is at the verge of being approved as of now.

9.2.1 Overview of IEEE 802.11

A voice-oriented connection-based standard, like many (cellular systems), begins with the first step of specifying the services to be provided. Then the reference system architecture and its interfaces are defined, and finally the detailed layered interfaces are specified to accommodate all the services. The situation in connectionless data-oriented networks, such as IEEE 802.11, is quite different. The IEEE 802.11 standard provides for a general PHY and MAC layer specification that can accommodate any connectionless application whose transport and network layers accommodate the IEEE 802.11 MAC layer. Today, TCP/IP is the dominant transport/network layer protocol hosting most popular connectionless applications, such as web access, e-mail, FTP, or Telnet, and it works over all MAC layers of LANs, including IEEE 802.11. Therefore, the IEEE 802.11 standard does not need to specify applications or services. However, IEEE 802.11 provides for local and privately owned WLANs with a number of competing solutions. In this situation, the first step in the standardization is to group all the solutions into one set of requirements with a reasonable number of options. The next step, in a manner similar to connection-based standards, is to define a reference system model and its associated detailed interface specifications.

The number of participants in the IEEE 802.11 standards soon exceeded 100 with a number of solutions. The finalized set of requirements, which did not came about easily, were:

- single MAC to support multiple PHY layers;
- should allow multiple overlapping networks in the same area;
- handle the interference from other ISM-band radios and microwave ovens;
- mechanism to handle “hidden terminals”;
- options for time-bounded services;
- provisions on privacy and access control.

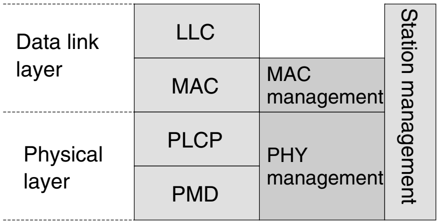
In addition, it was decided that the standard would not be concerned with licensed-band operations. These requirements set the overall direction of the standard in adopting different alternatives. However, as often happens in these types of standard, the actual adoptions were based on successful products that were already available in the market.

The 802.11 standard, like most LAN standards, is concerned only with the lower two layers of the OSI stack, namely the PHY and MAC layers. The MAC and PHY layers operate under the IEEE 802.2 LLC layer that supports many other LAN protocols. In the case of wired LAN standards such as 802.3, there are several physical layers that correspond to the same MAC specifications. A good example is IEEE 802.3, which was originally designed for thick coaxial cable, but was subsequently revised to include thin coaxial cable, a variety of twisted-pair cables, and even fiber optic links. In the same way, the IEEE 802.11 standard specifies a common MAC protocol that is used over many different PHY standards. The PHY standards are the “base” IEEE 802.11 standard, the 802.11b and g standards and the 802.11a standard. A new 802.11n physical layer is under consideration by the 802.11 working group. The MAC protocol is based on CSMA/CA. An optional polling mechanism called point coordination function (PCF) is also specified. In addition to the MAC and PHY layers, the IEEE 802.11 standard also specifies a management plane that transmits management messages over the medium and can be used by an administrator to tune the MAC and PHY layers. The MAC layer management entity (MLME) deals with management issues such as roaming and power conservation. The PHY layer management entity (PLME) assists in channel selection and interacts with the MLME. A station management entity (SME) handles the interaction between these management layers. Figure 9.4 shows the protocol stack associated with IEEE 802.11.

The base IEEE 802.11 standard specifies three different PHY layers: two using RF and one using IR communications. The RF PHY layers are based on spread spectrum, either direct sequence (DS) or frequency hopping (FH), while the IR PHY layer is based on pulse position modulation (PPM). Two different data rates are specified, 1 and 2 Mb/s, for each of the three PHY layers. The RF physical layers are specified in the 2.4 GHz ISM unlicensed frequency bands.

Example 9.2: Origins of PHY Layer Solutions The DSSS solution for IEEE 802.11 is based on WaveLAN, which was designed at NCR, Netherlands [Tuc91, WPI]. The FHSS solution was highly affected by RangeLAN designed by Proxim, CA, and products from Photonics, CA, and Spectrix, IL, affected the DFIR standard.

The IEEE 802.11b standard specifies the physical layer at 2.4 GHz for higher data rates: 5.5 Mb/s and 11 Mb/s. The PHY layer makes use of a modulation scheme called



PLCP: Physical layer convergence protocol
PMD: Physical medium dependent

FIGURE 9.4 Protocol Stack of IEEE 802.11.

complementary code keying (CCK). The transmission rate depends on the quality of the signal and it is backwards compatible with the DSSS-based base-802.11 standard. Depending on the signal quality, the transmission rates could fall back to lower values. The 802.11g standard further increases the data rates to up to 54 Mb/s in the 2.4 GHz ISM bands using OFDM. The IEEE 802.11a standard [Kap02] deals with the PHY layer in the 5 GHz U-NII bands. Once again, data rates up to 54 Mb/s are specified in these bands with OFDM as the modulation technique. Depending on the PHY layer alternative, the frequency band is divided into several channels. Each channel supports the maximum data rate allowed by that PHY layer alternative. The proposal for a very high rate PHY layer (>100 Mb/s) called 802.11n employs multiple input and output antennas at the transceivers. The technology is popularly called MIMO and also uses OFDM as the modulation scheme.

In the next few sections, the IEEE 802.11 standard is discussed in a top-down manner. First, the different topologies possible in IEEE 802.11 are considered, with the focus on understanding some of the management functions. Then detailed discussions of the MAC layer of 802.11 and different PHY layer alternatives are presented. Once the basic operation of the 802.11 WLAN has been considered, security issues in IEEE 802.11 will be discussed, as also the recent ongoing activities to extend the standard further.

9.2.2 IEEE 802.11 Wireless Local-Area Network Operations

The topology of an IEEE 802.11 WLAN can be one of two types: infrastructure or ad hoc (see Fig. 9.5). In the infrastructure topology, an AP covers a particular area called the basic

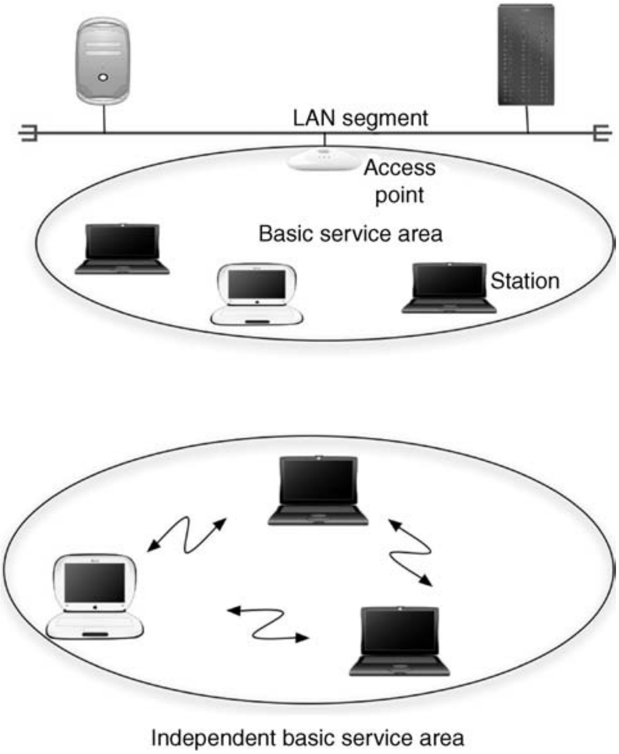


FIGURE 9.5 Topologies in IEEE 802.11.

service area (BSA) and MSs communicate with each other or with the Internet through the AP [Cro97a,Cro97b]. The AP is connected to a LAN segment and forms the *point of access* to the network. All communications go through the AP. So an MS that wants to communicate with another MS first sends the message to the AP. The AP looks at the destination address and sends it to the second MS. The AP and all the MSs associated with it are called a basic service set (BSS). In the ad hoc topology (also called independent BSS or IBSS), MSs that are in range of each other can communicate directly with one another without a wired infrastructure. However, it is not possible for an MS to forward packets meant for another MS not in the range of the source MS. Figure 9.5 shows schematics of both topologies. MSs and APs are identified by a 48-bit MAC address that is similar to other MAC addresses at the link layer. In an infrastructure topology, the MAC address of the AP also forms the BSSID, a unique identifier of the BSS.

If we assume that the range of communication of any WLAN device, be it an MS or an AP, is a region of radius R, we can look at the comparative advantages of the two topologies. An MS can communicate with another MS that is up to $2R$ away using an AP, provided that both MSs are within a distance R of the AP. The cost here is the additional transmission from the AP to the destination. In the ad hoc topology, a destination MS cannot be more than a distance R from the source MS. The advantage is that the information can be received in one hop.

Extending the Coverage in Infrastructure Topology. Depending on the environment in which it is deployed and the transmit powers that are used, an AP can cover a region with a radius anywhere between 30 and 250 feet (9.1–76.2 m). The coverage depends upon radio propagation characteristics in the environment and the antenna features (see Chapter 2). The presence of obstacles such as walls, floors, equipment, and so on can reduce the coverage. Many new 802.11-equipped devices have integrated antennas that additionally reduce coverage. To cover a building or a campus, it often becomes necessary to deploy multiple APs that are connected to the same LAN. A group of such APs and the member MSs is called an *extended service set* (ESS). The coverage area is called the extended service area (ESA). The wired backbone that connects the different APs along with services that enable the ESS is called the *distribution system.* The distribution system, for example, supports roaming between APs so that MSs can access the network over a wider coverage area than before. This is similar to cellular telephone systems, where multiple BSs provide coverage to a region, each BS covering only a cell. Note, however, that cellular telephone systems have a far more complex infrastructure to handle roaming and handoff. In 802.11 WLANs, it is easy to roam within a single LAN and requires support from higher layers (such as mobile IP) to roam across different LANs.

Network Operations in an Infrastructure Topology. When an MS is powered up and configured to operate in an infrastructure topology, it can perform a passive scan or an active scan. In the case of a passive scan, the MS simply scans the different channels to detect the existence of a BSS. The existence of a BSS can be detected through *beacon* frames that are broadcast by APs pseudo-periodically. The reason why it is called pseudo-periodic is that the beacon is supposed to be transmitted regularly at certain intervals. However, the AP cannot preempt an ongoing transmission in order to transmit a beacon. When we discuss the MAC layer, we will see that any device has to wait for the medium to be free before transmitting a frame. If the medium is busy, then the AP will transmit the beacon after the medium becomes free, in which case the beacon may not be precisely periodic. The beacon is a management frame that announces the existence of a network. It contains information about the network – the BSSID and the capabilities of the network (the PHY alternatives it supports, if security is mandatory, whether the MAC layer supports polling, the interval at which beacons are transmitted, timing parameters, and so on). The beacon is similar to certain control channels in cellular telephone systems (for instance, the broadcast control channel, BCCH, in GSM). The MS also performs signal strength measurements on the beacon frame. In the case of an active scan, the MS already knows the ID of the network that it wants to connect to. In this case, the MS sends a *probe request* frame on each channel. APs that hear the probe request respond with a *probe response* frame that is similar in nature to the beacon. In either case, the MS can create a scan report that provides it with information about the available BSSs, their capabilities, their channels, timing parameters, and other information. The MS makes use of this information to determine a compatible network that it can associate itself with.

In order to associate itself with an AP, the MS must authenticate itself if this is part of the capability of the network, and we will look at this in a later section. Otherwise, as long as the MS satisfies the announced capabilities of the network, it can send an *association request* frame to the AP. The association request informs the AP of the intention of the MS to join the network and it also provides additional information about the MS, such as its MAC address, how often it will listen to the beacon (called the *listen interval*), the supported data rates, and so on. If the AP is satisfied with the capabilities of the MS, then it will reply with an *association response* frame. In this message, the MS is given an association ID and this frame confirms that the MS is now able to access the network. During this association phase, the MS can be authenticated by the network and vice versa. Unlike the ad hoc mode of operation, administrators can control access to the network in the infrastructure mode of operation.

If an MS moves across BSSs or if it moves out of coverage and returns to the BSA of an AP, then it will have to reassociate itself with the AP. For this purpose, it will use a *reassociation request* frame similar in form to the association request frame, except that the MAC address of the old AP will be included in the frame. The AP will respond with a *reassociation response* frame. There are three mobility types in IEEE 802.11. The "no transition" type implies that the MS is static or moving within a BSA. A "BSS transition" indicates that the MS moves from one BSS to another within the same ESS. The most general form of mobility is "ESS transition," when the MS moves from one BSS to another BSS that is part of a new ESS. In this case, upper layer connections may break (it will need mobile IP for continuous connection). An MS moving from one BSS to another will have to detect the drop in signal strength from the old AP and detect the beacon of the new AP before the reassociation request. It could also use a probe request message instead of detecting the beacon from the new AP. This simple handoff between two APs is MS initiated. In cellular telephone systems, the MS is instructed by entities in the network (such as BSCs) to perform the handoff from one BS to another. They may use information supplied by the MS, such as the signal strength from different BSs. The handoff procedures in a WLAN are as shown in Fig. 9.6.

One of the important issues in wireless networks is roaming between different points of access to the wired network. This is possible only if equipment from different vendors supports the same set of protocols and is interoperable. Previously, there existed a Task group F that had proposed an inter-AP protocol (IAPP) to achieve multivendor interoperability. For example, when an MS moves from one AP to another and sends a reassociation request, the new AP must be able to converse with the old AP over the distribution system to inform it of the handoff and to free the resources in the old AP. This is achieved using the IAPP, now standardized as 802.11f in July 2003. The 802.11f standard specifies the

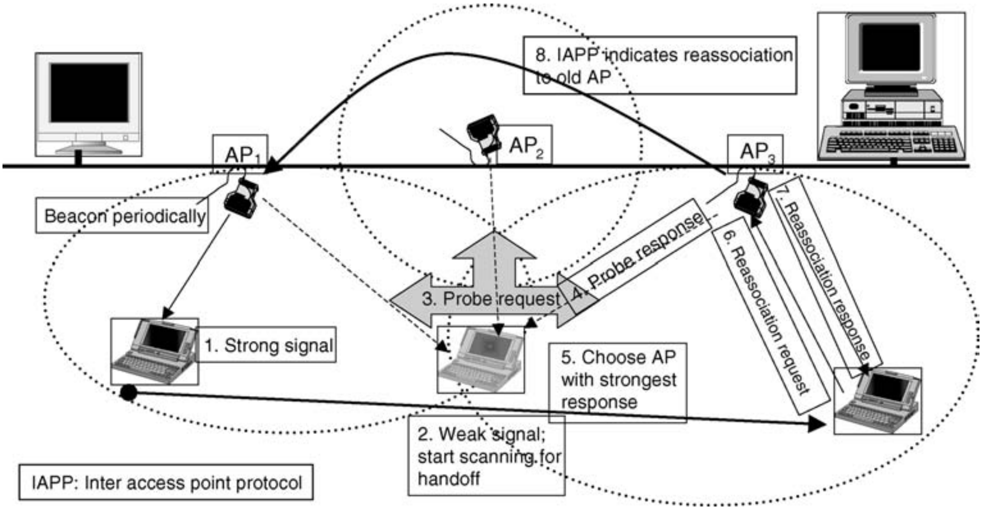

8. IAPP indicates reassociation to old AP
AP1
AP2
AP3
Beacon periodically
3. Probe request
4. Probe response
6. Reassociation request
7. Reassociation response
1. Strong signal
5. Choose AP with strongest response
2. Weak signal; start scanning for handoff
IAPP: Inter access point protocol

FIGURE 9.6 Handoff in a WLAN.

information and format of the information to be exchanged between APs and includes the recommended practice for multivendor AP interoperability via the IAPP across distribution systems.

Power management is an important component of network operations in an IEEE 802.11 WLAN. The idle receive state dominates the LAN adaptor power consumption. The challenge is how we can power off during the idle periods and maintain the session. The IEEE 802.11 solution is to put the MS in sleeping mode, buffer the data at the AP, and send the data when the MS is awakened. Compared with the continuous power control in cellular telephones, this is a solution tailored for bursty data applications. When MSs have no frames to send, they can enter a sleep mode to conserve power. If an MS is sleeping when frames arrive at an AP for it, then the AP will buffer such frames. A sleeping MS wakes up periodically and listens to the beacon frames. How often it wakes up is specified by the listen interval mentioned earlier (Fig. 9.7). The beacon frame also contains a field called the traffic indication map (TIM). This field contains information about whether or not packets are buffered in the AP for a given MS. If an MS detects that it has some frames waiting for it, then it can wake up from the sleep mode and receive those frames before going back to sleep. The MS uses a *power-save poll* frame to indicate to the AP that it is ready to receive buffered frames. The AP sends the buffered data when the station is in active mode.

If the MS chooses to leave the network or shutdown, then it will send a *dissociation* frame to the AP. This frame will terminate the association between the MS and the network,

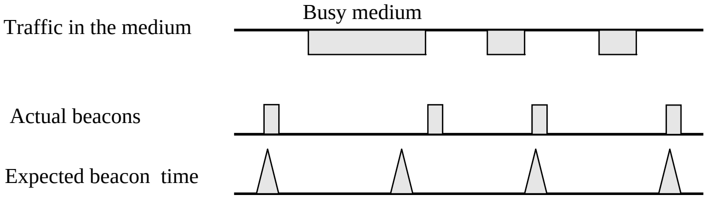


FIGURE 9.7 Power management in IEEE 802.11.

enabling the network to free resources that were previously reserved for the MS), such as the association ID, buffer space, etc.).

Network Operations in an Ad Hoc Topology. In an ad hoc topology, there is no fixed AP to coordinate transmissions and define the BSS. An MS that operates in ad hoc mode will power up and scan the channels to detect beacons from other MSs that may be in the vicinity and that may have set up an IBSS. If it does not detect any beacons, then it may declare its own network. If it does detect a beacon, then the MS can join the IBSS in a manner similar to the process in the infrastructure topology. MSs in an IBSS may choose to rotate the responsibility of transmitting a beacon. Power management works in a similar way, except that the source MS itself has to send an announcement TIM (ATIM) frame to the recipient MS.

Network Operations in Mesh Topology. Recently, wireless mesh networking has received a lot of attention as it enables the deployment of wireless networks over a large area without the need for an extensive fixed (wired) infrastructure. A wireless mesh network consists of entities that connect to each other over an air interface and relay packets in the network thus created. This eliminates the need for a wired backbone to relay packets. Some of these entities may act like APs, creating infrastructure WLANs and becoming points of access to the mesh network. Other entities will connect to the Internet and enable any device in the mesh network to access the Internet. Wireless mesh networks can use a variety of technologies, such as IEEE 802.16 or WiMAX-based devices (see Section 9.3) or IEEE 802.11-based devices. In 2004, a task group “S” of 802.11 was set up to investigate mesh networking with 802.11 and propose a standard for using a *wireless distribution system* unlike the wired distribution system in the ESS. While the standard is not finalized as of the time of writing, some elements of operations in a mesh network have been proposed [Lee06]. In a mesh network, APs (or MSs) are required to be capable of relaying packets to one another using the air interface so that packets may be delivered from a source MS to a destination MS through multiple wireless hops. Entities with relay capabilities are called mesh points. A mechanism for determining a path from one mesh point to another is also necessary and it is expected to be implemented at the MAC layer. Mesh portals enable connectivity to other mesh networks, LANs, or the Internet. Multicast and broadcast capability at the MAC layer is also another aspect that the IEEE 802.11s standard is expected to address. This task group is also supposed to develop enhancements to the current IEEE 802.11 standard to provide a method to configure the distribution system using the four MAC addresses thereby enabling some form of mesh networking between APs. This could be a wired or wireless mesh network that allows for automatic topology learning and dynamic path configuration over self-configuring multi-hop topologies. There are already proprietary protocols performing this task to extend coverage within homes, but the standard will allow for different scenarios with different requirements (e.g. quick setup and tear down, maximizing throughput, etc.).

9.2.3 The IEEE 802.11 Medium Access Control Layer

MSs in an IEEE 802.11 network have to share the transmission medium, which is air. If two MSs transmit at the same time and the transmissions are both in range of the destination, then they may collide, resulting in the frames being lost. The MAC layer is responsible for controlling access to the medium and ensuring that MSs can access the medium in a fair manner with minimal collisions. The medium access mechanism is based on CSMA, but there is no collision detection, unlike the wired equivalent LAN standard (IEEE 802.3). In IEEE 802.3, sensing the channel is very simple. The receiver reads the peak voltage on the wire or cable and compares that against a threshold. Collisions are extremely hard to detect in RF because of the dynamic nature of the channel. Detecting collisions also incurs difficulties in hardware implementation, because an MS has to be transmitting and receiving at the same time. Instead, the strategy adopted is to avoid collisions to the greatest extent possible. In IEEE 802.11, there are two types of carrier sensing: physical sensing of energy in the medium and virtual sensing. Physical sensing is through a clear channel assessment (CCA) signal produced by the physical layer convergence protocol (PLCP) in the physical layer of the IEEE 802.11. The CCA is generated based on "real" sensing of the air interface, either by sensing the detected bits in the air or by checking the RSS of the carrier against a threshold. Decisions based on the detected bits are made slightly slower, but they are more reliable. Decisions based on the RSS may create false alarms caused by high interference levels. The best designs take advantage of both carrier sensing and detected data sensing. In addition to physical sensing, IEEE 802.11 also provides for virtual carrier sensing. Virtual sensing is implemented by decoding a duration field in the 802.11 frame that allows an MS to know the time for which a frame will last. A "length" field in the MAC layer is used to specify the amount of time that must elapse before the medium can be freed. This time is stored in a *network allocation vector* (NAV) that counts down to zero to indicate when the medium is free again. To illustrate the IEEE 802.11 MAC layer, we will use the ad hoc topology as an example. However, the procedures are identical in an infrastructure topology as well.

The Distributed Coordination Function. We will first describe the basic medium access process in IEEE 802.11, called the DCF. Consider Fig. 9.8, which shows the basic method for accessing the medium in IEEE 802.11. An MS will initially sense the channel before transmission. If the medium is free, then the MS will continuously monitor the medium for a period of time called the *DCF IFS* (DIFS). If the medium is still idle after the DIFS, then the MS can transmit its frame without waiting. Otherwise, the MS will enter a back-off process. The rationale is that if another MS senses the medium after the first MS, then it will also wait

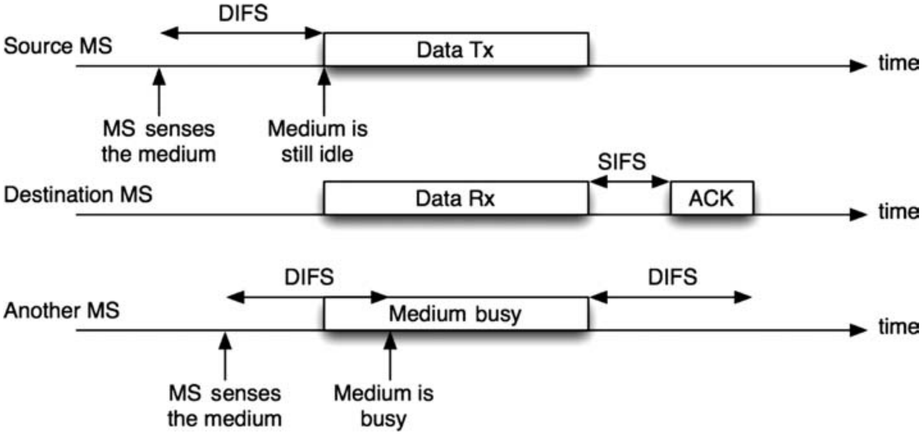


FIGURE 9.8 Basic medium access in IEEE 802.11.

for the DIFS. However, before a time DIFS expires, the first MS would have started its transmission. Upon hearing the transmission, the second MS will have to back off. The wireless medium is harsh and unreliable and hence all transmissions are acknowledged. The destination of the frame will send an acknowledgement (ACK) back to the source if the frame is successfully received as follows. It will wait for a time called the *short IFS* (SIFS) and transmits the ACK. The SIFS value is smaller than the DIFS value. All IFS values depend on the physical layer alternative. Thus, any other MS that senses the channel as idle after the original frame was transmitted will still be waiting and ACK frames have priority over their transmissions. In order to maintain fairness and avoid collisions, the MS that senses the medium as free for a time DIFS and transmits a frame will have to enter the back-off process if it wants to transmit another frame immediately. The exception is when it is transmitting one frame in many fragments. In such a case, the MS can indicate the number of fragments in the first frame to be transmitted and occupy the channel until the frame is completely transmitted.

The back-off process works as follows. Once an MS enters the back-off process, it picks a value called the *back-off interval* (BI) that is a random value uniformly distributed between zero and a number called the CW. The MS will then monitor the medium. When the medium is free for at least a time DIFS, the MS will start counting down from the BI value as long as the medium is free. The counter is decremented every so often (called a slot). If the medium is sensed as occupied before the counter goes down to zero, then the MS will freeze the counter and continue to monitor the medium. As soon the counter becomes zero, the MS can transmit its frame. This process is shown in Fig. 9.9.

The IEEE 802.11 MAC supports *binary exponential back-off* like IEEE 802.3. Initially, the CW is maintained at a value called CW_{min}, which is typically $2^5 - 1 = 31$ slots. So the BI will be uniformly distributed between 0 and 31 slots. The slot time varies depending on the physical layer alternative. For example, it is 20 μs in the IEEE 802.11b standard and 9 μs in the 802.11a standard. If a packet is not successfully transmitted (this could be due to collisions or a channel error), then the value of CW is essentially doubled. The MS will now pick a BI value that is uniformly distributed between 0 and $2^6 - 1 = 63$ slots. This process can be continued until CW reaches a value that is CW_{max} (usually 1023 slots). The rationale behind this approach is as follows. If there are many MSs contending for the medium, then it is likely that one or more MSs may pick the same BI value. Their transmissions will then collide. By increasing the value of CW, it is likely that this probability will go down, thereby reducing collisions.

Frames may be lost due to channel errors or collisions. A positive ACK from the destination is necessary to ensure that the frame has been successfully received. In IEEE 802.11, each MS maintains retry counters that are incremented if no ACKs are received. After a retry threshold is reached, the frame is discarded as being undeliverable.

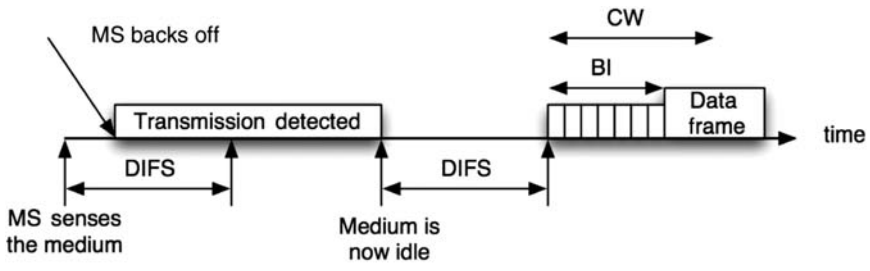


FIGURE 9.9 Back-off process in IEEE 802.11.

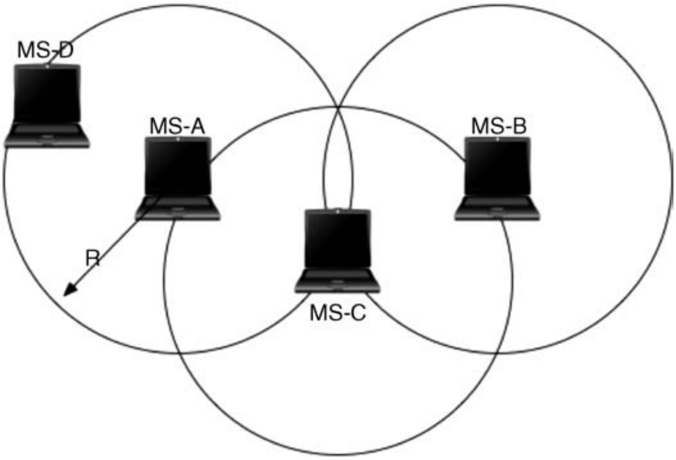


FIGURE 9.10 Illustrating the hidden and exposed terminal problems.

The Hidden Terminal Problem and Optional Mechanism. In wireless networks that use carrier sensing, there is a unique problem called the hidden terminal problem. Suppose all MSs are identical and have a transmission and reception range of *R*, as shown in Fig. 9.10.

The transmission from MS-A can be heard by MS-C but not by MS-B. So, when MS-A is transmitting a frame to MS-C, MS-B will not sense the channel as busy and MS-A is *hidden* from MS-B. If both MS-A and MS-B transmit frames to MS-C at the same time, the frames will collide. This problem is called the hidden terminal problem. There is a dual problem called the exposed terminal problem. In this case, MS-A is transmitting a frame to MS-D. This transmission is heard by MS-C, which then backs off. However, MS-C could have transmitted a frame to MS-B and the two transmissions would not interfere or collide. In this case, MS-A is called an *exposed terminal*. Both hidden and exposed terminals cause a loss of throughput.

To reduce the possibility of collisions due to the hidden terminal problem, the IEEE 802.11 MAC has an optional mechanism at the MAC layer, as shown in Fig. 9.11. Suppose

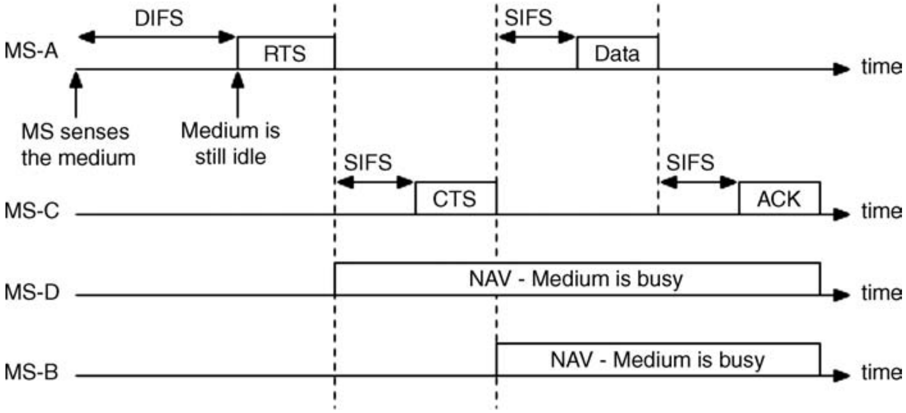


FIGURE 9.11 Operation of the RTS–CTS mechanism.

MS-A wants to transmit a frame to MS-C. It will first transmit a short frame called the RTS frame. The RTS frame is heard in the transmission range of MS-A and includes MS-C and MS-D, but not MS-B. Both MS-C and MS-D are alerted to the fact that MS-A intends to transmit a frame and they will not attempt to simultaneously use the medium. This is achieved by the virtual carrier sensing process that sets the NAV to a value equal to the time it will take to complete the exchange of frames successfully. In response to the RTS frame, MS-C will send a *clear-to-send* (CTS) frame that will be heard by all MSs in its transmission range. This includes MS-B and MS-A but not MS-D. The CTS frame lets MS-A know that MS-C is ready to receive the data frame. It also alerts MS-B to the fact that there will be a transmission from some MS to MS-C. Consequently, MS-B will defer any frames that it wishes to transmit in anticipation of the completion of the communication to MS-C. This way, even though MS-B is outside the transmission range of MS-A, the CTS message can be used to *extend* the carrier sensing range, thereby reducing the hidden terminal problem. Of course, it is quite possible that the RTS frame itself collided with a transmission from MS-B. In such a case, both MS-A and MS-B will have to enter the back-off process and retransmit their frames.

The RTS–CTS mechanism can be controlled in IEEE 802.11 by using an RTS threshold. All unicast and management frames larger than this threshold will always be transmitted using RTS–CTS. By setting this value to 0 bytes, all frames will use RTS–CTS. The default value is 2347 bytes, which disables RTS–CTS for all packets. When RTS–CTS signals are used, the CTS frame is transmitted by the destination MS after waiting simply for a time equal to SIFS. This way, the CTS frame has priority compared with all other transmissions, which have to wait for at least a time DIFS and perhaps an additional waiting time in back-off. Using the RTS–CTS signals reduces the throughput of a WLAN, but it may be essential to use this in dense environments.

The Point Coordination Function. One consequence of using CSMA/CA as described above with DCF is that it is impossible to have any bounds on the delay or jitter suffered by frames. Depending on the traffic load and the BI values that are picked, a frame may be transmitted instantaneously or it may have to be buffered until the medium becomes free. For real-time applications, such as voice or multimedia, this can result in performance degradation, especially when strict delay bounds are necessary. To provide some bounds on the delay, an optional MAC mechanism called the PCF is part of the IEEE 802.11 standard [Cro97a, Cro97b]. The PCF provides contention-free access to frames using a polling mechanism, described below.

The process starts when the AP captures the medium by sending a beacon frame after it is idle for a time called the *PCF IFS* (PIFS). The PIFS is smaller than the DIFS and larger than SIFS. In the beacon frame, the AP, also called the point coordinator, announces a *contention-free period* (CFP) where the usual DCF operation will be preempted. All MSs that use only DCF will set a NAV to indicate that the medium will be busy for the duration of the CFP. The AP maintains a list of MSs that need to be polled during the CFP. MSs get onto the polling list when they first associate with the AP using the association request. The AP then polls each MS on the list for data. The polls are sent after a time SIFS and the ACKs to the poll and any associated data will be transmitted by the corresponding MS also after a time SIFS. If there is no response from an MS to a poll, then the AP waits for a time PIFS before it sends the next poll frame or data. The AP can also send management frames whenever it chooses within the CFP. An example of PCF operations is shown in Fig. 9.12.

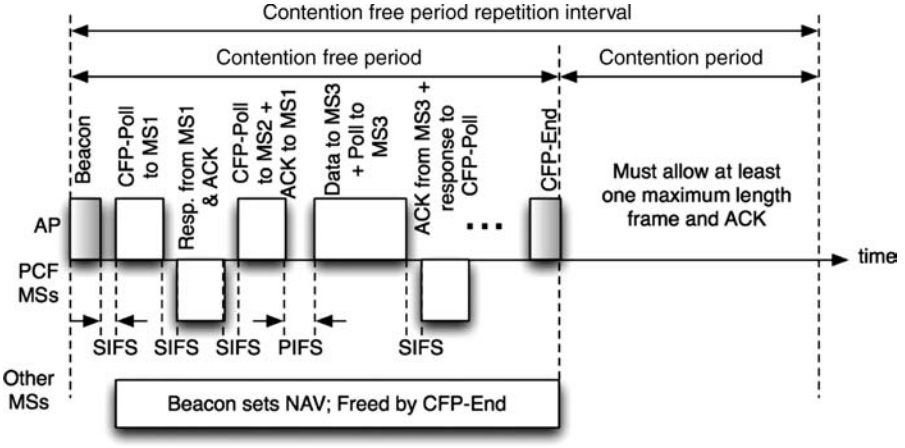

Contention free period repetition interval
Contention free period
Contention period
Beacon
CFP-Poll to MS1
Resp. from MS1 & ACK
CFP-Poll to MS2 + ACK to MS1
Data to MS3 + Poll to MS3
ACK from MS3 + response to CFP-Poll
...
CFP-End
Must allow at least one maximum length frame and ACK
AP
PCF MSs
time
SIFS
SIFS
SIFS
PIFS
SIFS
Other MSs
Beacon sets NAV; Freed by CFP-End

FIGURE 9.12 Operation of the PCF.

The AP indicates the culmination of the CFP via a message called CFP-End. This is a broadcast frame to all MSs and frees the NAV in MSs that are only DCF based. Following the CFP, a contention period starts. In this period, it must be possible for an MS to transmit at least one maximum-length frame using DCF and receive an ACK. The CFP can be resumed after completion of the contention period. The PCF mechanism is optional in IEEE 802.11. Most commercial systems deployed today do not support PCF, and real-time services do not have very good support in WLANs today. Note that polling has a lot of overhead, especially if MSs do not have frames to send when they are polled.

Medium Access Control Frame Formats. While this article will not define all of the different frame formats of an IEEE 802.11 MAC frame and discuss the fields in great detail, it will consider some examples to illustrate the MAC frame formats. Figure 9.13 shows the general format of a MAC frame. The most significant bit is last (rightmost) and the bits are transmitted from left to right. The *frame control* field has two bytes and comprises many fields. It carries information such as the protocol version, the type of frame (management: probe request, association, authentication, and so on; control: RTS, CTS, and so on; or data: pure data, CFP poll and data, null, and so on), the number of retries, and whether the frame is encrypted (discussed later). The duration field is important to set the NAV during virtual carrier sensing.

Frame control 2 bytes	Duration ID 2 bytes	Address 1 6 bytes	Address 2 6 bytes	Address 3 6 bytes	Seq-Ctl 2 bytes

Address 4 6 bytes	Frame body 0-2312 bytes	Frame check sequence 4 bytes

FIGURE 9.13 General format of a MAC frame.

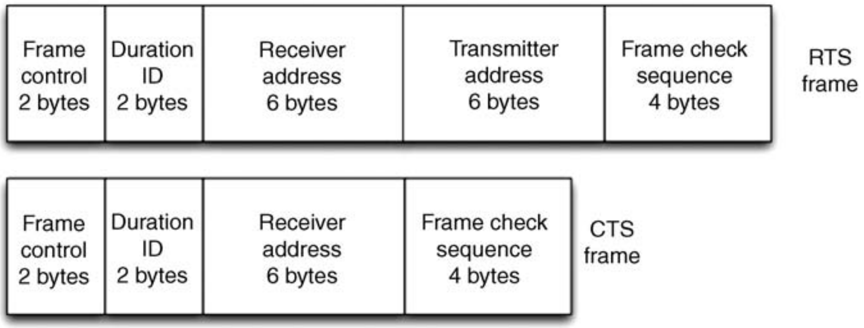
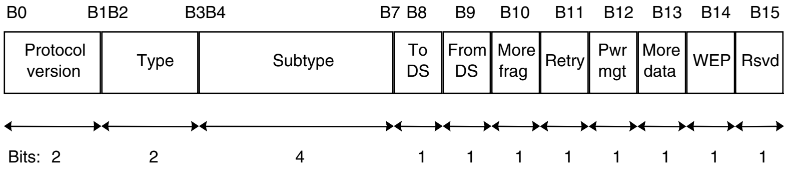


FIGURE 9.14 RTS and CTS frame formats.

There can be up to four address fields in the frame [Gas02]. The addresses can be different, depending on the type of frame. Common addresses used are the source and destination addresses, the receiver address if the destination is different from the receiver (e.g. the receiver is the AP, but the destination is a wired node on the LAN segment), the transmitter address (once again, if the transmitter is different from the source – it is the AP), and the BSSID. The sequence control field is used in case there is fragmentation of frames. The frame body carries the payload from the upper layers and the frame check sequence is a 32-bit CRC used to verify the integrity of the frame at the receiver. The frame format in Fig. 9.13 is used in an infrastructure topology. In an IBSS, only three address fields are used.

The RTS and CTS frames are very short frames, 20 bytes and 14 bytes respectively, and are shown in Fig. 9.14. The ACK frames are very similar to the CTS frame.

Compared with Ethernet, IEEE 802.11 is a wireless network that needs to have control and management signaling to handle registration process, mobility management, power management, and security. To implement these features, the frame format of 802.11 should accommodate a number of instructing packets, similar to those we described in WANs. The capability of implementing these instructions is embedded in the control field of the MAC frames. Figure 9.15 shows the overall format of the control field in the 802.11 MAC frame



Protocol Version: currently 00, other options reserved for future

To DS/from DS: “1” for communication between two APs

More Fragmentation: “1” if another section of a fragment follows

Retry: “1” if packet is retransmitted

Power Management: “1” if station is in sleep mode

Wave Data: “1” more packets to the terminal in power-save mode

Wired Equivalent Privacy: “1” data bits are encrypted

FIGURE 9.15 Details of the frame control field in the MAC header of IEEE 802.11.

TABLE 9.3 Type and Subtype Fields and their Associated Instructions

- Management type (00)
  - Association request/response (0000/0001)
  - Reassociation request/response (0010/0011)
  - Probe–request/response (0100/0101)
  - Beacon (1000)
  - ATIM: announcement TIM (1001)
  - Dissociation (1010)
  - Authentication/deauthentication (1011/1100)
- Control type (01)
  - Power save poll (1010)
  - RTS/CTS (1011/1100)
  - ACK (1101)
  - CF end/CF end with ACK (1110/1111)
- Data type (10)
  - Data/data with CF ACK/no data (0000/0001)
  - Data poll with CF/data poll with CF and ACK (0010/0011)No data/CF ACK (0100/0101)
  - CF poll/CF poll ACK (0101/0110)

with description of all fields except type and subtype. These two fields are very important because they specify various instructions for using the packet. The 2-bit *type* field specifies four options for the frame type:

- management frame (00)
- control frame (01)
- data frame (10)
- unspecified (11).

The 4-bit *subtype* provides an opportunity to define up to 16 instructions for each type of frame. Table 9.3 shows all six bits used for the type and subtypes in the frame control field. Combinations that are not used provide an opportunity to incorporate new features in the future.

Figure 9.16 illustrates the frame body of the beacon frame. The time stamp allows MSs to synchronize to a BSS. The beacon interval says how often the beacon can be expected to be heard. It is typically 100 ms, but could be changed by an administrator. The capability information (2 bytes) provides information about the topology (whether infrastructure or ad hoc), whether encryption is mandatory, and whether additional features are supported. One such feature is *channel agility*, where the AP hops to different channels after a predetermined amount of time.

Time stamp 8 bytes	Beacon interval 2 bytes	Capab. Info. 2 bytes	SSID variable	FH parameter set 7 bytes	DS param set 2 bytes	CF parameter set 8 bytes	IBSS parameter set 4 bytes	TIM var.

FIGURE 9.16 Frame body of the beacon frame.

We have not discussed PHY layer alternatives yet. The parameter sets in the beacon provide information about the PHY layer parameters that are necessary to join the network. For instance, if FH is used, then the FH parameter set will specify the hopping pattern. The TIM field is used to support MSs that may be sleeping, as described earlier.

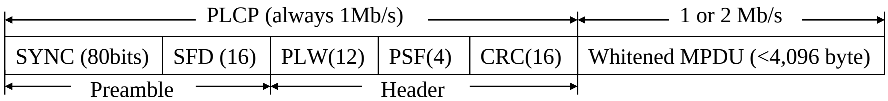
9.2.4 The Physical Layer

The IEEE 802.11 standards body has standardized several different PHY layer alternatives. When it was first standardized in 1997, there were three PHY layer options. We will call these options the "base" IEEE 802.11 PHY layer alternatives. The IEEE 802.11b standard supports up to 11 Mb/s in the 2.4 GHz ISM bands, the IEEE 802.11g standard supports up to 54 Mb/s in the 2.4 GHz ISM bands, and the IEEE 802.11a standard supports up to 54 Mb/s in the 5 GHz U-NII bands. Before we discuss these alternatives, let us look at the PHY layer in IEEE 802.11.

The PHY layer in IEEE 802.11 is broken up into two sublayers: the *PLCP* and the *physical medium-dependent* (PMD) layers. The PLCP includes a function that adapts the underlying medium-dependent capabilities to the MAC-level requirements. The PLCP would, for instance, add some additional fields to the frame to enable synchronization at the physical layer. The PMD layer actually determines how information bits are transmitted over the medium. When the MAC protocol data units (MPDUs) arrive at the PLCP layer a header is attached that is designed specifically for the PMD layer of the choice for transmission. The PLCP packet is then transmitted by the PMD layer according to the specification of the signaling techniques.

The Base IEEE 802.11 Standard. The base IEEE 802.11 standard specifies three different PHY layer alternatives. Two of these use RF transmissions in the 2.4 GHz ISM bands and one uses DFIR.

The Frequency Hopping Option. The first option for transmission in the 2.4 GHz ISM bands makes use of FHSS. The entire band is divided into 1 MHz width channels and the specification makes it important to confine 99% of the energy to one such channel during transmission to reduce interference to the other channels. These restrictions are also due to the rules imposed by the FCC in the USA. The standard specifies 95 such 1 MHz width channels and they are numbered accordingly. In the USA, only 79 of these channels are allowed. Devices that use the FH option hop between these channels when transmitting frames. The dwell time in each channel is approximately 0.4 s or the minimum hop rate of the IEEE 802.11 FHSS system is 2.5 hops per second, which is rather slow. The hop sequences (the channel hopping pattern) depend on mathematical functions. An example hopping pattern is $\{3, 26, 65, 11, 46, 19, 74, \ldots\}$. In the USA, each set of hopping patterns can have at most 26 different channels. This means that it is possible to create three orthogonal hopping sets (since there are 79 channels in the USA). If three APs use these three orthogonal hopping sets, then there will be no interference between these networks. The modulation scheme used with FHSS is called Gaussian frequency-shift keying (GFSK). This modulation scheme makes use of the frequency information to encode data. It is possible to use either two frequencies within the channel or four frequencies within the channel. In the former case, the data rate will be 1 Mb/s; in the latter case, the data rate will be 2 Mb/s. The advantage of the FHSS system is that the receivers are less complex to implement. The PLCP for the FHSS PMD layer introduces an 80-bit field for synchronization, a frame delimiter, and some fields to indicate the data rate. Depending on this field, the



SYNC: Alternating 0,1
SFD: 0000110010111101
PLW: Packet length width
PSF: Data rate in 500Kb/s steps
CRC: PLCP header coding

FIGURE 9.17 PLCP frame for the FH option in IEEE 802.11.

data rate can be modified in steps of 500 kb/s from 1 Mb/s to 4.5 Mb/s. However, the standard only supports 1 and 2 Mb/s. The values of SIFS and the slot for back-off in this option are 28 μs and 50 μs respectively.

Figure 9.17 shows the details of the PLCP header, which is added to the whitened MPDU to prepare it for transmission using FHSS physical layer specifications of the IEEE 802.11. The PLCP additional bits consist of a preamble and a header. The *preamble* is a sequence of alternating 0 and 1 symbols for the 80 bits that are used to extract the received clock for carrier and bit synchronization. The start of the frame delimiter (SFD) is a specific pattern of 16 bits, shown in the figure, indicating start of the frame. The next part of the PLCP is the header, which has three fields. The 12-bit packet-length width (PLW) field identifies the length of the packet, which could be up to 4 kbytes. The four bits of the packet-signaling field (PSF) identifies the data rate in 0.5 Mb/s steps starting with 1 Mb/s.

Example 9.3: Specification of Data Rate on the Physical Layer The existing 1 Mb/s is represented by 0000 as the first step. The 2 Mb/s is represented by 0010, which is $2 \times 0.5\,\text{Mb/s} + 1\,\text{Mb/s} = 2\,\text{Mb/s}$. The maximum 3-bit number represented by this system is 0111, which is associated with $7 \times 0.5 + 1 = 4.5\,\text{Mb/s}$. If all four bits are used then we have $15 \times 0.5 + 1 = 8.5\,\text{Mb/s}$. These limitations imply that data rates cannot even reach 10 Mb/s.

The rest of the rates are reserved for the future. The 16-bit CRC code is added to protect the PLCP bits. It can recover from errors of up to two bits and otherwise identify whether the PLCP bits are corrupted or not. The total overhead of the PCLP is 16 bytes (128 bits), which is less than 0.4% of the maximum MPDU load, justifying the low impact of running the PCLP at lower data rates. The received MPDU is passed through a scrambler to be randomized. Randomization of the transmitted bits, which is also called whitening because the spectrum of a random signal is flat, eliminated the DC bias of the received signal. A scrambler is a simple shift register finite -state machine with special feedback that is used both for scrambling and de-scrambling of the transmitted bits.

The Direct Sequence Option. The DSSS modulation technique has been the most popular commercial implementation of IEEE 802.11. DSSS has some inherent advantages in multipath channels and can increase the coverage of an AP for this reason [Tuc91]. We will briefly discuss the features of this PMD layer.

In a DSSS system, the data stream is “chipped” into several narrower pulses (chips), thereby increasing the occupied spectrum of the transmitted signal. One common way of doing this is to multiply the data stream (typically a series of positive and negative rectangular pulses) by a spreading signal (typically another series of positive and negative

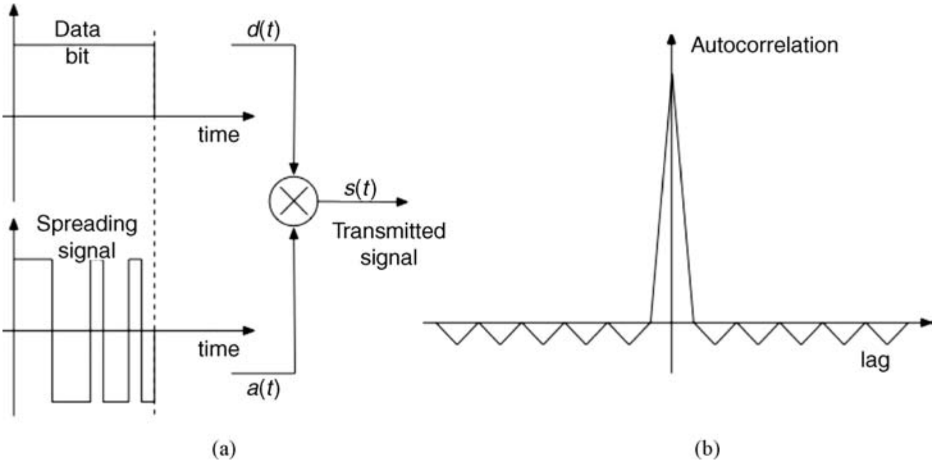


FIGURE 9.18 (*a*) DSSS and (*b*) autocorrelation of the Barker pulse.

rectangular pulses, but with much narrower pulses than the data stream). While the data stream is random and depends on what needs to be transmitted, the spreading signal is deterministic. Figure 9.18*a* shows an example where the data stream $d(t)$ is multiplied by a spreading signal $a(t)$ to produce a signal $s(t)$ that is then modulated over an RF carrier. In this figure, 11 narrow pulses are contained within one broad data pulse. The pulses could have a positive (+) or negative (−) amplitude. This results in the bandwidth expanding by a factor of 11, and this is also called the *processing gain.* A specific pattern of pulses in the spreading signal is used. The pattern used in the IEEE 802.11 standard is a Barker sequence. The interesting property of the Barker sequence is that its autocorrelation has a very sharp peak and very narrow side lobes, as shown in Fig. 9.18*b*. Because of this property, it is possible for a receiver to reject interference from multipath signals and recover information robustly in a harsh wireless environment. The Barker sequence with differential BPSK (DBPSK) is used for data rates of 1 Mb/s and the Barker sequence with differential QPSK (DQPSK) is used for data rates of 2 Mb/s. In either case, the chip rate is 11 Mc/s (megachips per second).

Unlike FHSS, a signal carrying 2 Mb/s now occupies a bandwidth that is as large as 25 MHz. In the IEEE 802.11 standard, 14 channels are specified for the DSSS PMD layer. Channel 1 is at 2.412 GHz, Channel 2 at 2.417 GHz, and so on (see Fig. 9.19). Only the first 11 channels are available for use in the USA. Figure 9.19 shows the channelization in

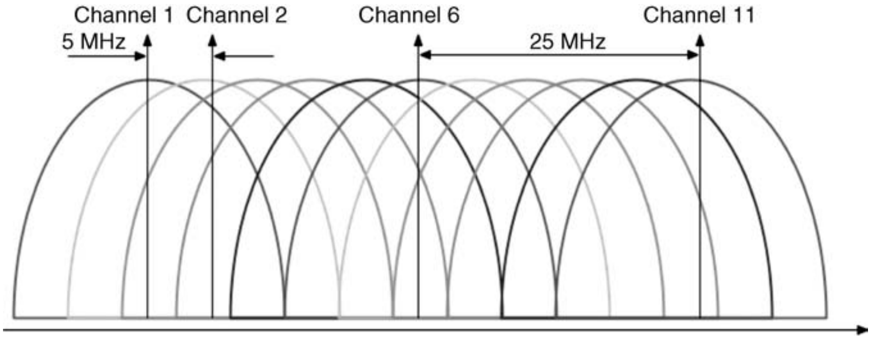
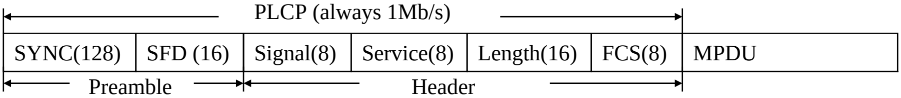


FIGURE 9.19 Channelization for the IEEE 802.11 DS option.



SYNC: Alternating 0,1
SFD: 1111001110100000
Signal: Data rate in 100kHz steps
Service: reserved for future use
Length: Length of MPDU in microseconds
FCS: PLCP header coding

FIGURE 9.20 PLCP frame for the DSSS of the IEEE 802.11.

the USA. Since each channel occupies roughly 25 MHz bandwidth and the channel separation is only 5 MHz, there is significant overlap between channels. If two WLANs in the same vicinity were to use adjacent channels, then there would be severe interference and throughput degradation. There are three *orthogonal* channels, Channels 1, 6 and 11, in the USA that can be deployed without interference. The FH option is easier in terms of implementation because the sampling rate is on the order of the symbol rate of 1 MS/s. The DS implementation requires sampling rates on the order of 11 Mc/s. However, because of the wider bandwidth, DSSS provides a better coverage and a more stable signal.

Figure 9.20 shows the details of the PLCP frame for the DSSS version of the IEEE 802.11. The overall format is similar to the FHSS, but the length of the fields is different because transmission techniques are different and different manufacturers designed the model product for development of the FHSS and DSSS standards. The PLCP sublayer once again introduces some fields for synchronization (128 bits), frame delimiting, and error checking. The PLCP header and preamble are always transmitted at 1 Mb/s using DBPSK. The rest of the packet is transmitted using either DBPSK or DQPSK, depending on the data rate. The values of SIFS and the slot for back-off in this option are 10 μs and 20 μs respectively. The MPDU from the MAC layer is transmitted either at 1 or 2 Mb/s; however, analogous to the FHSS version of the standard, the PLCP of the DSSS version also uses the simpler BPSK modulation at 1 Mb/s all the time. The MPDU for the DSSS does not need to be scrambled for whitening because each bit is transmitted as a set of random chips that is a whitened transmitted signal. The length of the SYNC in the DSSS is 128 bits, which is longer than FHSS because DSSS needs a longer time to synchronize. The format of the SFD of the DSSS is identical to that of the FHSS, but the value of the code, shown in Fig. 9.20, is different. The PSF of the FHSS is called the *signal* field and it uses 8 bits to identify data rates in steps of 100 kb/s (five times more precision than FHSS).

Example 9.4: Frame Formats for Various Data Rates in IEEE 802.11 Using the above encoding, we represent 1 Mb/s for DSSS by 00001010 (10 × 100 kb/s) and 2 Mb/s by 00010100 (20 × 100 kb/s), and 11 Mb/s (used in IEEE 802.11b) by 001101110 (55 × 100 kb/s) and 01101110 (110 × 100 kb/s). The maximum number in this system is 11111111, which represents 255 × 100 kb/s = 25.5 Mb/s.

The *service* field in the DSSS is reserved for future use and it does not exist in the FHSS version. The *length* field of the DSSS is analogous to the PLW in the FHSS; however, the length field specifies the length of the MPDU in microseconds. The frame correction sequence (FCS) field of the DSSS is identical to the CRC field of the FHSS.

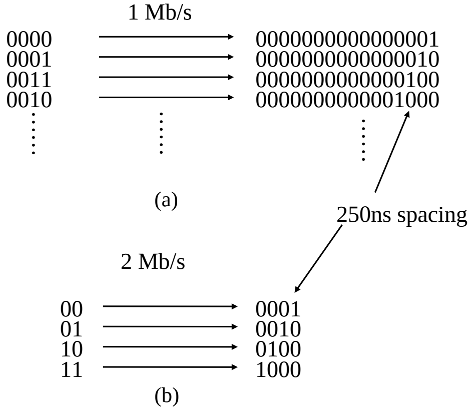


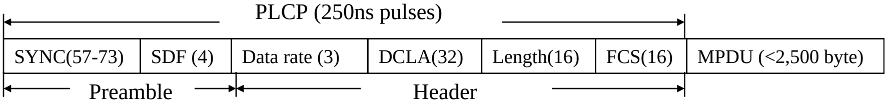
FIGURE 9.21 PPM using 250 ns pulses in the DFIR version of the IEEE 802.11: (a) 16-PPM for 1 Mb/s; (b) 4-PAM for 2 Mb/s.

The Diffuse Infrared Option. The third option in IEEE 802.11 is to use IR for transmission [Val98]. The spectrum occupied by the IR transmission is at wavelengths between 850 and 950 nm. The technique used for transmission is DFIR – that is, communications are omnidirectional. The range specified is around 20 m, but the transmissions cannot penetrate physical obstacles. The modulation scheme used is PPM. A data rate of 1 Mb/s is supported using 16-PPM and a data rate of 2 Mb/s is supported using 4-PPM. This is in comparison with the IrDA (Infrared Data Association) standard, which primarily allows communications at a few hundred kilobits per second to a few megabits per second between two devices (like a laptop and a personal digital assistant) that are within a few feet of one another.

The PMD layer of DFIR operates based on transmission of 250 ns pulses that are generated by switching the transmitter LEDs on and off for the duration of the pulse. Figure 9.21 illustrates the 16-PPM and 4-PPM techniques recommended by the IEEE 802.11 for 1 Mb/s and 2 Mb/s respectively. In 16-PPM, blocks of 4 bits of the information are coded to occupy one of the 16 slots of a 16-bit length sequence according to their value. In this format, each $16 \times 250\,\text{ns} = 4000\,\text{ns}$ carries 4 bits of information, which supports a $4\,\text{bits}/4000\,\text{ns} = 1\,\text{Mb/s}$ transmission rate. For the 2 Mb/s version, every 2 bits are pulse position modulated into four slots of duration $4 \times 250\,\text{ns} = 1000\,\text{ns}$, which generates data at $2\,\text{bits}/1000\,\text{ns} = 2\,\text{Mb/s}$. The peak transmitted optical power is specified at 2 W with an average of 125 or 250 mW.

The PLCP packet format for DFIR is shown in Fig. 9.22. The PLCP signals are shown in the unit of slots of 250 ns for one basic pulse. The SYNC and SDF fields are shorter because noncoherent detection using photosensitive diode detectors do not need carrier recovery or elaborate random code synchronizations. The three-slot data rate indication system starts by 000 for 1 Mb/s and 001 for 2 Mb/s. The Length and FCS are identical to the DSSS. The only new field is the DC level adjustment (DCLA) that sends a sequence of 32 slots allowing the receiver to set its level of the received signal to a set threshold for deciding between received 0 s and 1 s. The MPDU length is restricted to 2500 bytes.

IEEE 802.11b and 802.11g. The DS option for IEEE 802.11, although successful, consumed a lot of bandwidth for the given data rate. The chip rate is 11 Mc/s, but the maximum data rate is 2 Mb/s. That is, one Barker Sequence of 11 chips, transmitted every microsecond, can at most carry 2 bits of information. To increase the data rate, the IEEE



SYNC: Alternating 0,1 pulses
SFD: 1001
Data rate: 000 and 001 for
DCLA: DC level adjustment sequences
Length: length of MPDU in microseconds
FCS: PLCP header coding

FIGURE 9.22 PLCP frame for the DFIR of the IEEE 802.11.

802.11b standard adopted a slightly different method. Instead of transmitting one 11-chip sequence every microsecond, with IEEE 802.11b, the device transmits one 8-chip codeword every 0.727 μs. Each 8-chip codeword can carry up to 8 bits of information for a maximum data rate of $8/(0.727 \times 10^{-6}) = 11$ Mb/s. If the codeword carries only 4 bits of information, then the data rate will be 5.5 Mb/s. The codewords are derived from a technique called CCK [Hal99].

CCK works as follows for the case when 8 bits are mapped into an 8-chip codeword. The incoming data stream is broken up into units of 8 bits. Suppose the least significant bit is labeled d0 and the most significant bit is labeled d7. Then, four phases are defined to correspond to the four possible values of a pair of bits, as shown in the first two columns of Table 9.4. Depending on what the bits are, the phases then take on a value as shown in the third and fourth columns in Table 9.4. For example, if d5 = 0 and d4 = 1, then the phase $\varphi_3 = \pi$. Once the phases are determined, the 8-chip codeword is given by the vector

$$C = \{e^{\mathrm{j}(\varphi_1+\varphi_2+\varphi_3+\varphi_4)},\quad e^{\mathrm{j}(\varphi_1+\varphi_3+\varphi_4)},\quad e^{\mathrm{j}(\varphi_1+\varphi_2+\varphi_4)},\quad -e^{\mathrm{j}(\varphi_1+\varphi_4)},\quad e^{\mathrm{j}(\varphi_1+\varphi_2+\varphi_3)},$$
$$e^{j(\varphi_1+\varphi_3)}\quad -e^{j(\varphi_1+\varphi_2)},\quad e^{j(\varphi_1)}\}$$

This vector has elements that belong to the set $\{+1, -1, +\mathrm{j}, -\mathrm{j}\}$, where j is the square root of -1. These four elements can be mapped in RF to the phase of the carrier, and the receiver can decode this phase information to recover the data bits. CCK can be though of either as a modulation scheme or as a coding scheme. All the terms of the above equation share the first phase, if we factor that out we have

$$\mathbf{C} = \{e^{\mathrm{j}(\varphi_2+\varphi_3+\varphi_4)},\quad e^{\mathrm{j}(\varphi_3+\varphi_4)},\quad e^{\mathrm{j}(\varphi_2+\varphi_4)},\quad -e^{\mathrm{j}(\varphi_4)},\quad e^{\mathrm{j}(\varphi_2+\varphi_3)},$$
$$e^{\mathrm{j}(\varphi_3)}\quad -e^{\mathrm{j}(\varphi_2)},\quad 1\}\, e^{\mathrm{j}(\varphi_1)}$$

TABLE 9.4 Mapping for CCK

Di-bit	Phase parameter	Di-bit (d_{i+1}, d_i)	Phase
(d_1, d_0)	φ_1	(0, 0)	0
(d_3, d_2)	φ_2	(0, 1)	π
(d_5, d_4)	φ_3	(1, 0)	π/2
(d_7, d_6)	φ_4	(1, 1)	−π/2

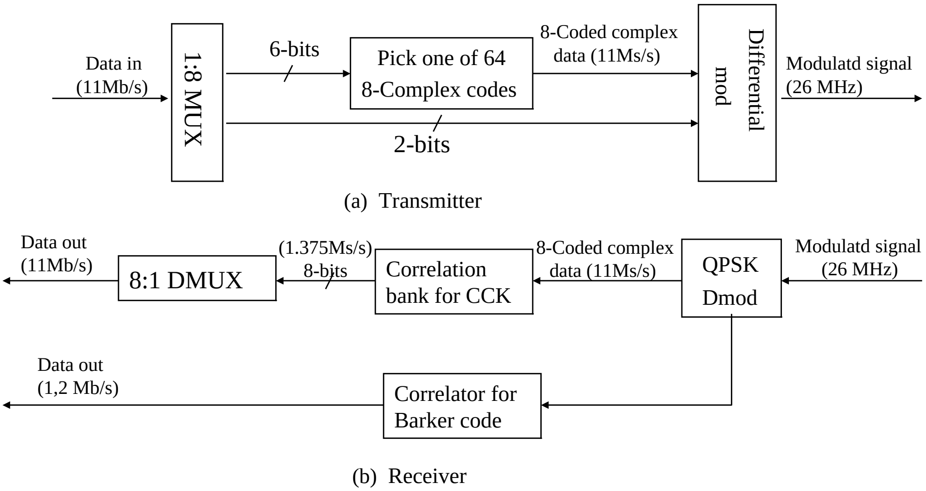


FIGURE 9.23 Simplified implementation of the CCK for IEEE 802.11b.

This coding suggests that our 256 transformation matrix can be decomposed into two transformations: one a unity transformation that maps 2 bits (one complex phase) directly and the other one that maps 6 bits (three phases) into an eight-element complex vector with 64 possibilities determined by the inner function of the above equation. The above decomposition leads to a simplified implementation of the CCK system that is shown in Fig. 9.23. At the transmitter the serial data in multiplied into 8-bit addresses. Of the 8 bits, 6 bits are used to select one of the 64 orthogonal codes produced as one of the 8-complex code and 2 bits are directly modulated over all elements of the code that are transmitted sequentially. The receiver actually comprises two parts: one the standard IEEE 802.11 DSSS decoder using Barker codes and one a decoder with 64 correlators for the orthogonal codes and an ordinary demodulator for IEEE 802.11b. By checking the PLCP data rate field, the receiver knows which decoder should be employed for the received packets. This scheme provides an environment for implementation of a WLAN that accommodates both 802.11 and 802.11b devices.

The advantage of CCK is that it maintains the channelization of IEEE 802.11 while increasing the data rate by a factor of 5. CCK is also fairly robust to the degradations caused by multipath in the wireless environment. The values of SIFS and the slot for back-off in this option are 10 µs and 20 µs respectively. IEEE 802.11b also has an optional modulation method called packet binary convolutional coding (PBCC) that is not widely implemented. The advantage of PBCC over CCK is the use of powerful convolutional coding for FEC.

The IEEE 802.11g standard [Vas05] maintains backwards compatibility with IEEE 802.11b and IEEE 802.11 DS options by adopting minimal PHY layer frame changes and by including some mandatory and optional physical layer components. In the PLCP layer, 802.11g allows for the use of short preambles to reduce packet overhead. The four physical layers specified in this standard are prefixed by the term ERP, which stands for *extended rate physical*. The standard specifies OFDM and CCK as the mandatory modulation schemes with a data rate of 24 Mb/s as the maximum mandatory data rate. With OFDM, IEEE 802.11g also provides for optional higher data rates of 36, 48, and 54 Mb/s. OFDM is the same modulation scheme that is used in IEEE 802.11a, and we discuss it below. PBCC is an optional modulation scheme in 802.11g that allows for raw data rates of 22 and 33 Mb/s. We discuss OFDM in the context of 802.11a below, but the discussion could also apply to 802.11g with some modifications.

The IEEE 802.11a Standard. One of the primary problems for huge data rates in wireless channels is what is called the coherence bandwidth of the wireless channel caused by multipath dispersion. The coherence bandwidth limits the maximum data rate of the channel to that which can be supported within this bandwidth (for example, if the coherence bandwidth is B Hz and the channel bandwidth $W \gg B$ Hz, a transmission bandwidth of W Hz will result in irrecoverable errors unless equalization or spread spectrum is used). In order to overcome this limitation, we can send data in several sub-channels each on the order of the coherence bandwidth or less, so that many of them will get through correctly. Using several sub-channels and reducing the data rate on each channel increases the symbol duration in each channel. If the symbol duration in each channel is larger than the multipath dispersion, then errors will be smaller and it will be possible to support larger data rates. This principle can be exploited while maintaining bandwidth efficiency using a fairly old technique called OFDM. OFDM has been used in DSLs as well to overcome the variations in attenuation with frequency over copper lines. OFDM enables spacing carriers (sub-channels) as closely as possible and implementing the system completely in digital, eliminating analog components to the greatest extent possible. OFDM is used as the physical layer in IEEE 802.11a, HIPERLAN/2, and IEEE 802.11g [Kap02].

IEEE 802.11a specifies eight 20 MHz channels [vNee99]. As shown in Fig. 9.24, several sub-channels are created in OFDM using orthogonal carriers in each channel: 52 sub-channels are specified for each channel, with a bandwidth of approximately 300 kHz each, 48 sub-channels are used for data transmission, and four are used as pilot channels for synchronization. One OFDM symbol (consisting of the sum of the symbols on all carriers) lasts for 4 μs and carries anywhere between 48 and 288 coded bits. For example, at 54 Mb/s, the OFDM symbol has 216 data symbols. With a code rate of 3/4, the number of coded bits/symbol will be 4 × 216/3 = 288. This is possible by using different modulation schemes – ranging from BPSK, where we have 1 bit per sub-channel, to more complex modulation

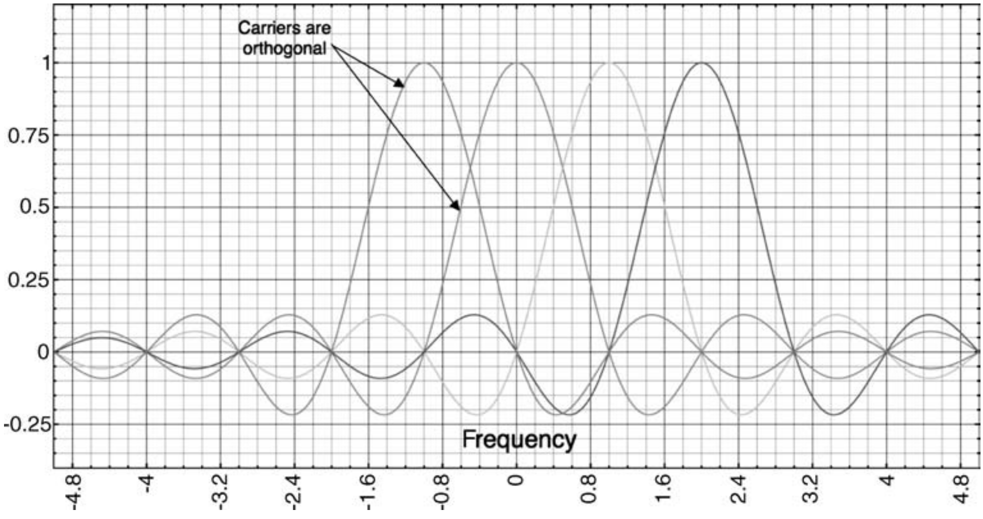


FIGURE 9.24 Orthogonal carriers in OFDM.

TABLE 9.5 Data Rates and Associated Parameters in IEEE 802.11a

Data rate (Mb/s)	Modulation	Code rate	Data bits/ symbol	Coded bits/ sub-channel	PLCP rate field
6	BPSK	1/2	24	1	1101
9	BPSK	3/4	36	1	1111
12	QPSK	1/2	48	2	0101
18	QPSK	3/4	72	2	0111
24	16-QAM	1/2	96	4	1001
36	16-QAM	3/4	144	4	1011
48	64-QAM	2/3	192	6	0001
54	64-QAM	3/4	216	6	0011

schemes like QAM. Error control coding also plays an important role in determining the data rate. Table 9.5 summarizes some features of the different supported data rates.

The PLCP in the case of 802.11a is a bit different, in that there is no synchronization field. A rate field with 4 bits indicates the data rate that is being transmitted. This field is shown in Table 9.5 for different data rates. The preamble and header are always modulated using BPSK (lower data rates). The values of SIFS and the slot for back-off in this option are 16 µs and 9 µs respectively.

The IEEE 802.11n Standard. The current MAC and PHY layers of the IEEE 802.11 standard constrain the raw data rate to 54 Mb/s and the throughput to a fraction of that, depending on traffic load, channel conditions, and so on. A new task group is looking at an IEEE 802.11n standard that will look at both MAC and PHY enhancements to improve the throughput to more than 100 Mb/s (to up to 600 Mb/s). Note that this is not the raw data rate on the air, but the actual throughput of the network. Some of the ideas being floated to improve throughput are to use directional antennas, MIMO with OFDM, and throughput enhancements at the MAC layer. MIMO enables the spectral efficiency of links to go well above the 1 bit/(s Hz) that is usually the order in traditional systems. This increase in spectral efficiency is possible through the use of *space–time techniques* (see Chapter 3), such as STC, beamforming, and spatial multiplexing. These techniques either increase the reliability of the link through diversity or increase capacity by canceling interference from the simultaneous transmission of multiple data streams from multiple antennas.

The primary MAC enhancement to improve throughput in 802.11n is the use of frame aggregation to reduce overhead [Xia05]. At very high data rates, the overhead of waiting times, back-off, and frame headers can reduce throughput significantly. One method of reducing this overhead is to aggregate frames, either at the MAC or PHY layer. Similarly, acknowledgments can also be delayed and aggregated. Frame aggregation can be used for single destinations, multiple destinations, and use multiple rates for multiple destinations.

For some time, there were two competing proposals in Task Group N for the PHY and MAC layers, namely WWiSE (World-Wide Spectrum Efficiency) and TGnSync, with many vendors in each group. Both proposals used MIMO at the physical layer. Products conforming to parts of these proposals were also available in the market. Neither of the proposals was successful in obtaining 75% of the vote. Recently (in January 2006), these two proposals merged with a third proposal that has been now submitted for approval. The industry consortium driving this new proposal is called the *Enhanced Wireless Consortium* (EWC). The goal of this organization is to “help accelerate the IEEE 802.11n development

TABLE 9.6 Summary of PHY Alternatives in IEEE 802.11

Standard	Spectrum, USA	Data rates (Mb/s)	Modulation scheme
Base IEEE 802.11	2.402–2.479 GHz	1, 2	GFSK, FHSS
	2.402–2.479 GHz	1, 2	B/QPSK, DSSS
	850–950 nm	1, 2	PPM, IR
802.11a	5.15–5.35, 5.725–5.825 GHz	6–54	OFDM
802.11b	2.402–2.479 GHz	1, 2, 5.5, 11	CCK
802.11g	2.402–2.479 GHz	1–54	OFDM, CCK
802.11n	2.4 and 5 GHz	> 100	MIMO/OFDM

process and promote a technology specification for interoperability of next-generation WLAN products.”

Summary of Physical Layer Alternatives. Table 9.6 summarizes the different PHY layer alternatives in IEEE 802.11.

9.2.5 Deployment of Wireless Local-Area Networks

When WLAN standardization activities started in late 1980s the main issue related to installation was the selection of topology, not large-scale cellular deployment. At the time, WLANs were perceived as an extension to wired LANs that avoided wiring challenges. With the popularity of Internet access starting in the mid 1990s, the corresponding growth of the WLAN industry, and thoughts on integrating WLANs into 3G cellular systems, more attention was paid to large-scale WLAN deployment for wide-area coverage. Since commercially successful cellular telephone networks were deployed very carefully based on relatively accurate channel models for path loss, research efforts in systematic deployment of WLANs in large areas using automated coverage prediction software attracted attention at that time [For95].

However, in practice, WLANs were installed in small areas, such as residential homes or small shops, in a random manner by users in the most convenient location where the WLAN AP (or router) could connect to the backbone Internet connection points such as a cable or DSL modem. In large-area applications, such as wireless mobile access inside a large office building or a warehouse area, or outdoor deployments for wireless Internet access, WLANs were deployed in grid formation [Unb02]. APs are installed every 20–30 m inside a building in convenient locations, such as corridors and large open areas where larger traffic was expected. In outdoor applications, APs are installed in much wider grids on top of utility posts, high on the outside walls of multifloor buildings, or on the roof of buildings to optimize coverage. The grid installation, deployed by building owners or independent service providers (for instance, in the hallways of shopping malls), provides excellent coverage with large overlap between adjacent cells and a very low outage probability. The basic difference between user deployment and grid installation is that, in the grid installation, some primitive network planning by visual inspection, measurements of the RSS in selected locations, or a study of the construction drawings of the building is performed to provide for a better coordination among the locations of the APs. In user deployments, the AP positions are mainly decided based on the convenience of installation. Figure 9.25 illustrates these differences. This figure also shows *optimal* deployment, which uses

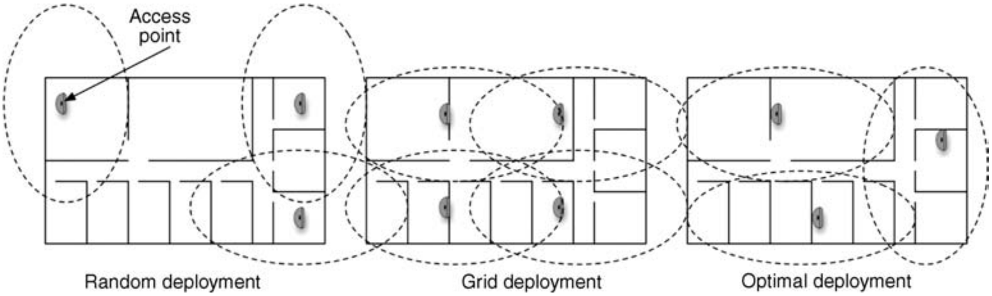


FIGURE 9.25 Random, grid, and optimal deployment of large-scale WLANs.

software tools to determine the best places for deploying APs (perhaps minimizing interference or optimizing for user loads).

For outdoor grid installations in a street in a typical downtown or industrial area, we need specialized technicians and coordination with the city and utility companies for the installation of APs on the tops of lampposts. In addition, extensive wiring would be needed to connect the APs to a backbone network. The physical installation using a coverage optimization technique will look very similar to the grid installation.

Differences between Deployment of Wireless Local-Area Networks and Cellular Networks. Next, we consider the differences among issues related to deployment of WLANs and cellular telephone networks.

WLANs operate in unlicensed bands, whereas cellular telephone networks use licensed bands. As we discussed in Chapters 5 and 7, the network capacity in FDMA- and TDMA-based cellular networks depends on the frequency reuse factor during deployment, which is determined by calculations of the interference from neighboring cells. In the case of CDMA networks, again we had a parameter for the calculation of capacity that was related to interference from neighboring cells, which again relates capacity to interference. All these calculations are based on the assumption that the band is licensed to one operator, which technically means that the network planner has control on the interference. WLANs operate in unlicensed bands in which a network planner does not have control on the interference. Network managers in a university campus or a corporation, to control interference, may restrict students or employees in the deployment of WLANs other than those owned by the university or the corporation, which does not comply with the government regulations on using these bands and can be considered illegal.

For optimal deployment, cellular networks use relatively accurate statistical coverage prediction models, such as the Okumura–Hata model discussed in Chapter 2; statistical channel models for coverage in indoor areas are much less accurate, and this poses a challenge in the analysis of coverage of WLANs unless we resort to labor-intensive empirical measurements or use computationally intensive ray-tracing algorithms.

APs for WLANs are very inexpensive, around a few hundred dollars or less at the time of writing, and they do not need expensive antenna towers and site landscape for installation. BSs are orders of magnitude more expensive than APs, and for large area coverage (macrocells) they need expensive towers and appropriate land for cell sites. Radio resource management in WLANs is much simpler because WLANs have limited numbers of nonoverlapping frequency bands (for example, three nonoverlapping channels in the case of the popular 802.11g standard); cellular networks have an order of magnitude more channels to handle. Mobility management for cellular networks is much more complex than WLANs because most popular WLAN applications are quasi-stationary, whereas cellular networks were designed to operate inside a vehicle. In addition, traditional WLAN data traffic is in bursts, it is nonsymmetric and location- and time-selective; therefore, the MAC is not performed centrally through radio resource management techniques. As a result of all of these differences, an 802.11 AP is simple and connects to the Internet backbone directly through a wired LAN or cable/DSL modem, whereas a cellular BS is connected to the PSTN using a hierarchy that includes a BSC and an MSC. The BSC and MSC are needed because, in cellular networks, we have more complex radio resource management, mobility management, and connection management techniques. Cells are larger, voice connections need more quality control, and cellular telephones have higher mobility.

It is true that at the time of writing the dominant bands for using WLANs are the 2.4 GHz ISM bands. However, as the popularity of WLANs increases, many researchers envision that at some point this industry will resort to higher frequencies where wider portions of bandwidth are available. The path of migration is first to 5 GHz and then to tens of gigahertz (around 60 GHz) and perhaps to the IR domain. Cellular planning for higher frequencies would involve other challenges. In some situations, the desired signal is attenuated by obstacles, whereas an interfering signal arrives unobstructed, and there is a consequent degradation in the carrier to interference ratio (denoted by C/I). This situation may prevent the design of a logical layout for cells in an indoor area. However, if the signal is contained in one room and does not penetrate the walls of the room, then the walls can be used to define the boundaries of the cells. This situation exists for infrared and microwave (above 20 GHz) WLANs, wherein each room constitutes one cell of the network. A more quantitative example will further clarify this situation.

Example 9.5: WLAN Coverage in Different Environments A systematic comparative performance evaluation of random, grid, and coverage optimization techniques for deployment of WLANs in office buildings, shopping centers, and campus areas at 5, 17 and 60 GHz frequencies is analyzed by Unbehaun [Unb02,Unb03]. In this work, ray-tracing software was used to generate channel profiles in different environments. Figure 9.26 shows the sample building layouts for an office, a shopping mall, and a campus area used by the ray-tracing software. Considering different APs densities, the received SNR in different location is calculated to form the cumulative distribution function of the SNR for various conditions. Figure 9.27 shows two sets of cumulative distribution functions at 5 GHz in a manufacturing floor. Each set consists of user deployment, grid installation, and optimal network planning installation for 0.1 and 0.15 APs per 1000 m^2. Table 9.7 provides the 10th percentile point for the density function of the SNR in different environments and 5, 17, and 60 GHz. Since 17 GHz has coverage difficulties in shopping areas and even 60 GHz is not suitable, unlike 17 GHz operation in open areas, these values are not included in the table.

In general, office buildings provide a simple environment for WLAN deployment. For 2.4 and 5 GHz systems, since the signal penetrates through walls, the coverage is relatively unproblematic and a few APs opportunistically installed in convenient areas, such as corridors, sitting areas, and large lecture halls, typically cover a floor of a building. At 17 and 60 GHz, propagation is mostly limited to LOS operation and essentially one AP per room is needed. Since the size of the cells is very small and more bandwidth at higher frequencies is available, the throughput per user can be increased well beyond those achieved with 2.4 or 5 GHz systems. This additional throughput is at the expense of higher infrastructure cost. In office

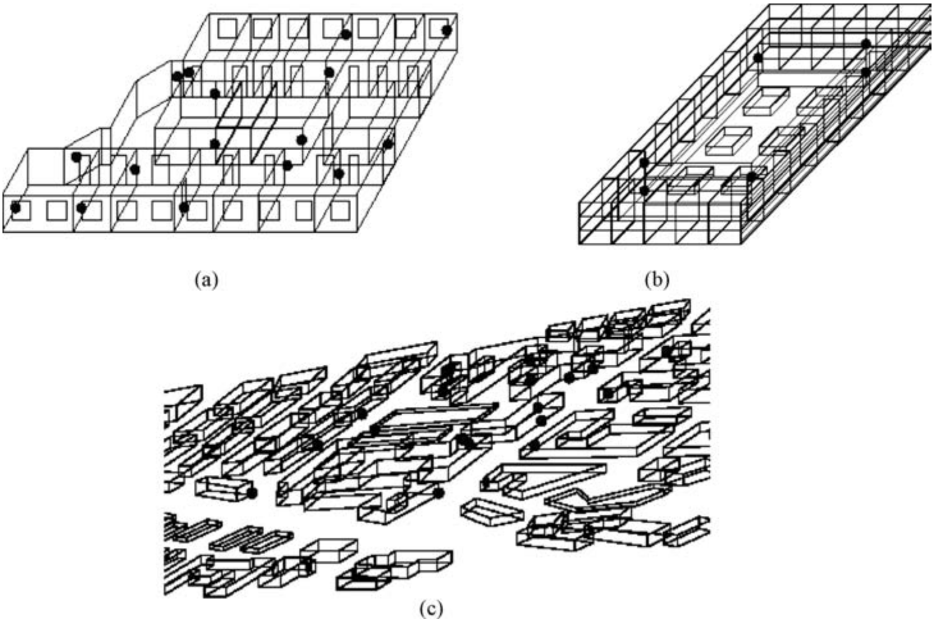


FIGURE 9.26 Typical building layouts used for the performance analysis of WLAN using ray-tracing software in (*a*) an office. (*b*) a shopping mall, and (*c*) a campus area [Unb02].

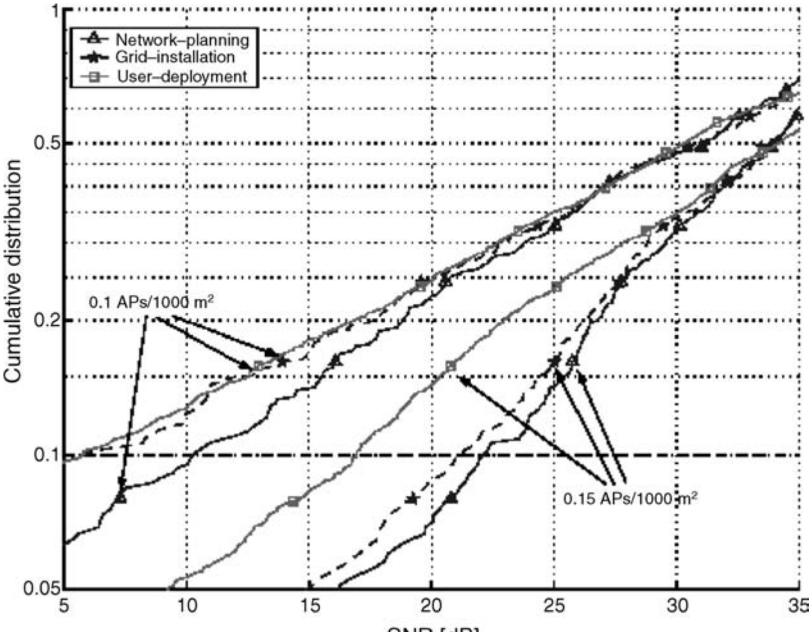


FIGURE 9.27 Cumulative distribution function of the received SNR at 5 GHz in a shopping area [Unb02].

TABLE 9.7 Required 10th Percentile SNR in Different Environments and Different Frequencies

Deployment method	Frequency (GHz)	Minimum AP density/1000 m^2	10th percentile SNR (dB)
Office environment			
User deployment	5	1.85	14.6
Grid installation	5	1.85	17.5
Coverage optimization	5	1.85	21.2
User deployment	17	14.8	16.1
Grid installation	17	7.4	16.9
Coverage optimization	17	7.4	19.9
User deployment	60	74	18.0
Grid installation	60	52	15.2
Coverage optimization	60	52	14.9
Shopping mall environment			
User deployment	5	0.5	21.3
Grid installation	5	0.17	19.0
Coverage optimization	5	0.17	22.2
User deployment	17	6	16.2
Grid installation	17	3	16.0
Coverage optimization	17	3	18.5
Campus environment			
User deployment	5	0.15	14.2
Grid installation	5	0.15	21.0
Coverage optimization	5	0.15	22.0

areas, the deployment methodology is not very critical, and performing sophisticated network planning or coverage optimization does not result in substantial performance gain. At such high frequencies, deployment must be dense; and in fact, user deployment can outperform grid installation or coverage optimization. In the campus environment, coverage with an acceptable infrastructure density can be provided if a 2.4 or a 5 GHz system is used. The small cell sizes at 17 and 60 GHz basically preclude an outdoor installation of such networks. The deployment method does in fact have a significant influence on the system capacity in the campus environment, and some form of network planning should preferably be performed. The shopping mall is a problematic and very complex environment, both with respect to providing coverage and regarding interference issues. The potentially very large user populations in such environments cause not only strong shadowing, but also result in a relatively low average throughput per user. The strong fragmentation of the environment and the fact that the network layout needs to be done in three dimensions (the shopping mall comprises multiple floors) makes it very difficult to devise a reasonably good network plan that provides adequate coverage. Coverage optimization, and in some cases also grid installation, can somewhat improve the performance. However, the capacity is mainly limited by the intense interference, which also fluctuates heavily due to the strong fragmentation of the environment. The potential improvements by using sophisticated network planning are, however, limited.

Capacity of Infrastructure Wireless Local-Area Networks. As a mobile user moves inside the coverage area of a cell, the received SNR changes and can be modeled as a random variable. Figure 9.27, for example, illustrates the statistical behavior of the SNR as an MS is located in different locations of a shopping mall. For the traditional circuit-switched voice applications, each user has a single data rate independent of its received SNR and we calculate the capacity in terms of the number of available voice channels in a cell or in the entire network, which is a function of total available bandwidth (see Chapter 5). All of the modern wireless data applications are multirate systems for which the data rate is adjusted depending on the received SNR and packet losses. Examples of multirate WLAN and mobile data systems are IEEE 802.11a/g and HDR services (see Section 7.7). In multirate systems, each MS can operate at one of the multiple choices of the data rates according to the value of its received SNR. In a single-user environment, the data rate of an MS is the average of all data rates it operates at while moving in the area. In a multiple-user environment, the average data rate per MS is also a function of the MAC technique and the number of MSs in the area. In the rest of this section we provide a framework for understanding the capacity of a wireless data network regardless of its MAC.

In wireless data applications, a multirate MS has its highest data rate at short distances from the AP or BS. As the distance increases, the RSS, and consequently the SNR, reduces until a point where the necessary SNR for the highest data rate is not available and the modem has to be switched to the next lower data rate. As the distance continues to increase, the data rate continues to fall to lower rates until the signal strength falls below the coverage of the AP or the BS at the lowest allowed rate. In an infrastructure WLAN with multiple APs providing a comprehensive coverage, we expect that there is another AP or BS to connect to when the signal falls below one of the lower thresholds. Therefore, if we consider an area that is covered by a wireless data service, the data rate available to the user has a spatial distribution in which associated with any location is one of the available multiple rates of the system. In other words, the data rate in a random location in the area of the coverage forms a *discrete random variable*. One way to define a capacity for this multirate system with a statistical data rate is to define the *spatial capacity* as the *average* of the data rates that a user randomly located in the area of the coverage observes. With this definition, the spatial capacity will be given by

$$R_{\mathrm{av}} = \sum_{n=1}^{N} p_n R_n \tag{9.1}$$

where R_{av} is the average spatial data rate, R_n is one of the available multirates, and p_n is the probability of occurrence of that data rate, which is the ratio of the areas in which we have that specific data rate to overall area of the coverage of the AP or the BS.

Example 9.6: Spatial Capacity of IEEE 802.11b IEEE 802.11b supports four data rates, namely 11 Mb/s, 5.5 Mb/s, 2 Mb/s, and 1 Mb/s. In a semi-open indoor area these data rates can be used up to distances of 50 m, 70 m, 90 m, and 115 m respectively. Figure 9.28*a* shows the area of coverage and circles around an AP that provide different data rates. If a terminal is located randomly in the area of coverage, then the probability of being in each of the areas is given by the ratio of the area for the specific data rate to the total coverage area. That is, $p_n = A_i/\pi r^2$, where r is the radius of the largest circle (115 m). Figure 9.28*b* shows the data rate, distance coverage, annular area, and the probability of having a certain data rate. Figure 9.28*c* shows the PDF of the data rates calculated from the ratio of the coverage area for a data rate and the overall coverage area. If we substitute the data rates and their probabilities from the density function in Fig. 9.28 into Eq. (9.1), then the average data rate or the spatial capacity of the AP is 2.584 Mb/s, which is well below the expected 11 Mb/s.

Data rate (Mb/s)	*Coverage distance (m)*	*Area of coverage (m^2)*	$p_n = \frac{A_i}{\pi D_4^2}$
R_1=11	D_1=50	A_1=7850	0.19
R_2=5.5	D_2=70	A_2=7536	0.18
R_3=2	D_3=90	A_3=10048	0.24
R_4=1	D_4=115	A_4=16092	0.39

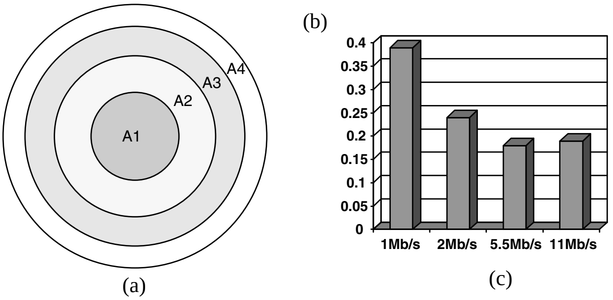


FIGURE 9.28 Data rates and coverage areas for the IEEE 802.11b in a semi-open indoor area: (a) coverage area for different data rates; (b) calculation of probabilities for data rates; (c) PDF of the data rates.

The above scenario provides a worst-case condition for an infrastructure WLAN. In actual deployment, network designers try to have overlapping cells and overlaps are at the low data rates, reducing their contribution to the average data rate. The upper bound for spatial capacity in a cellular deployment is achieved when the adjacent APs are close enough that the user always observes the highest data rate (11 Mb/s in Example 9.6). For the popular grid installation, as shown in Fig. 9.29, the minimum distance between the APs (size of the grid) to support the maximum spatial diversity is $r_g = \sqrt{2}r_1$, where r_1 is the radius of the circle where the maximum data rate (50 m in Example 9.6) is available. For this situation, the PDF of the data rates for the system is only an impulse at 11 Mb/s with a probability of one. As the grid distance between the APs increases beyond r_g the PDF of the data rates starts to include lower data rates, resulting in decreases in the spatially averaged data rate. As mentioned earlier, if we have no coverage gaps, then the minimum spatial throughput cannot be less than the spatial throughput of a single AP (2.584 Mb/s for the 802.11b in Example 9.6). In our discussions in this chapter we addressed single-floor indoor environments. For some discussions on multifloor environments the reader can refer to Hills [Hil01].

9.2.6 Security Issues and Implementation in IEEE 802.11

Security in wireless networks is an important problem, especially because it is extremely difficult to contain radio signals within a protected perimeter [Edn04]. Anyone can listen to radio signals and anyone can also potentially inject signals into the network. Typically, in any network, wireless or wired, it is common to deploy security features or services like confidentiality, entity authentication, data authentication and integrity, and so on to protect against security threats [Sti02]. The IEEE 802.11 standard has some mechanisms to provide confidentiality, integrity, and authentication at the link level. All data that leaves the 802.11 link will not be protected. For instance, an MS communicating with an AP can have all its

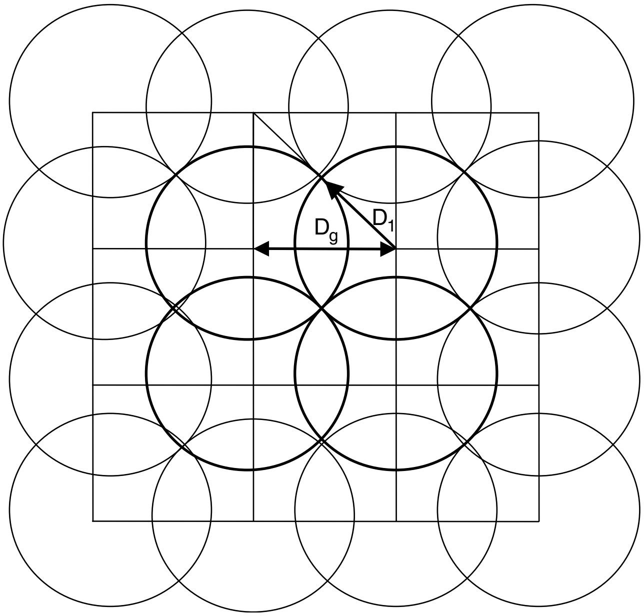


FIGURE 9.29 Grid deployment with optimal spatial capacity.

IEEE 802.11 frames that are on the air protected. Once the AP receives the frame, all protection is removed before it is transmitted on the distribution system. So, additional security at the higher layers (such as IPSec or the secure sockets layer (SSL)) may be required for some applications if the payload needs to be secure.

The original mechanism for providing confidentiality and authentication in IEEE 802.11 is called *wired equivalent privacy* (WEP) [Gas02]. Over the last few years, several techniques for compromising WEP have been published in the literature. Tools such as AirCrack, Kismet, and WEPcrack are freely available that can be used to extract the secret key used in WEP encryption. WEP makes use of the RC4 stream cipher with 40-bit keys (although there are options to use 128-bit keys in most commercial products today). Both the implementation of WEP and the RC4 algorithm itself have vulnerabilities that have rendered WEP not secure for today's applications. WEP was initially proposed in the standard as a self-synchronizing, exportable, efficient option. While it does satisfy these three properties, its security has left much to be desired.

In what follows we discuss the original implementation of security in IEEE 802.11 and discuss enhancements that have been recently proposed.

Entity Authentication in IEEE 802.11. The mandatory entity authentication mechanism in IEEE 802.11 is called *open-system authentication*. In this case, there is no real authentication. If one IEEE 802.11 device sends a frame to another, then it is implicitly accepted. For example, an MS may simply send a frame to the AP choosing "open system" as the authentication algorithm (authentication algorithm = 0). The AP will simply accept it if open-system access is allowed and send a response. From this transaction, the AP will obtain the MAC address of the MS for communication purposes.

A better authentication procedure is called *shared-key authentication*, where WEP is implemented [Gas02]. If the network is using WEP, shared-key authentication is mandatory.

The assumption is that all devices in the network share a secret key. An MS will send a frame for authentication with sequence number 0, the authentication algorithm set to 1 (to indicate shared-key authentication). The AP will then send a challenge message (128 bits) in clear text to the MS along with its response. The MS will respond with an encrypted version of the challenge text. If the AP is able to verify the integrity of the reply, then the MS is authenticated and it has the shared WEP key configured in it. Sometimes an MS will authenticate itself with several APs before associating itself with one of them. This process is called preauthentication. This authentication scheme is still not very secure, however, and also creates weaknesses in the protocol due to the way in which it is employed with a stream cipher.

Several commercial products also implement address filtering where only certain MAC addresses are allowed access to the network. This is not part of the standard, and it is also possible for malicious users to spoof MAC addresses easily. However, address filtering is an additional security measure that is available for IEEE 802.11 networks.

Confidentiality and Integrity in IEEE 802.11. Confidentiality is simply provided in IEEE 802.11 by encrypting all packets using the RC4 stream cipher. Stream ciphers operate as follows. Using a secret key, a pseudo-random sequence of bits (called the key stream) is generated. If this sequence has a very long period and the algorithm is strong, then it will be computationally impossible for someone to generate the sequence without knowing the secret key.

A pseudorandom generator is used along with the 40-bit secret key to create a key sequence that is simply XOR-ed with the plain-text message. The pseudorandom sequence thus generated will be XOR-ed with the MAC frame to make the contents of the frame secure from interception. RC4 is one algorithm to generate the pseudorandom key stream. This algorithm makes use of a secret key and in the case of WEP an initialization vector (IV) that is 24 bits long. Since the key is constant for all transactions, the same pseudorandom key stream is generated if the IV is not changed. An attacker could capture two streams of encrypted frames, XOR them together and eliminate the key stream. They would then have an XOR of two data frames. If by some chance they know the contents of one data frame, then they can get the other as well. Since the IV is only 24 bits long, it is possible for an attacker to break the encryption scheme. One well-publicized attack is the Fluhrer, Mantin, Shamir (FMS) attack on RC4. In addition, there are several weak keys that could make the encryption scheme easier to break. It is also possible for an attacker to replay packets depending on the sequence numbers that are being used. In order to ensure that an attacker has not modified a message, the WEP protocol uses the in-built CRC to verify the integrity of the message. Checking the integrity of the message using the CRC has vulnerabilities that have been publicized in recent years.

Key Distribution in IEEE 802.11. The IEEE 802.11 standard does not specify how the shared keys must be distributed to devices (AP and MSs). It is usually a manual installation of keys where a user will type the key in the device driver software. This process is unfortunately not scalable and also has several human vulnerabilities. Users may write down the key on a piece of paper when they buy a new device and lose this paper. Some vendors have automated methods of key distribution. Cisco's Light Extensible Authentication Protocol makes use of the challenge–response mechanism to generate a key at the AP and an identical matching key locally in the MS that could then be used in a successful encrypted communication.

Security Features in 802.11i. The Task Group I of the IEEE 802.11 working group has prepared an enhanced security framework for IEEE 802.11 called 802.11i that was approved as a standard in June 2004. Several vendors have already implemented elements of this standard. This framework includes what is called a *robust security network* (RSN) that is similar to WEP, but has several new capabilities in devices [Edn04]. It is possible for both WEP and RSN devices to coexist in a *transitional security network* (TSN).

A consortium of major WLAN manufacturers called the Wi-Fi Alliance considered options to improve security in legacy devices while 802.11i was being standardized. The proposal from this alliance is called Wi-Fi protected access (WPA), which introduces an enhancement to WEP called the *temporal key integrity protocol* (TKIP). In this protocol, RC4 is still used as the encryption algorithm. However, this protocol adds some features to overcome the weakness of WEP. A message integrity code is used instead of the CRC check. This changes the way in which IVs are generated. It changes the encryption key for every frame, increases the size of the IV, and also adds a mechanism to manage keys.

In 802.11i, RC4 is replaced by the advanced encryption standard (AES). In particular, the key stream and message integrity check will be generated by a *c*ounter-mode *c*ipher-clock-chaining *M*AC *p*rotocol (CCMP). AES is a block cipher; it operates on fixed blocks of data, unlike a stream cipher that generates a key stream. However, any block cipher can operate in different *modes*, and cipher-block-chaining (CBC) is one such mode of operation. The counter mode is another mode of operation. It is expected that the counter mode will be used to generate the key stream and the CBC will be used to generate the message integrity check. Both these modes have been used in other systems with good security.

Both TKIP and AES-CCMP provide confidentiality and message integrity. In order to perform entity authentication, the IEEE 802.11 system still has to rely on challenge–response protocols. Over the years, there have been several protocols developed for dial-up entity authentication and for port security in wired LANs. These include 802.1X, the extensible authentication protocol (EAP) and remote authentication dial-in user service (RADIUS). Note that all these protocols are not equivalent. For instance, both 802.1X and RADIUS could use EAP for entity authentication and key distribution. EAP itself would use some challenge–response protocol like the challenge handshake authentication protocol (CHAP) or SSL to authenticate the devices. Both WPA and RSN mandate 802.1X and EAP as part of the access control mechanism for 802.11 networks. Note that access control is increasingly becoming an important problem with the emergence of hot-spot networks in airports, cafes, and so on.

9.2.7 Wireless Local-Area Networks Standards and 802.11 Standards Activities

Standards activities for WLANs have evolved geographically with ETSI's broadband radio access network (BRAN) working on HIPERLAN standards in Europe and the IEEE on the 802.11 series in the USA. In Japan, the Multimedia Mobile Access Communications Promotion Council (MMAC-PC) working group under the Association of Radio Industries and Businesses (ARIB) works on WLAN standards. The spectrum used by WLANs is mostly unlicensed spectrum, although some systems use licensed spectrum as well (e.g. HIPERLAN in some licensed bands). The only commercially successful WLAN standard has been the IEEE 802.11. HIPERLAN/1 was standardized in the mid 1990s and supported complex multi-hop ad hoc networking. The medium access mechanism in HIPERLAN/1 [Wil95] was based on a form of carrier sense multiple

TABLE 9.8 Summary of some WLAN Standards

Standard	Standard body	Spectrum	Data rate (Mb/s)	Primary medium access	Primary region
IEEE 802.11,b, g	IEEE	2.4 GHz ISM bands	1, 2, 5.5, 11, up to 54	CSMA/CA	North America
IEEE 802.11a	IEEE	5 GHz U-NII and ISM bands	Up to 54	CSMA/CA	North America
IEEE 802.11n[a]	IEEE	2.4 GHz ISM bands	>100	CSMA/CA	North America
HIPERLAN/1	ETSI	5 GHz bands	23	EY-NPMA	Europe
HIPERLAN/2	ETSI	5 GHz bands	Up to 54	TDMA reservation	Europe
Wireless access and WLAN	MMAC	3–60 GHz bands	20–25	Various	Japan

[a] The IEEE 802.11n standard is yet to be finalized.

access called elimination-yield non-preemptive multiple access (EY-NPMA). HIPERLAN/2 has adopted a physical layer that is very similar to IEEE 802.11a and a medium access mechanism based on reservation and TDMA. However, neither of these standards has been adopted successfully in commercial products. The MMAC-PC activities include a variety of wireless access networks (some compatible with the IEEE 802.11 standards) ranging from WPANs to outdoor fixed public networks. Recently, there have been efforts to harmonize the standards activities of ETSI, IEEE 802.11 and MMAC-PC. Table 9.8 summarizes some of the 802.11, HIPERLAN, and MMAC-PC standards.

As mentioned earlier, the competing standard for WLANs in Europe is the HIPERLAN/2 standard that is specified for the U-NII bands. HIPERLAN/2 uses the same physical layer as IEEE 802.11a, although it accommodates a few different data rates. In this standard, there are mechanisms suggested for power measurement and control and radio resources management. Previously, there existed Task Group H, which was enhancing the current 802.11 MAC and 802.11a PHY with network management and control extensions for spectrum and transmit power management in the U-NII bands with the possibility of dynamic channel selection capabilities. In this case, APs would be able to select channels dynamically based on information they can obtain about neighboring APs that may transmitting on the same channel. This way, a laborious network planning process could be simplified. The 802.11h standard, completed in October 2003, considers such transmit power control (TPC) and dynamic frequency selection (DFS) to satisfy the regulatory aspects in Europe.

Outside of the standards that have been specified already in this chapter, there are several ongoing activities in the IEEE 802.11 working group. There are several task groups that are engaged in enhancing aspects of the IEEE 802.11 standard. Some of these are as follows. *Task Group J* is considering enhancements to the current standard to provide operations in the frequency band between 4.9 and 5 GHz for use in Japan. The reason for this task group is to make changes in the 802.11 standard to accommodate the regulatory demands in this spectrum that exist in Japan. There will be expected changes to both the MAC and PHY layers to meet these regulations. *Task Group K* is looking at enhanced radio resource management outside the purview of 802.11h. The primary goal of this group is to provide mechanisms to higher layers that enable radio and network measurements as necessary. Once power measurements and reporting are possible in a standardized manner, they can be exploited to make better use of the spectrum, reduce interference, and so on. Future road transportation is expected to evolve into an ITS. The goal of *Task Group P* (called wireless access for the vehicular environment – WAVE) is to define enhancements to 802.11 that may be required to support ITS applications. Such definitions will include data exchange between high-speed vehicles and between these vehicles and the roadside infrastructure in the licensed ITS band (5.9 GHz). The newly formed Task Group *Y* is looking at operation in the 3.5 GHz bands that were recently opened up by the FCC for WLAN operation. One of the objectives of this protocol is to develop a fair contention protocol for access to the medium.

While not directly changing or adding to the standards, there are task groups that are involved in maintenance and other issues related to 802.11. Task Group M is performing the maintenance of the 802.11 standard. Task group D is looking at regulatory domain updates. There are also some new proposed activities that are pending approval at various levels as of the time of writing: an 802.11t that aims to recommend practices for wireless performance prediction, an 802.11u that is considering interworking with external networks, an 802.11v for network management of MSs, and an 802.11w that provides protection (data integrity and authentication) of management frames. In addition, there are several study groups looking at harmonizing the 802.11 and the ETSI standards and some investigating the possibility of improvements to the 802.11 standard to provide higher throughput.

9.3 IEEE 802.16 (WiMAX)

In the late 1980s, when the first generation of WLANs appeared in the market, work stations and PCs were dominant in computer communications, but laptops had not yet gain popularity. As a result, the first-generation WLANs were designed as boxes with a size close to a shoebox. One of these boxes connected a cluster of workstations and PCs in the vicinity of each other inside a building and the other box was connected to the Ethernet. This setting provided a wireless network connection to all computers in the cluster. The incentive for the design of these boxes was to avoid wiring difficulties involved in buildings with numerous terminals and limited walls to snake the LAN wiring into them. The market anticipation for these wire-replacement products did not meet the enthusiasm of the designers for emergence of this innovative technology. A few tens of start-up companies and small groups in large companies who were designing this first generation of WLANs went into financial trouble in the early 1990s and started to think of a new marketing strategy to sell their product. Motorola's Altair product, operating in licensed 18–19 GHz bands, was one of the first to add an external antenna to the boxes and use them as point-to-multipoint outdoor devices to bring network access to remote buildings. Examples of these applications were connecting the main campus building of a university to satellite buildings, such as students' dormitories. These applications were successful enough to generate adequate income for some of these financially troubled start-ups or groups in large companies to carry on the operation for a few more years. Eventually, some of these companies designed PCMCIA cards for the emerging market

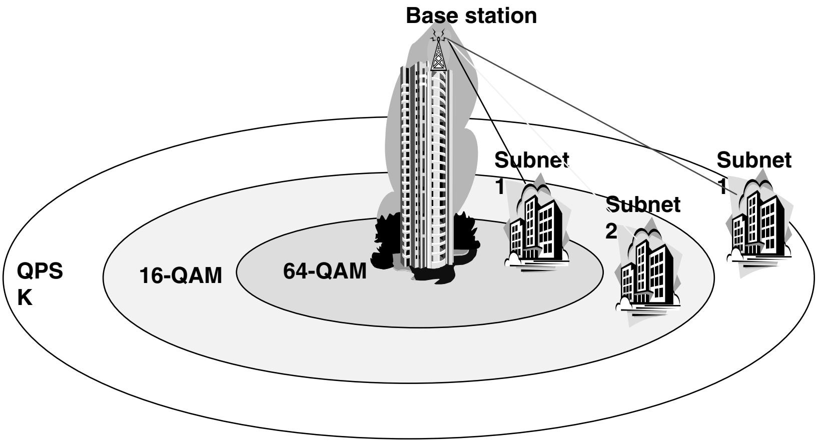


FIGURE 9.30 WiMax PHY and coverage.

of laptop computers which generated sizable income for some of the older surviving companies and attracted new companies to prosper in the WLAN industry in the late 1990s.

In 1998, the idea of a standard for point-to-multipoint outdoor applications, also referred to as a WMAN, was initiated in a National Institute of Standards and Technology (NIST) meeting and it was welcomed by the IEEE, resulting in initiation of the IEEE 802.16 standardization committee in 1999. The charter of this committee was to define an air interface standard for fixed and mobile broadband wireless systems using a point-to-multipoint design and/or mesh technology. The mesh technology portion of the standard was later moved to IEEE 802.20 working on mobile broadband wireless access (MBWA) as a separate standard. The first IEEE 802.16 standard for a point-to-multipoint WMAN was for operation in 10–66 GHz and was approved in 2001. This standard supported data rates up to 134 Mb/s using plain QPSK modulation, 16-QAM, and 64-QAM to cover 1–3 miles in LOS conditions. Figure 9.30 shows the relation between the coverage and modulation technique complexity and the application of point-to-multipoint networking. The original IEEE 802.16 standard, also called LMDS, did not turn out to be a commercially successful product. However, it attracted considerable attention from the cellular network equipment manufacturers.

Later on, in 2001, the WiMAX forum industrial group was formed to improve 802.16 and to promote conformance and interoperability of this standard. Today, the same way that IEEE 802.11 is also called WiFi, IEEE 802.16 is also called WiMAX technology. The WiMAX forum considers this technology as a competitor to cable and DSL technologies to provide for last-mile broadband access to the Internet. This movement resulted in the revival of the 802.16 standard. First, 802.16a was ratified in 2003 as an amendment to 802.16 to operate in the 2–11 GHz band using OFDM technology and extend the coverage to NLOS situations. Then 802.16b extended this standard to the 5–6 GHz bands and added measures to support QoS. The next amendment to 802.16 was 802.16c, which delivered a system profile for the 10–66 GHz 802.16 standard. The 802.16d, completed in 2004, was a revision project aligning with ETSI's pan-European HIPERMAN (the outdoor version of the HIPERLAN standard for WLANs) superseding the earlier 802.16a, b, and c amendments. Around this time the peak of interest in WiMAX started, and 802.16e was ratified in 2005 to

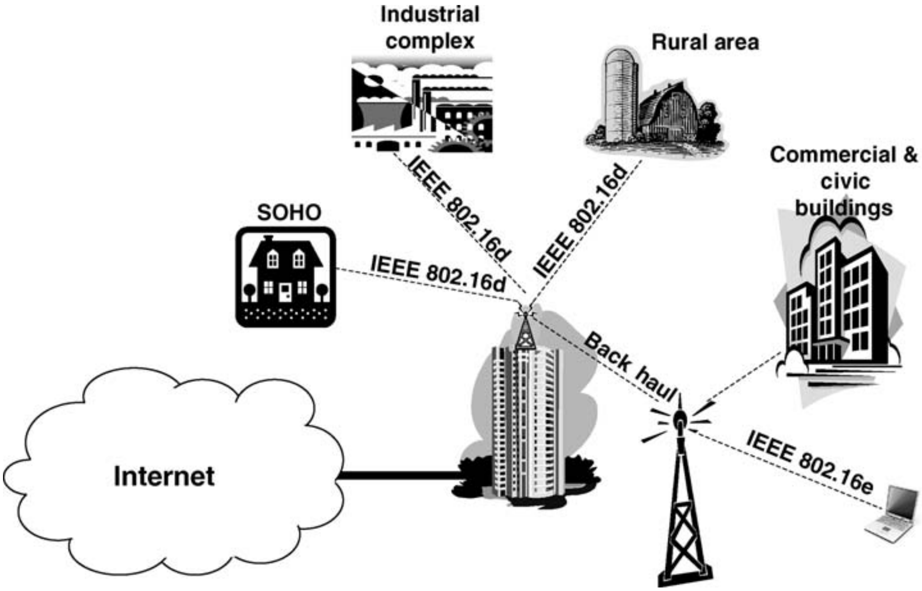


FIGURE 9.31 WiMAX vision for applications.

add mobility to this standard and make it more like a cellular network. The 802.16e is sometimes referred to as “Mobile WiMAX,” and it uses scalable OFDM and a more detailed QoS. Figure 9.31 illustrates the WiMAX view of the services after completion of 802.16d and e. This vision suggests a comprehensive wireless network connecting remote farms, factories, SOHOs, commercial buildings, and mobile terminals with a number of wireless connections suitable for deployment in remote areas and developing countries with limited existing wiring for the backbone of the network. This MAN technology is something in between WLANs and cellular networks. WiMAX defines a more complex architecture than WLANs to support QoS and mobility using outdoor antennas, which makes it closer to cellular networks. On the other hand, the transmission technology is based on OFDM and it is affected by HIPERLAN and other WLAN standards. In the rest of this section we provide more details on technical aspects of the 802.16 WiMAX technology.

9.3.1 General Architecture

The general architecture of WiMAX is defined in Fig. 9.32. There are three subsystems: the mobile subscriber station (MSS), the access service network (ASN), and the connectivity service network (CSN). The ASN modules connect BSs to each other and a gateway to connect to the CSN subsystem. The CSN subsystem includes home and visitor databases to support mobility management and users’ profiles. The CSN is connected to the access service providers or the Internet as the backbone of the network. The architecture differentiates the access and network providers. A network access provider (NAP) consists of a number of ASNs and a network service provider (NSP) owns the CSN to support ASNs. WiMAX defines a number of interconnections between subsystems to support multivendor operation. The architecture is flexible and can adapt to different hardware configurations, allowing adaptation to variety of fixed and mobile terminals and different sizes of BS.

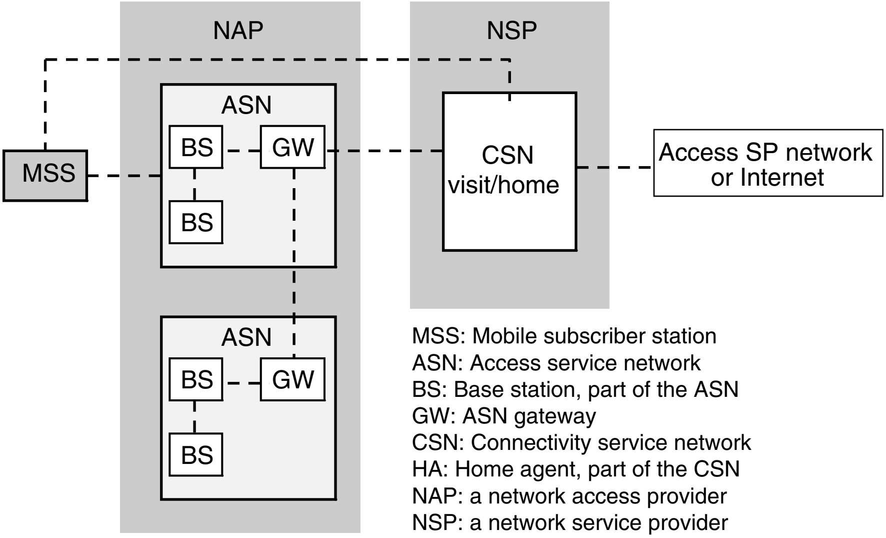

NAP
NSP
ASN
BS
GW
BS
MSS
CSN visit/home
Access SP network or Internet
ASN
BS
GW
BS
MSS: Mobile subscriber station
ASN: Access service network
BS: Base station, part of the ASN
GW: ASN gateway
CSN: Connectivity service network
HA: Home agent, part of the CSN
NAP: a network access provider
NSP: a network service provider

FIGURE 9.32 General architecture of WiMAX.

Figure 9.33 shows an overview of the protocol stack of 802.16. Similar to all other 802 standards, this standard also defines two layers. The MAC layer sandwiches the common MAC sublayer with convergence and security sublayers. A variety of *convergence sublayer* services for differentiated treatment of data are defined to support QoS and to describe how wireline technologies such as Ethernet, ATM, and IP are encapsulated in the 802.16 air interface. The MAC layer of the standard also describes a secure communications procedure using secure key exchange during authentication, and encryption during data transfer. These features facilitate application development, as well as faster and more reliable security, and they do not exist in the 802.11 MAC protocol stack. In addition, the MAC layer of WiMAX specifies power-saving mechanisms for sleeping and idle terminals and supports the handoff mechanisms. These features of WiMAX are very similar to those of cellular networks. The physical layer also has its own convergence protocols to support rate-adaptive operation for different modulation techniques.

9.3.2 Physical Layer

The physical layer of the legacy 802.16 LMDS systems was defined for 10–66 GHz using three single-carrier modulation options, QPSK, 16-QAM and 64-QAM, with bandwidths of

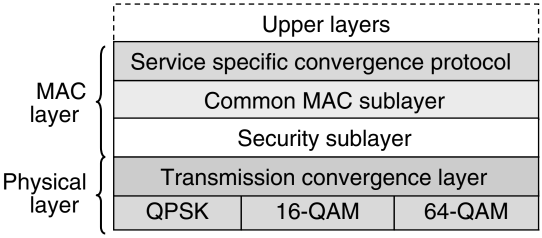


FIGURE 9.33 WiMAX protocol stack.

20 MHz, 25 MHz, and 28 MHz respectively. The highest data rate was 134 Mb/s, which was associated with 64-QAM and 28 MHz of bandwidth. At those high carrier frequencies and using simple modulation techniques the coverage would be limited to at most a few miles in LOS environments, which limits the practicality of the technology. This feature was considered to be one of the major reasons for failure of the original 802.16 to become a commercially successful standard. In terms of coverage, the IEEE 802.11 technology operating at 2.4 GHz already existed in the market at a reasonable price and it could provide a cheaper and possibly better solution than LMDS for last-mile wireless Internet access.

The revised 802.16d resorted to he OFDM technology with 256 carriers at 2–11 GHz using QPSK modulation, 16-QAM, and 64-QAM to support up to 75 Mb/s in 20 MHz of bandwidth. The carriers from different users are combined using orthogonal frequency division multiple access (OFDMA) with 2048 carriers for all users. OFDMA is a multiple-access or multiplexing scheme that provides multiplexing operation of data streams from multiple users. In an OFDMA system, resources are allocated using time slots for each OFDM symbol and the number of sub-carriers allocated to the individual user. The time slots and frequency sub-carriers are organized into sub-channels for allocation to individual users. Lower frequency of operation and using a more robust OFDM modulation, which had emerged as the modulation of choice for the 802.11a and g standards in indoor applications, were expected to extend the outdoor coverage for point-to-multipoint 802.16d BSs to 3–5 miles with support of non-LOS conditions. The 802.16e operates at frequencies between 2 and 6 GHz to support mobile operation. This standard uses scalable OFDMA (S-OFDMA) to carry data supporting channel bandwidths of between 1.25 and 20 MHz with up to 2048 sub-carriers. The scalability is supported by adjusting the FFT size while fixing the sub-carrier frequency spacing. Changing the size of FFT and subsequently the number of sub-carriers adds another order of flexibility to the choice between QPSK, 16-QPSK, and 64-QAM for individual carriers, providing a more flexible environment for rate-adaptive transmissions.

9.3.3 Medium Access Control Layer of WiMAX

The IEEE 802.16 common MAC uses a scheduling algorithm for centrally assigned connection-based access. In the original 802.16 the access was TDMA; OFDMA is used in the later versions of 802.16d/e, which can be interpreted as a modern implementation of the FDMA techniques.

The 802.16 PHY supports TDD and FDD operation operating in both licensed and unlicensed bands. FDD is considered to address applications where local spectrum regulatory requirements either prohibit TDD or are more suitable for FDD deployments. While FDD is suitable for wider coverage for two-way interactive and delay-sensitive telephony with symmetric traffic requirement, TDD enables adjustment of the downlink and uplink ratio to support asymmetric traffic efficiently for applications such as Internet access. TDD assures channel reciprocity for better support of link adaptation, MIMO, and other closed-loop advanced antenna technologies. TDD only requires a single channel for both downlink and uplink, resulting in a less complex implementation of the RF circuitry of the network and providing greater flexibility for adaptation to varied global spectrum allocations.

Compared with the contention-based IEEE 802.11 MAC, the IEEE 802.16 MAC uses connection-oriented centrally assigned access, which is better suited for the support of QoS and, consequently, for better voice quality traditional telephone connections and emerging VoIP and IP-TV applications. The IEEE 802.11 uses contention-based CSMA/CA with a restricted support mechanism for QoS provided by variable IFS, PCF, and RTS/CTS in the IEEE 802.11e extension. In IEEE 802.16, when a slot is allocated by the BS or the AP, the time slot can enlarge and contract, but other subscribers cannot use it. In high traffic loads this approach is more stable and bandwidth efficient than the IEEE 802.11 contention access. The advantage with contention access is its simplicity. This contrast between the assigned and random access MAC is one of the fundamental differences governing design of short-range wireless networks. In general, assigned access techniques are better suited for real-time streaming applications, such as telephone, and contention-based access is more suited for data applications. As we show in Chapter 10, this difference is also one of the differentiating factors between the Bluetooth and ZigBee technologies for WPANs.

QUESTIONS

1. Name three categories of unlicensed bands used in the US and compare them in terms of size of the available band and coverage.
2. Explain the differences between WLAN and WPAN.
3. How does the current state of the art data rate of the wired (Ethernet) and wireless LANs (802.11) compare with one another?
4. Why are unlicensed bands essential for the WLAN and WPAN industries?
5. Explain the difference between the wireless Inter-LAN bridges and wireless LANs.
6. What three topologies can IEEE 802.11 WLANs operate in? What are the differences?
7. What are the differences between IEEE 802.11 and HIPERLAN standards?
8. Name four major transmission techniques considered for WLAN standards and give the standard activity associated with each of them.
9. Compare OFDM and spread spectrum technology as PHY alternatives for WLANs.
10. Give the physical specification summary of the DSSS and FHSS used by the IEEE 802.11.
11. What are the MAC services of IEEE 802.11 that are not provided in the traditional LANs such as 802.3.
12. Why does the MAC layer of 802.11 have four address fields compared to 802.3 that has two?
13. What is the PCF in 802.11, what services does it provide, and how is it implemented?
14. Explain the difference between a hidden terminal and an exposed terminal.
15. What is the difference between an ESS and a BSS in the IEEE802.11?
16. Explain why an AP in 802.11 also acts as a bridge?
17. What are the differences between carrier sensing in the 802.11 and 802.3?
18. What are the responsibilities of the MAC Management sublayer in 802.11?
19. What is the difference between the backoff algorithms in the 802.11 and 802.3?
20. What is the purpose of PIFS, DIFS, and SIFS time intervals and how they are used in the IEEE 802.11?
21. What is the difference between a Probe and a Beacon signal in 802.11?
22. Explain the operation of the timing of the beacon signal in 802.11.
23. What is the difference between power control in 802.11 and power control in cellular systems?

24. How is authentication and integrity provided in IEEE 802.11?
25. What modulation scheme and code rates are used in IEE 802.11a to provide a raw data rate of 36 Mbps? Explain how this rate is achieved in terms of number of symbols and bits per second.
26. What are the differences between the different methods of deployment of infrastructure WLANs?
27. Compare the PHY layer of WiMax with that of 802.11.
28. Compare the MAC layer of WiMax with that of 802.11.

PROBLEMS

Problem 1:

You want to transmit the information sequence 00111100 using CCK as in 802.11b. What is the CCK codeword in vector form? Show all steps. Assume that bit d0 is the left most bit.

Problem 2:

(a) Use the equation for generation of CCK to generate the complex transmitted codes associated with the data sequence {0,1,0,0,1,0,1,1}

(b) Repeat (a) for the sequence {1,1,0,0,1,1,0,0}

(c) Show that the two generated codes are orthogonal.

Problem 3:

Write Matlab code to generate all 256 codewords of CCK in 802.11b. The aperiodic autocorrelation of a sequence $[\alpha_0\ \alpha_1\ \alpha_2\ \ldots\ \alpha_{N-1}]$ is given by:

$$R(k) = \begin{cases} \sum_{j=0}^{N-1-k} \alpha_j \alpha_{j+k}, & 0 \le k \le N-1 \\ \sum_{j=0}^{N-1+k} \alpha_j \alpha_{j+k}, & 1-N \le k \le 0 \\ 0, & |k| \ge 0 \end{cases}$$

Compute the aperiodic autocorrelation vectors of the 256 codewords using the function xcorr in Matlab. Add the autocorrelation vectors element-wise. Verify that the result shows a value of 2048 (256 × 8) for the center element and zeros for all other elements.

Problem 4:

The original WaveLAN, the basis for the IEEE 802.11 uses an 11-bit Barker code of [1,−1,1,1,−1,1,1,1,−1,−1,−1] for DSSS.

(a) Sketch the aperiodic autocorrelation of the code (see problem 3).

(b) If we use the system using random codes with the same chip length in a CDMA environment, how many simultaneous data users we can support with an omnidirectional antenna and one access point?

Problem 5:

(a) If in the PPM-IR PHY layer used for the IEEE 802.11 instead of PPM we were using baseband Manchester coding what would be the transmission data rate? Your reasoning must be given.

(b) What is the symbol transmission rate in the IEEE 802.11b? How many complex QPSK symbols are used in one coded symbol? How many bits are mapped into one transmitted symbol? What is the redundancy of the coded symbols (the ratio of the coded symbols to total number of choices)?

(c) What is the symbol transmission rate of the coded symbols per channel in the IEEE802.11a? How does this symbol rate relate to the data rates (6, 9, 12, 18, 27, 36, and 54Mbps) and convolutional coding rates ($^1/_2$, $^3/_4$, and 9/16)?

Problem 6:

Redraw the timing diagram of Fig. 9.8, assuming that all MSs use RTS/CTS mechanism to send packets.

Problem 7:

A voice over IP application layer software generates a 64Kbps coded voice packet every 20 ms. This software is installed in two laptops with WLAN PCMCIA cards communicating with an AP connected to a Fast Ethernet (100 Mbps).

(a) What is the length of the voice packets in msec, if the PCMCIA cards were DSSS IEEE 802.11?

(b) If the two terminals start to send voice packets almost at the same time. Give the timing diagram to show how the first packets are delivered though the wireless medium to the AP using CSMA/CA mechanism.

(c) Repeat (b) and (c) if 802.11b at 11 Mbps was used instead of DSSS 802.11. How would this change if 5.5 Mbps was the data rate?

Problem 8:

Figure P9.1 shows the layout of an office building. If the distance between the AP and the MSs 1, 2 and 3 are 50, 65, and 25 meters respectively, determine the path loss for between the AP and MSs:

(a) Using path loss per wall model and free space loss (assume that the wall loss is 3 dB per wall).

(b) Using the 802.11 path loss model from Chapter 2.

Using the 802.11 transmitted and received power specifications, determine whether a single AP can cover the entire building.

Problem 9:

We want to install a LAN in a 5 story office building with identical floor plans. Each floor of the building is a 80 m × 80 m square with a height of 4 m. There are 15 terminals in each floor of the building and the external wiring comes to the first floor.

(a) What is the total cost of wiring, equipment and installation of the entire network if an IEEE 802.3 star network with one 240 port switch in the first floor connects all terminals to each other and the external connection? Assume a charge of $150 per run of wiring between two locations, a $6,000 cost for the switch, and a $35 cost for the network interface card per terminal.

(b) To avoid wiring costs, assume that we use an IEEE 802.11 wireless LAN access point with 100 mW (20 dBm) transmitter power and a -80dBm receiver sensitivity in the center of the third floor. What is the total cost of the wireless LAN if the access point is $1,500 and each network interface card is sold for $200?

(c) Use the 802.11 path loss model from Chapter 2 to calculate the coverage of the access point. Can we cover the entire building with one access point?

Compare the advantages and disadvantages of the two solutions.

Problem 10:

You are designing a wireless LAN for an office building. You are not able to perform measurements or site surveys and have to rely on statistical models and certain other information. There are also certain constraints on where you can actually place the access point(s). You have the following information available to you:

Maximum number of walls between an access point and a mobile terminal = 4
Maximum number of floors between an access point and the mobile terminal = 2
Transmit power possibilities = 250 mW and 100 mW
Sensitivity of receiver is -90 dBm
Maximum distance from access point to building edge = 30m
Building has office walls, brick walls, and metallic doors
Shadow fading margin = 8dB

What would be a conservative estimate of the number of access points required for the WLAN set-up? Why? State your assumptions, models, and provide reasons for all your assumptions and calculations. *Hint: Use path loss models from Chapter 2 that are applicable to indoor areas.*

Problem 11:

Suppose the coverage areas where 11, 5.5, 2, and 1 Mbps data rates are reliably available have radii of 20, 30, 40 and 50 m instead of the values in Example 9-6. What is the spatial capacity of the access point? How would the spatial capacity change if the 1 Mbps data rate was available up to 75m? Plot the spatial capacity Vs the range of the 1 Mbps coverage area as it varies from a radius of 50m to 100m assuming all other values remain the same.

Problem 12:

Figure P9.2(a) shows the overhead for packet formation and applications using TCP packets. Each TCP packet can have a length of up to 65495 byte that should be fragmented to fit the maximum MAC packet of 2312 byte. The TCP/IP header is 40 byte, the 802.2 LLC/SNAP header is 8 byte, and the 802.11 MAC and PLCP headers and synchronization preamble are 34 and 24 bytes respectively. The TCP ACK is a TCP header with no application data and the MAC ACK is shown in Fig. P9.2(b). Assuming SIFS and DIFS intervals of 10µsec and 50µsec respectively, determine the application throughput of the 802.11b for data rates of 11, 5.5, 2, and 1 Mbps for data packets of length 100 and 1000 bytes.

Problem 13:

In reality the throughput of a WLAN is a function of the channel characteristics and it fluctuates in time. Figure P9.3 shows a typical application throughput of an 802.11b terminal in a one minute observation time. Due to the channel fading and other imperfections this throughput varies in time as we measure it in a certain distance between the transmitter and the receiver. As the distance between the transmitter and the receiver increases and the RSS reduces this average throughput also reduces. The throughput (Mbps) versus distance (meters) relation of an IEEE 802.11b in an office building is empirically determined to follow the following approximated equation:

$$S_u(r) = -0.2r + 5.5 \tag{P9.1}$$

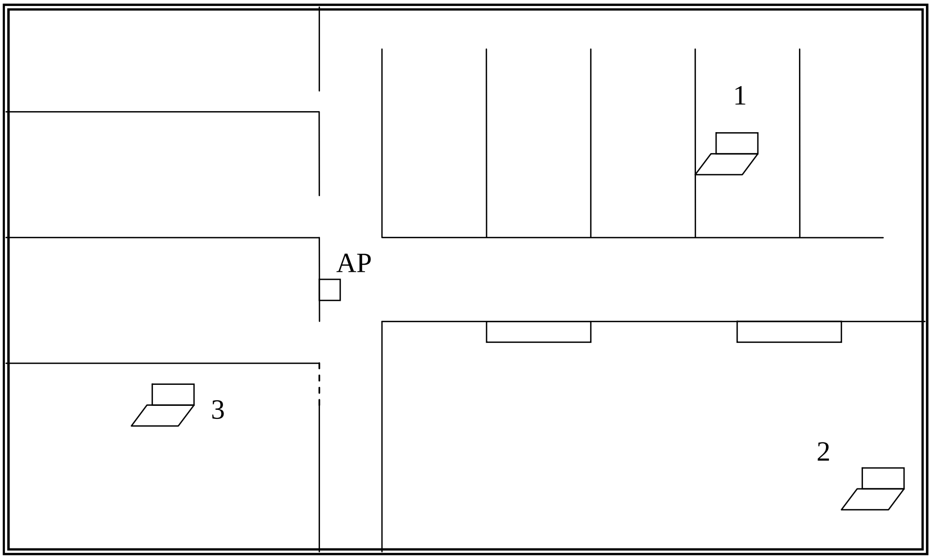


FIGURE P9.1 Layout of an office building.

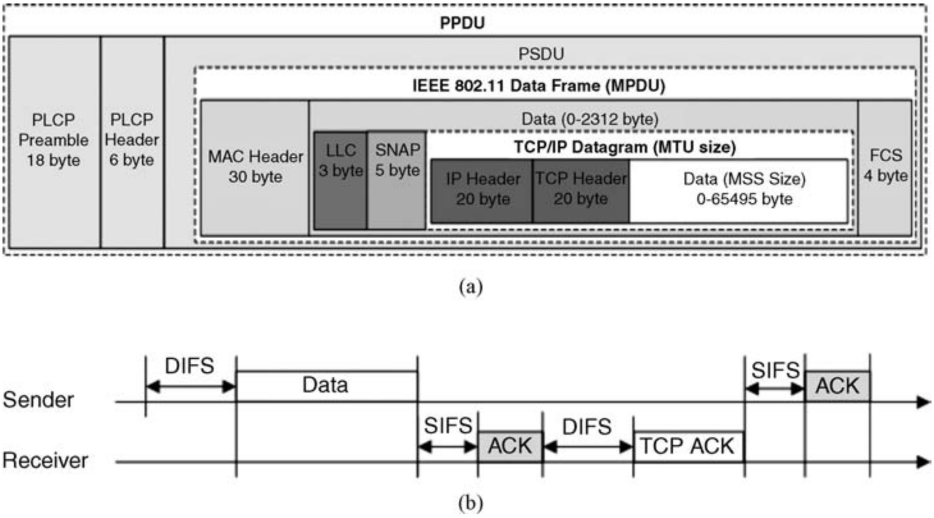


FIGURE P9.2 Packet transmission in the IEEE 802.11 b. (a) Overheads for the formation of a packet (b) overheads for successful transmission of a TCP packet.

(a) Determine the maximum throughput in Mbps and maximum coverage in meters?

(b) Show that we can find the average throughput of a user randomly walking in the coverage area of an access point by:

$$\bar{S_u} = \frac{2\int_0^{R_L} rS_u(r)dr}{R_L^2}[Mbps] \qquad \text{(P9.2)}$$

In which R_L is the distance for which the thoughput of the WLAN approaches to zero and the WLAN has no coverage anymore.

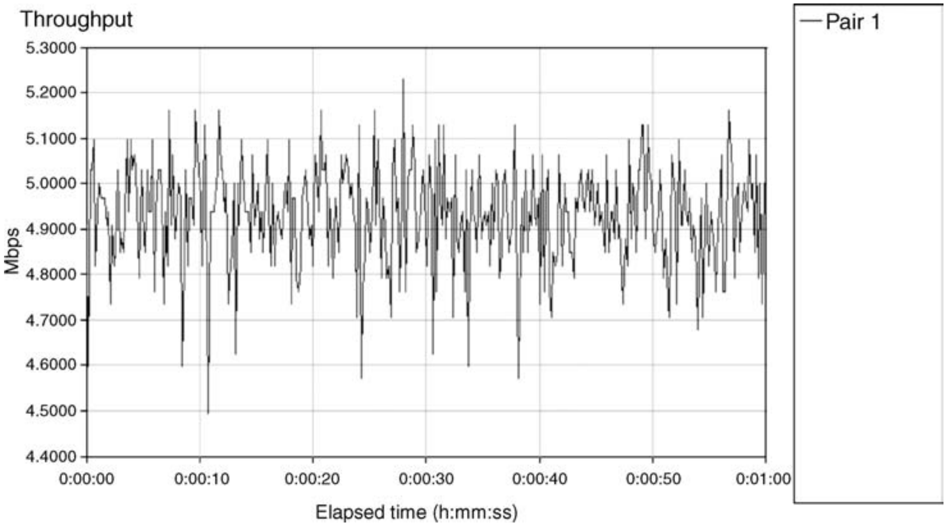


FIGURE P9.3 Throughput variations in one location for 802.11b.

(c) Use Eqs. (P9.1) and (P9.2) to calculate the average throughput of a user randomly walking in the coverage area of the WLAN.

(d) Compare your results with the minimum and maximum nominal data rates of the IEEE 802.11b. Explain the difference between your results.

Problem 14:

The average throughput versus distance relationship of an IEEE 802.11g in a typical office building is measured to fit the following function:

$$S_u(r) = \begin{cases} 22; & 0 < r < 1 \\ -22\log 10r + 25; & r > 1 \end{cases} \tag{P9.3}$$

(a) Use Eqs. (P9.1) and (P9.2) to calculate the average throughput of a user randomly walking in the coverage area of the WLAN.

(b) Compare your results with the minimum and maximum nominal data rates of the IEEE 802.11g. Explain the difference between your results.

PROJECTS

These projects assume that you have a wireless network at home and a laptop that has the ability to connect to WiFi networks.

Project 1:

Use the free tool Netstumbler (http://www.netstumbler.com/) or Inssider (http://www.metageek.net/products/inssider) on Windows or the equivalent istumbler (http://www.istumbler.net/) with MAC OS X to map all WiFi access points near your house. Enumerate the ESSID, the MAC address, and channel numbers that the access points are using. How many networks can you detect from inside your house? Does the number of networks change if you try scanning for them from outside your house? How many interfering networks are near your house (i.e. those networks using the same channel number)?

Project 2:

Note that the networks detected in Project 1 are only those networks that choose to broadcast their SSIDs. There are WiFi networks that do not broadcast their IDs that may also be in the area. Use the Kismet tool (http://www.kismetwireless.net/) to see if there are any hidden networks in your area.

Project 3:

Use one of the tools in Project 1 to check the received signal strength from your access point at different locations inside your house. What is the best RSS you receive and the worst inside your house? How does this change outside your house? Collect enough data and plot the average RSS values as a function of approximate distance from the AP. Can you develop your own path-loss model from this data?

10

IEEE 802.15 WIRELESS PERSONAL-AREA NETWORK

10.1 INTRODUCTION

In Chapters 8 and 9 we provided an overview of the wired LAN and WLAN technologies. In this chapter we provide an overview of WPAN activities. At the time of writing, WPANs are differentiated from WLANs by their smaller area of coverage, ad hoc-only topology, plug-and-play architecture, support of voice and data devices, and low power consumption.

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

WPANs started as BodyLANs, which connect sensors and information devices attached to the body to neighbors for military application and as personal networks to connect a perrson's personal equipment such as laptops, notepads, and cell phones in commercial applications. WPANs are also used in sensor networks, a topic that we consider in Chapter 13. Some of the material in this chapter is presented again, albeit differently in that chapter for fluency and completeness.

The very first PAN to be announced was the BodyLAN, which emerged from a DARPA project in the mid 1990s. This was a low-power, small-sized, inexpensive, WPAN with modest bandwidth that could connect personal devices in many co-located systems with a range of around 5 feet ($\sim$1.5 m) [Den96]. Motivated by the BodyLAN project, a WPAN group was originally started in June 1997 as a part of the IEEE 802.11 standardization activity. In May 1998, the Bluetooth development was announced and a Bluetooth special group was formed within the WPAN group [Sie00]. In March 1999, the IEEE 802.15 was approved as a separate group in the 802 community to handle WPAN standardization. By the early 2000s, IEEE 802.15 WPAN had four subcommittees: Bluetooth, coexistence, high data rate, and low data rate. Bluetooth has been selected as the base specification for IEEE 802.15. More recently, two other subcommittees have been added to IEEE 802.15 for mesh networking and body-area networks (BANs).

A study of the details of the IEEE 802.15 standards for PHY and MAC layers provides a good overview of the applications of the technical material we presented in Part One of the book, because a variety of technologies have been developed under this standardization committee.

10.1.1 IEEE 802.15 Wireless Personal-Area Network Standardization Series

The 802.15 WPAN group is focused on development of short-distance wireless networks used for networking of portable and mobile computing devices such as PCs, PDAs, cell phones, printers, speakers, microphones, and other consumer electronics. The WPAN group intends to publish standards that allow these devices to coexist and interoperate with one another and other wireless and wired networks in an internationally acceptable frequency of operation.

The original functional requirement, published 22 January 1998, was based on the BodyLAN project and specified devices with [Hei98]:

- power management – low current consumption;
- range – 0–10 m;
- speed – 19.2–100 kb/s (actual);
- small size, e.g. $\sim$0.5 cubic inches ($\sim$8.2 cm^3) no antenna;
- low cost, i.e. relative to target device;
- should allow overlap of multiple networks in the same area;
- networking support for a minimum of 16 devices.

These specifications fit the Bluetooth specification that was announced after this premier announcement. Until today, IEEE 802.15 has had six groups with variety of subgroups and standard specifications. Task group one was based on Bluetooth and defines PHY and MAC specifications for wireless connectivity with fixed, portable, and moving devices within or entering a space about a person or object that typically extends up to 10 m in all directions and envelops the person whether stationary or in motion. Several versions of the Bluetooth specifications have been released by the IEEE 802.15.1. The latest and the last of this standard series was finalized in 2005.

Task group two was chartered to focus on the coexistence of WPAN and 802.11 WLANs. This group developed a coexistence model to quantify the mutual interference and a coexistence mechanism to facilitate coexistence of an IEEE 802.11 WLAN and the Bluetooth. A goal of this group was to achieve a level of interoperability that could allow the transfer of data between a Bluetooth device and an 802.11 device. This committee is no longer active and the resulting recommendations were not used in any popular product.

Task group three of the IEEE P802.15 works on PHY and MAC layer for high-rate (HR) WPANs that operate at data rates in the order of gigabits per second using UWB technology. This standard was expected to provide for low-power, low-cost solutions addressing the needs of portable consumer digital imaging and multimedia applications. IEEE 802.3 resulted in a number of group activities. IEEE 802.15.3a started in 2002 and worked on multiband OFDM (MB-OFDM) UWB and DS-UWB technologies. In 2006, impeded by regulatory issues and market uncertainty, this task group withdrew from the project until the technology proves itself to be commercially viable. IEEE 802.3b worked on interoperability among different MACs, and IEEE 802.15.3c worked on millimeter-wave PHY alternatives at frequencies up to 60 GHz to support data rates on the order of several gigabits per second for applications such as short-range cable replacements to carry video streaming for home theatre applications.

Task group four was chartered to investigate an ultralow complexity, ultralow power-consuming (durations of months and years), ultralow cost PHY and MAC layers for data rates around 250 kb/s. Potential applications were sensors, interactive toys, smart badges, remote controls, and home automation. The project is also keen on addressing the location tracking capabilities required to support uses of smart tags and badges. The first edition of the 802.15.4 standard was released in May 2003, and in 2004 the task group put itself into hibernation. The ZigBee set of higher level communication protocols is based upon the specification produced by IEEE 802.15.4.

Task force five works on mesh networks with the IEEE 802.11s group. Mesh networks are expected to extend the coverage of the WLAN and WPAN technologies by supporting the implementation of a multi-hop ad hoc networking topology.

A recent task group of the IEEE 802.15 is Working Group 6, initiated in 2007, which focuses on BANs. Compared with IEEE 802.15.1 Bluetooth, IEEE 802.15.6 and its BAN system are focused on operating at relatively low frequencies (less than 1 MHz) and at shorter distances and lower data rates. This technology is geared towards applications such as communications between a pacemaker and a wristwatch or an embedded RFID tag inside a body and an RFID reader outside.

10.2 IEEE 802.15.1 BLUETOOTH

Bluetooth is an open specification for short-range wireless voice and data communications that was originally developed for cable replacement in personal area networking to operate all over the world. The initial study for development of this technology started at Ericsson, Sweden, in 1994. In 1998, Ericsson, Nokia, IBM, Toshiba, and Intel formed a special interest group (SIG) to expand the concept and develop a standard under IEEE 802.15 WPAN. In 1999, the first specification, v1.0b, was released and then accepted as the IEEE 802.15 WPAN standard for 1 Mb/s networks. At the time of writing, over 1000 companies participate as members in the Bluetooth SIG, and a number of companies all over the world are developing Bluetooth chip sets. Marketing forecasts indicate penetration of Bluetooth in more than 100 million cellular phones and several millions of other consumer devices. The IEEE 802.15 standard is also studying coexistence among and interference between Bluetooth and IEEE 802.11 products operating at 2.4 GHz.

Bluetooth is the first popular technology for short-range ad hoc networking that is designed for integrated voice and data applications. Unlike WLANs, Bluetooth has a lower data rate, but it has an embedded mechanism to support voice applications. Unlike 3G cellular systems, Bluetooth is an inexpensive personal-area ad hoc network operating in unlicensed bands and owned by the user.

The Bluetooth SIG considers three application basic scenarios, which are shown in Fig. 10.1 [BLU00]. The first scenario, shown in Fig. 10.1*a*, is the wire replacement to connect a personal computer or laptop to its keyboard, mouse, microphone, and notepad. As the name of the scenario indicates, it avoids multiple short-range wiring surrounding today's personal computing devices. The second scenario is ad hoc networking of several different users at very short range from each other, such as in a conference room. As we saw in the last three chapters, WLAN standards and products also commonly consider this scenario. The third scenario is use as an AP to the wide-area voice and data services provided by the cellular networks, wired connection, or satellite links. The 802.11 community also

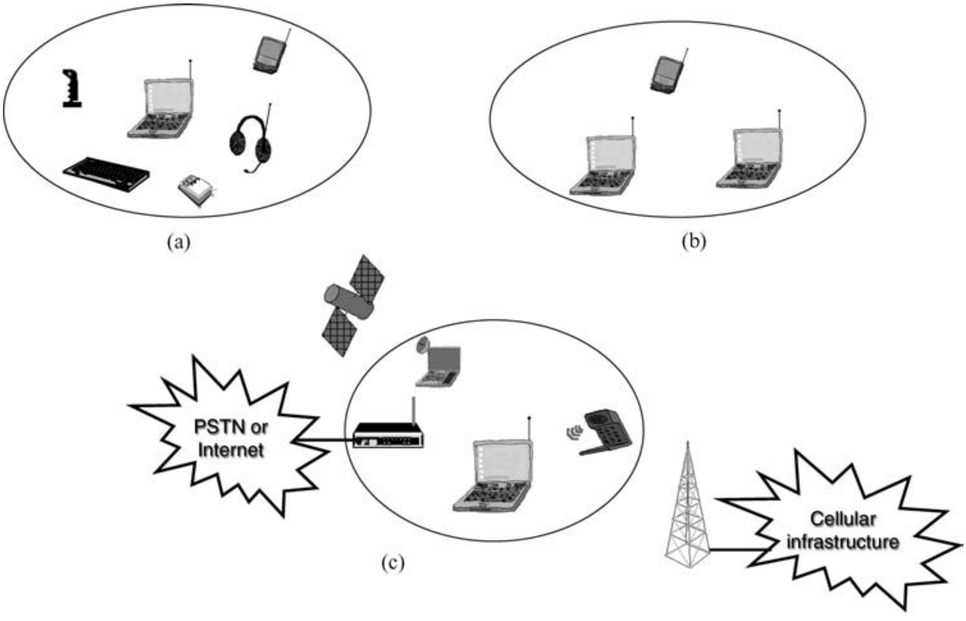


FIGURE 10.1 Bluetooth application scenarios: (*a*) cable replacement; (*b*) ad hoc personal network; (*c*) integrated AP.

considers this overall concept of the AP. However, the Bluetooth AP is used in an integrated manner to connect to both voice and data backbone infrastructures. The HIPERLAN/2 standard will provide a more expensive version of similar connections that supports a larger number of users and wider bandwidths.

10.2.1 Overall Architecture

The topology of the Bluetooth is referred to as a *scattered ad hoc topology* and is illustrated in Fig. 10.2. In a scattered ad hoc environment, a number of small networks, each supporting a few terminals, coexist or possibly interoperate with one another. To implement such a network we need a plug-and-play environment. The network should be self-configurable, providing an easy mechanism to form a new small network, as well as participation in an existing small network. To implement that environment the system should be capable of providing different states for connecting to the network. The terminals should have options to associate with multiple networks at the same time. The access method should allow formation of small independent ad hoc cells, as well as the possibility of interacting with large voice and data networks considered by Bluetooth.

To accommodate these features the Bluetooth specification defines a small cell as a *piconet* and identifies four states: master (M), slave (S), standby (SB) and parked/hold (P) for a terminal. Like other ad hoc topologies, such as the one supported by IEEE 802.11, each terminal can be an M or an S. As shown in Fig. 10.2, the Bluetooth topology, however, allows S terminals to participate in more than one piconet. An M terminal in the Bluetooth can handle seven simultaneous and up to 200 active slaves in a piconet. If access is not available, then a terminal can enter the SB mode, waiting to join the piconet later. A radio can also be in a P mode, in a low-power connection. In the P mode the terminal releases its MAC address while in the SB state it keeps its MAC address. Up to 10 piconets can operate in one area [Blu00]. Bluetooth specifications have selected the unlicensed ISM bands at 2.4 GHz for operation. The advantage is the worldwide availability of the bands and the disadvantage is the existence of other users, in particular IEEE 802.11 and 802.11b products, in the same band. At the time of writing, a

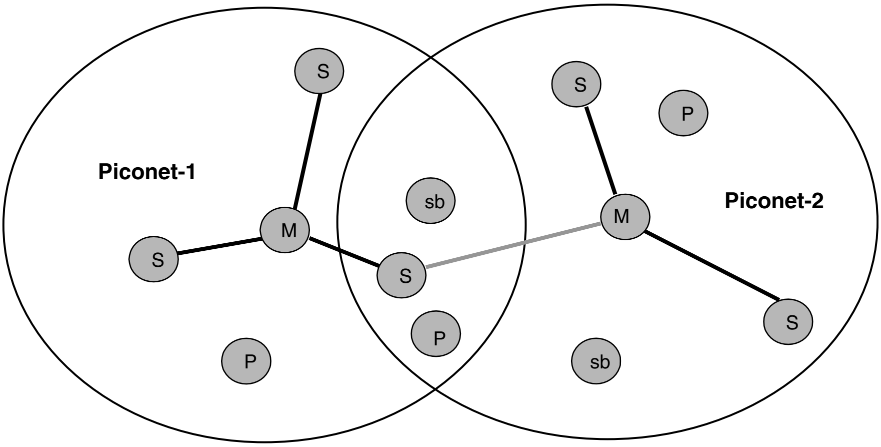


FIGURE 10.2 Bluetooth's scattered ad hoc topology.

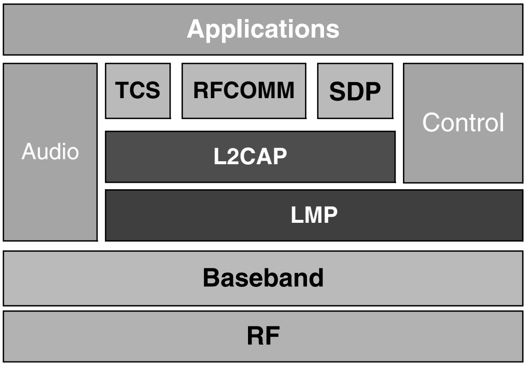


FIGURE 10.3 Protocol stack of the Bluetooth.

subcommittee of the IEEE 802.15 is working on the interference issues related to Bluetooth and IEEE 802.11 and 11b.

10.2.2 Protocol Stack

One of the distinct features of Bluetooth is that it provides a complete protocol stack that allows different applications to communicate over a variety of devices. Other wireless local networks, such as IEEE 802.11, specify the three lower layers for communications. The protocol stack for voice, data, and control signaling in Bluetooth is shown in Fig. 10.3 [Haa00]. The *RF layer* specifies the radio modem used for transmission and reception of the information. The *baseband layer* specifies the link control at bit and packet level. It specifies coding and encryption for packet assembly and FH operation. The *link management protocol* (LMP) configures the links to other devices by providing for authentication and encryption, state of units in the piconet, power modes, traffic scheduling, and packet format. The *LLC and adaptation protocol* (L2CAP) provides connection-oriented and connectionless data services to the upper layer protocols. These services include protocol multiplexing, segmentation and reassembly, and group abstractions for data packets up to 64 kbytes in length. The audio signal is directly transferred from the application to the baseband. Also, the LMP and the application exchange control messages interact to prepare the physical transport to the application.

Different applications may use different protocol stacks but, nevertheless, all of them share the same physical and data link control mechanisms. There are three other protocols above the L2CAP. The *service discovery protocol* (SDP) finds the characteristics of the services and connects two or more Bluetooth devices to support a service such as faxing, printing, teleconferencing, or e-commerce facilities. The *telephony control protocol specification* (TCS) defines the call control signaling and mobility management for the establishment of speech for cordless telephone application. Using these protocols, legacy telecommunication applications can be developed.

Example 10.1: TCS in Bluetooth Figure 10.4 shows the protocol stack for implementation of the cordless telephone application. The audio signal is directly

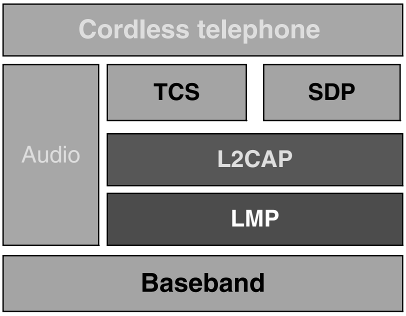


FIGURE 10.4 Protocol stack for implementation of cordless telephone over Bluetooth.

transferred to the baseband layer while SDP and TCS protocols operating over L2CAP and LMP handle signaling and connection management.

The *RFCOMM* is a "cable replacement" protocol that emulates the standard RS-232 control and data signals over Bluetooth baseband. Using RFCOMM, a number of non-Bluetooth-specific protocols can be implemented on the Bluetooth devices to support legacy applications.

Example 10.2: Lightweight Applications in Bluetooth Figure 10.5 shows the implementation of a vCard application for credit card verification. This application protocol runs over the object exchange (OBEX) protocol that is accommodated by the RFCOMM protocol in the Bluetooth protocol stack. Therefore, the sequence of protocols for implementation of credit card verification over Bluetooth is: vCard, OBEX, RFCOMM, L2CAP, Baseband, RF. This protocol stack implementation contains both the internal object representation convention of vCard and the over-the-air transport protocols of Bluetooth.

Example 10.3: Wireless Application Protocol over Bluetooth Figure 10.6 shows the implementation of a wireless application environment (WAE) protocol that defines applications over the wireless application protocol (WAP). The WAP packets use the TCP/UDP protocols for Internet access on top of the PPP that runs over the RFCOMM.

The overall Bluetooth protocols can be divided into three classes. The Bluetooth SIG developed the core Bluetooth-specific protocols for baseband, LMP, L2CAP, and SDP exclusively. Protocols that were also developed by the Bluetooth SIG but which were based on existing protocols include RFCOMM and TCS. The third group consists of existing protocols that are adopted by the Bluetooth SIG. Examples of these protocols include PPP,

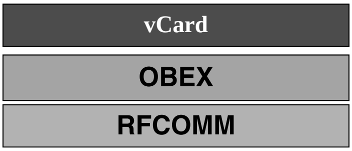


FIGURE 10.5 Protocol stack for implementation of vCard over Bluetooth.

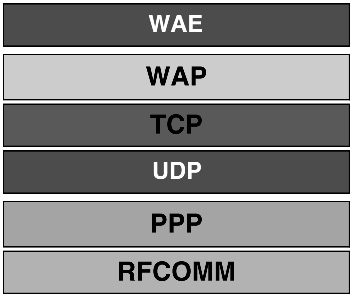


FIGURE 10.6 Protocol stack for implementation of WAE over Bluetooth.

UDP/TCP/IP, OBEX, WAP, vCard, vCal, IrMC-1, and WAE. The Bluetooth specification is open, and other legacy protocols such as hypertext transfer protocol (HTTP) and FTP can be accommodated on top of the existing protocol stack.

Example 10.4: FTP over Bluetooth Figure 10.7 provides a protocol stack for implementation of the FTP application. OBEX and RFCOMM manage the data transfer, while SDP provides for the establishment of the link.

The overall structure of the protocol stack in Bluetooth does not clearly follow the OSI model and its acronyms. Therefore, the division of the following section may appear somewhat different from other wireless local networks described in the last two chapters. However, we make every effort to make them as close as possible to the previous chapters to aid the reader to understand the details and relate them to other details for similar systems.

10.2.3 Physical Connection

The OSI equivalent physical layer of Bluetooth is embedded in the RF and baseband layers of the Bluetooth protocol stack. The physical connection of Bluetooth uses an FH-SS modem with a nominal antenna power of 0 dBm (10 m coverage) that has an option to operate at 20 dBm (100 m coverage). Like the 1 Mb/s option of the IEEE 802.11 FH-SS standard, the Bluetooth specification uses a two-level GFSK modem with a transmission rate of 1 Mb/s that hops over 79 channels in the ISM bands starting at 2.402 GHz and

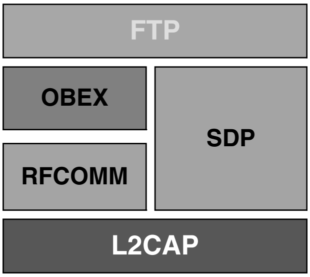


FIGURE 10.7 Protocol stack for implementation of FTP over Bluetooth.

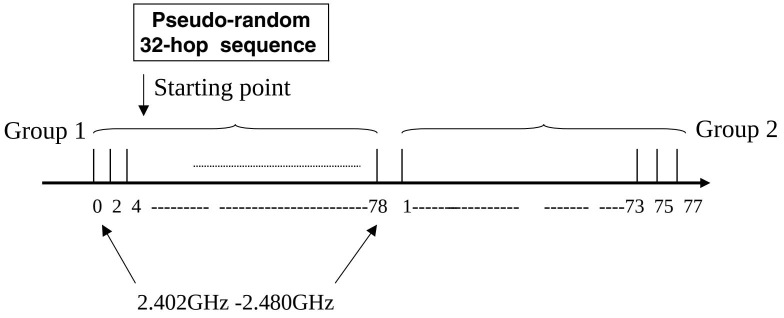


FIGURE 10.8 The hopping sequence mechanism in Bluetooth.

stopping at 2.480 GHz. The hopping rate and pattern and number of hops used in Bluetooth, however, are different from IEEE 802.11. The Bluetooth hopping rate is 1600 hops per second (625 μs dwell time), compared with the 2.5 hops per second (400 ms dwell time) system adopted by 802.11. The two-level GFSK modem allows simple noncoherent detection implementation using simple FM demodulators. The 0 dBm modem with the Bluetooth hopping pattern complies with the FCC rules in the USA; owing to local regulations, the bandwidth is reduced in Japan, France, and Spain. An internal software switch (that allows an environment for implementation of a system that works universally) handles this transition.

The Bluetooth specification assigns a specific FH pattern for each piconet. This pattern is determined by the piconet identity and master clock phase residing in the master terminal in the piconet. Figure 10.8 illustrates the elements of the FH strategy in Bluetooth. The overall hopping pattern is divided into 32 hop segments. The 32-hop pseudorandom hopping pattern segment is generated based on the master identity and clock phase. The 79 frequency hops at the ISM bands are arranged in odd and even classes. Each 32-hop sequence starts at a point in the spectrum and hops over the pattern, which covers 64 MHz because it hops either on odd or even frequencies. After completion of each segment the sequence is altered and the segment is shifted 16 frequencies to the forward direction. The 32 hops are concatenated and the random selection of the index is changed for each new segment. This way, segments slide through the carrier list to maintain the average time each frequency is used at an equal probability. Change of the clock or identity of the piconets will change the sequence and segment mapping, allowing different piconets to operate with a different set of random codes. These codes are not orthogonal to one another, but they are randomized against each other. With 79 hops it is difficult to find a large number of orthogonal codes anyway [Haa00].

To protect the integrity of the transmitted data, Bluetooth uses two error-correction schemes in the baseband controllers. An FEC code is always applied to the header information, and if needed it is extended to the payload data for the voice-oriented synchronous packets. The FEC code generally reduces the number of retransmissions. It is always applied to the header, because header information is short and important. The flexibility of using FEC for payload provides an option to avoid overhead in favor of increased throughput when the channel is good and error free. An unnumbered ARQ scheme is also applied by the baseband layer for the asynchronous data-oriented information in which the recipient acknowledges data transmitted. For data transmission to be acknowledged, both the header error check and the payload check, if applied, must indicate no error condition. These functionalities implemented in the baseband layer of the Bluetooth protocol stack are often implemented in the data link layer of the OSI reference model for networks complying with that model.

10.2.4 Medium Access Control Mechanism

Although the modulation technique and frequency of operation of the Bluetooth radio system closely follows that of the FH-SS 802.11, the MAC mechanism in Bluetooth is widely different from 802.11. The Bluetooth access mechanism is a voice-oriented innovative system that is not identical either to the data-oriented CSMA/CA type or to voice-oriented CDMA or TDMA methods, yet it has elements that are somehow related to these access methods. The medium access mechanism of Bluetooth is a fast FH-CDMA/TDD system that employs polling to establish the link. The fast hopping of 1600 hops per second allows short time slots of 625 μs (625 bits at 1 Mb/s) for one packet transmission that allows a better performance in the presence of interference. Bluetooth is a CDMA system that is implemented using FH-SS. In the Bluetooth CDMA, each piconet has its own spreading sequence, while in the DSSS/CDMA system used for digital cellular systems each user link is identified with a spreading code. The DSSS/CDMA is not selected for Bluetooth because DSSS/CDMA needs central power control that is not possible in scattered ad hoc topology envisioned for Bluetooth applications. Without any need for centralized power control for CDMA operation, the FH/CDMA in Bluetooth allows tens of piconets to overlap in the same area, providing an effective throughput that is much larger than 1 Mb/s. As we discuss in Chapter 11, FH-SS 802.11 operates in the same 79 hops as Bluetooth, with only three sets of hopping patterns. The throughput of the Bluetooth FH/CDMA system, however, is less than the 79 Mb/s that could be achieved in a coordinated FDM or OFDM system, as employed in 802.11a and HIPERLAN/2 operating in the 5.2 GHz U-NII bands. In Bluetooth, the FH/CDMA is selected over simple FDM or OFDM because ISM bands at 2.4 GHz only allow spread-spectrum technology. The access method in each piconet of Bluetooth is TDMA/TDD. The TDMA format allows multiple voice and data terminals to participate in a piconet. The TDD eliminates crosstalk between transmitter and the receiver, allowing a single-chip implementation in which a radio alternates between transmitter and receiver modes. To share the medium among a larger number of terminals, at each slot a “master” decides and *polls* a “slave”; polling is used rather than contention access methods because contention provides too much overhead for the short packets (625 bits) that were selected for implementation of a fast FH system.

10.2.5 Frame Formats

The Bluetooth packet format is based on one packet per hop and a basic one-slot packet of 625 μs that can be extended to three slots (1875 μs) and five slots (3125 μs). This frame format and the FH/TDMA/TDD access mechanism allow an M terminal to poll multiple S terminals at different data rates for voice and data applications to form a piconet.

Example 10.5: Operation of Piconets Figure 10.9 illustrates several examples of Bluetooth operation in a piconet. In Fig. 10.9*a*, an M terminal is communicating with three

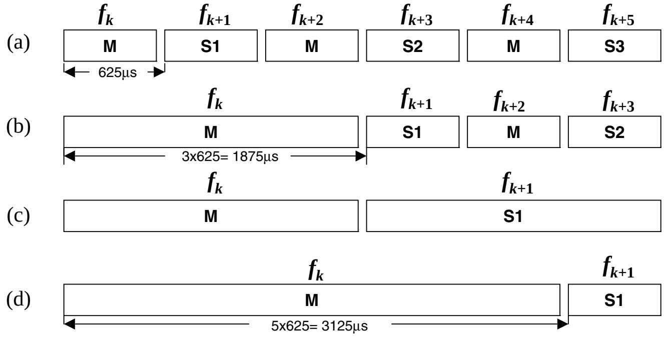


FIGURE 10.9 FH/TDMA/TDD multislot packet formats in Bluetooth: (*a*) one-slot packets; (*b*) asymmetric three-slot; (*c*) symmetric three-slot (1875 μs); (*d*) asymmetric five-slot (3125 μs).

S terminals. The TDMA/TDD format allows simultaneous operation of the three terminals, assigning 625 μs (equivalent to 625 bits at 1 Mb/s) for transmission and a time gap between the two packets in each direction. Terminals may run different applications (voice or data at different rates), but applications should run on one of the one-slot detailed packet formats that are specified by the Bluetooth SIG. The time gap is specified at 200 μs to allow a terminal to switch from transmitter to receiver mode for the TDD operation [Haa00]. Figure 10.9*b* shows an asymmetric communication in which the M uses a higher speed three-slot link while the S operates at lower rate with one-slot packets. Figure 10.9*c* represents a symmetric higher speed three-slot communication and Fig. 10.9*d* an asymmetric very high-speed five-slot with a return low-speed one-slot link.

The overall packet structure of Bluetooth is shown in Fig. 10.10. There are 74 bits for the access code field, 54 bits for the header field, and up to 2744 bits for different payloads, which can be as long as five slots. In IEEE 802.11 FH-SS packets, the preamble and header of the physical layer, shown in Fig. 9.17, were 96 bits and 32 bits respectively, while the payload could be as long as 4096 × 8 = 32 768 bits. The size of the overhead is more or less in the same range, but the maximum payload of 802.11 is at least an order of magnitude

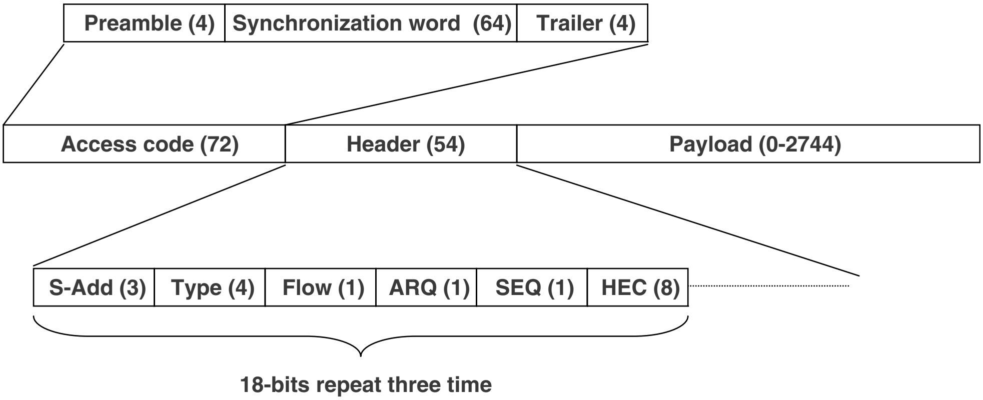


FIGURE 10.10 Overall frame format of the Bluetooth packets.

larger. Apparently, Bluetooth uses more-flexible shorter packets for ease of integration and better performance in fading, but these gains are at the expense of a higher percentage of overhead that reduces the throughput.

As shown in Fig. 10.10, the access code field consists of a 4-bit preamble and a 4-bit trailer plus a 64-bit synchronization PN sequence with a large number of codes with good autocorrelation and cross-correlation properties. The 48-bit IEEE MAC address unique to every Bluetooth device is used as the seed to derive the PN sequence for hopping frequencies of the device. There are four different types of access code. The first type identifies an M terminal and its piconet address. The second type of access code specifies an S identity that is used to page a specific S. The third type is a fixed access code reserved for the inquiry process, which will be explained later. The fourth type is the dedicated access code, which is reserved to identify a specific set of devices, such as fax machines, printers, or cellular phones.

As shown in Fig. 10.10, the header field has 18 bits that are repeated three times with a $^{1}/_{3}$ FEC code. The 18 bits start with a 3-bit S address identifier, 4-bits for packet type, 3-bits for status reports, and an 8-bit error check parity for the header. The 3-bit S-Add allows addressing the seven possible active Ms in a piconet. The 4-bit packet type allows 16 choices for different grades of voice service, data services at different rates, and four control packets. The 3-bit status reports are used to flag overflow of the terminal with information, acknowledgement of successful transmission of a packet, and sequencing to differentiate the sent and resent packets.

The Bluetooth SIG specifies different payloads and associated packet-type codes that allow implementation of a number of voice and data services. Different master–slave pairs in a piconet can use different packet types, and the packet type may change arbitrarily during a communication session. The 4-bit packet type identifies 16 different packet formats for the payloads of the Bluetooth packets. Six of these payload formats are asynchronous connectionless (ACL), primarily used for packet data communications. Three of the payload formats are synchronous connection oriented (SCO), primarily used for voice communications. One is an integrated voice (SCO) and data (ACL) packet and four are control packets common for both SCO and ACL links.

The three *SCO* packets, shown in Fig. 10.11, are high-quality voice (HV) packets numbered as HV1, -2, and -3 to designate the level of quality of the service. The SCO

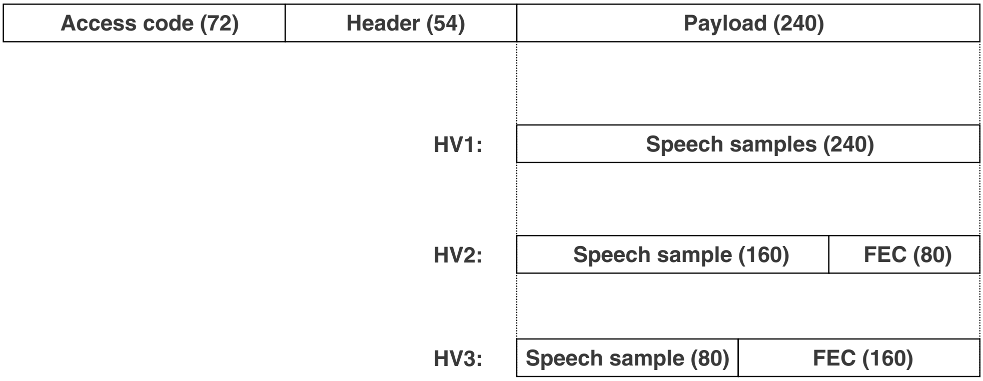


FIGURE 10.11 SCO one-slot packet frame formats.

packets are all single-slot packets, the length of the payload being fixed at 240 bits, and they do not use the status report bits, but they are transmitted over reserved periodic duplex intervals to support 64 kb/s per voice user. HV1 uses all 240 bits for the user voice samples, HV2 uses 160 bits for user voice samples and 80 bits of parity for a $^1\!/_3$ FEC code, and HV3 uses 80 bits of user voice samples and 160 bits of parity for a $^2\!/_3$ FEC code. To keep the data rate for voice samples at 64 kb/s, the HV1, HV2, and HV3 packets in each direction are sent every six, four and two slots respectively.

Example 10.6: Data Rate of HV Packets The HV1 packets are 240 bits long, and so they are sent every six slots. The packets are one-slot packets sent at the rate of 1600 slots/s. Therefore, we have

$$\frac{1600(\text{slots/s})}{6(\text{slots})} \times 240(\text{bits}) = 64\,\text{kb/s}$$

The overall format of the payload for the six *ACL* packets is shown in Fig. 10.12. The payload has its own 8- or 16-bit header, payload, and 16-bit CRC code. The header has information on the length and identity of the packet. If we want to compare the headers with those of 802.11, we may compare the overhead with the MAC overhead of the 802.11 shown in Fig. 9.13. This time the overhead of Bluetooth is significantly lower than the 34 bytes (272 bits) overhead of the 802.11 MAC frames. Most of the saving in the overhead of Bluetooth occurs because 802.11 employs four addresses: source, destination of the device, and the two intermediate APs. Bluetooth uses one 48-bit IEEE MAC address to identify a device that is embedded in the access code and is not needed in the payload.

The six ACL packets are data medium (DM) and data high (DH) rate packets, numbered DM1/DH1, −3, or −5 according to the length of the slot they take. Figure 10.12

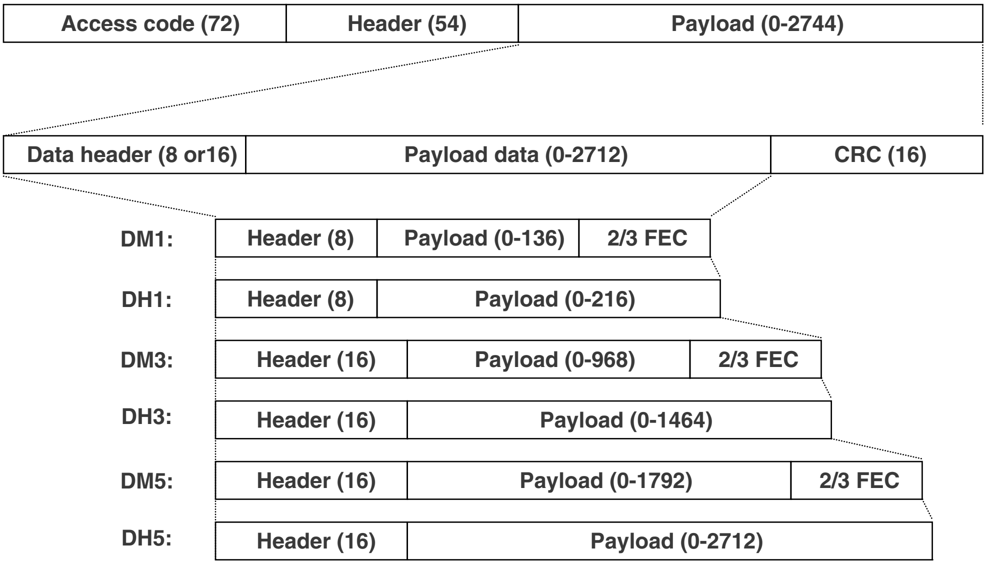


FIGURE 10.12 ACL's one-, three-, and five-slot packet frame formats.

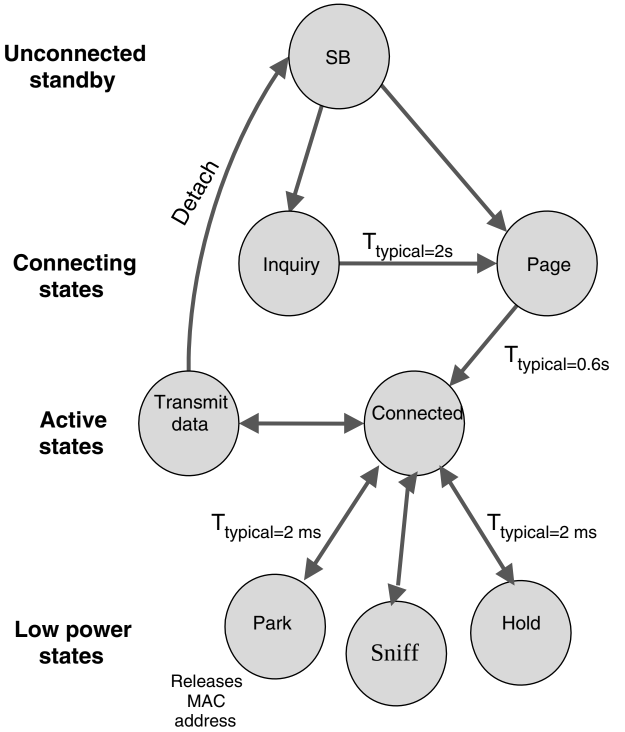


FIGURE 10.13 Functional overview of the Bluetooth specification.

shows the overall frame format of all DM and DH data-oriented packets. DM packets use a rate $^2/_3$ FEC that improves the QoS. DH packets do not employ coding to achieve higher data rates. By using different numbers of slots for a packet data payload size, exercising the coding option, and changing the symmetric nature of the transmitted packets in each direction, a number of packet data links can be implemented in the Bluetooth specification.

Example 10.7: High Data Rate in Bluetooth A symmetric one-slot DH1 link between an M and an S terminal carries 216 bits per slot at a rate of 800 slots per second (every other slot) in each direction. The associated data rate is 216 (bits/slot) × 800 (slots/s) = 172.8 (kb/s).

Example 10.8: Medium Data Rate in Bluetooth The asymmetric DM5 link, shown in Fig. 10.9*d* uses five-slot packets carrying 1792 bits per packet by the M and one-slot packet carrying 136 bits per packet by the S terminal. The number of packets per second in each direction is 1600/6 packets per second. Therefore, the data rate from M is given by

$$1792\ (\text{bits/packet}) \times \frac{1600}{6}\ (\text{packets/s}) = 477.8\ (\text{kb/s})$$

TABLE 10.1 ACL Packet Types and Associated Data Rates in Symmetric and Asymmetric Modes

Type	Symmetric	Asymmetric	
DM1	108.8	108.8	108.8
DH1	172.8	172.8	172.8
DM3	256	384	54.4
DH3	384	576	86.4
DM5	286.7	477.8	36.3
DH5	432.6	721	57.6

The data rate of the S terminal in this asymmetric connection is

$$136\ (\text{bits/packet}) \times \frac{1600}{6}\ (\text{packets/s}) = 36.3\ (\text{kb/s})$$

Table 10.1 shows all 12 symmetric and asymmetric data links that are supported with the frame format of the Bluetooth specification. The maximum data rate of 723.2 kb/s is available in an asymmetric channel for a single user, while the reverse channel carries 57.6 kb/s. The reader should remember that data applications operate in bursts; therefore, even if an M node communicates with the maximum seven S data terminals, then still most of the time only one of the S terminals will communicate with the M. When more than one S terminal simultaneously attempts to communicate with an M terminal, the QoS provided to the S terminals has to be compromised either by sharing the throughput or by providing additional delays. The decision-making process to reach a compromise in the voice-oriented access methods, such as the one used in Bluetooth, needs a complex algorithm to handle the QoS as negotiated at the start of a session. Comparing this situation with CSMA/CA used in 802.11, there is no negotiation at the starting point. When more than one terminal attempts to communicate with a single AP, the medium is shared and the compromise is made automatically through the CSMA/CA access method described in Chapter 4. Apparently, CSMA/CA is more appropriate for data-only applications, and that is why it was developed by the data-oriented networking industry. However, when voice applications become dominant, the TDMA/TDD-type access methods can guarantee QoS for voice, whereas CSMA/CA cannot do it easily.

The only remaining traffic packet in Bluetooth is a data voice (DV) packet, which is a mixed SCO and ACL packet with the same access code and overall header that must be transmitted in regular intervals. The voice part carries 80 bits of voice payload without any coding and the data part is a short packet of length 0–72 bits with a 16-bit $^{2}/_{3}$ CRC coding and an 8-bit data payload header. This packet also uses the three status-report bits.

The Bluetooth specification also defines four control packets: ID, NULL, POLL, and FHS. The ID packet occupies only half of a slot and it carries the access code with no data or even a packet type code. This packet is used before connection establishment only to pass an address. The NULL and POLL packets have the access code and the header, and so they have packet-type codes and status report bits. The NULL packet is used for ACK signaling and there is no ACK packet for it. The POLL packet is similar to the NULL packet, but it has an ACK. The M terminals use the POLL packet to find the S terminals in their coverage area.

The frequency hop synchronization (FHS) packet carries all the information necessary to synchronize two devices in terms of access code and hopping timing. This packet is used in the inquiry and paging process that will be explained later.

10.2.6 Connection Management

The link manager (LM) layer and L2CAP layer of Bluetooth perform the link setup, authentication, and link configuration. An important issue in a truly ad hoc network is how to establish and maintain all the connections in a network whose elements appear and disappear in an ad hoc manner and there is no central unit transmitting signals to coordinate these terminals. In both digital cellular systems and WLANs there is a common control or a beacon signal that allows a new terminal to lock to the network and exchange its identity with the networks identity. The Bluetooth specification achieves initiation of the network through a unique inquiry and page algorithm.

The overall state diagram of Bluetooth is shown in Fig. 10.13. In the beginning of the formation of a piconet, all devices are in SB mode, and then one of the devices starts with an inquiry and becomes the M terminal. During the inquiry process the M terminal registers all the SB terminals, which then become S terminals. After the inquiry process, the identification and timing of all S terminals is sent to the M terminal using the FHS packets. A connection starts with a PAGE message, with which the M terminal sends its timing and identification to the S terminal. When connection is established, the communication session takes place and, at the end, the terminal can be sent back to the SB, hold, park or sniff states. Hold, park, and sniff are power-saving options. The hold mode is used when connecting several piconets together or managing a low-power device. In the hold mode, data transfer restarts as soon as the unit is out of this mode. In the sniff mode, a slave device listens to the piconet at reduced and programmable intervals according to the application needs. In the park mode, a device gives up its MAC address but remains synchronized to the piconet. A parked device does not participate in the traffic, but it occasionally listens to the traffic of the M terminal to resynchronize and check on broadcast messages.

The main innovative part of the inquiry and paging algorithms in Bluetooth is a searching mechanism for two terminals that are not synchronized but which both know a common address. The following example explains this algorithm.

Example 10.9: Search Algorithm for Synchronization Two Bluetooth devices with a common 48-bit IEEE 802 address first use the common address to generate a common FH pattern of 32 hops and a common PN sequence for the access code of all their packets. Then they start their operation as depicted in Fig. 10.14. In the initial state, terminal 1 sends two ID packets carrying the common access code every half slot on a different hop frequency associated with the common FH pattern and listens to the response of the slave in the next slot. If there is no response, then it continues broadcasting the ID packets on the two new frequencies in the common hop pattern and repeats this procedure eight times for a period of 10 ms (eight 2-slot times). During these 10 ms, the common ID is broadcast at 16 of the total 32 different hop frequencies. If there is no response, then terminal 1 assumes that terminal 2 is in sleep mode and repeats the same broadcast again and again until the period of transmission becomes longer than the expected sleeping time of terminal 2. At this time, terminal 1 assumes that terminal 2 has scanned but its scan frequency was not

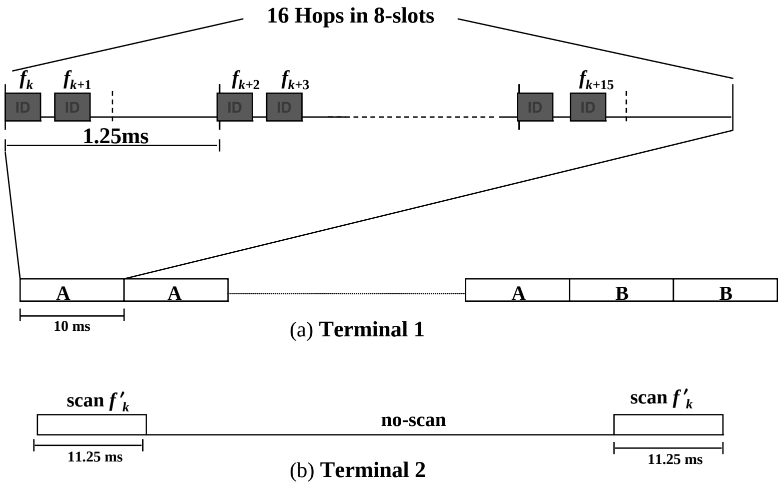


FIGURE 10.14 Basic search for paging algorithm in the Bluetooth.

among the 16 hops, designated by A in Fig. 10.14, and continues its broadcast with the second half of the 32 hop frequencies, designated by B in the figure. Terminal 2 is in sleeping mode; it wakes up periodically for a period of 11.25 ms to scan the channel at a given frequency for its desirable access code and sleeps again. In each scan period of 11.25 ms the sliding correlator in terminal 2 hears the desired address at 16 different frequencies. If one of these frequencies is the same as the scanning frequency, then the correlator peaks and synchronization is signaled. Depending on the operation, terminal 2 can scan the second time at the same frequency or at a new frequency for verification. In either case, the objective is to maximize the probability of hitting the same frequency as the broadcast frequency.

The basic principle explained in the above example is used during the inquiry and paging processes. The following two examples explain these applications for the above mechanism.

Example 10.10: Paging As in the previous example, the M terminal broadcasts repeating ID page trains carrying the access code of the paged terminal, two per slot, waits for the response in the next slot, repeats at new hopping frequencies of the paged terminal to cover 16 frequencies every 10 ms, and repeats this for the estimated length of the sleeping time. The S terminal scans for 11.25 ms with one of the 32 frequencies of its hopping pattern, sleeps, and scans at the next hopping frequency. When the frequencies are the same, a peak appears at the correlator output of the S terminal and the slave responds by sending its own ID packet as an acknowledgement for detection of FH timing. The M terminal then stops broadcasting ID packets and sends an FHS packet containing its own ID and timing information. The S terminal responds with another ID packet corresponding to the timing of the M terminal and then connection is established and the slave joins the piconet for information exchange. Usually, the M terminal knows the approximate timing of the hopping pattern and the 16 most-probable hops are adequate to establish the connection. In the case that this estimate is not correct, like the previous example, then the M terminal resorts to the second half of the 16 hops when there is no response after the estimated sleeping time.

Example 10.11: Inquiry The inquiry message is typically used for finding Bluetooth devices, including public printers, fax machines, and similar devices with an unknown address. The general format of the inquiry process is very similar to the paging mechanism. A unique access code and FH pattern are reserved for inquiry. In other words, the inquiry process is universally identified with all attributes of a device. Like paging, inquiry starts with an "inquirer" broadcasting an ID packet every half slot at a different hop frequency, covering 16 frequencies every 10 ms, and repeats the same process until it receives responses. The "inquiree" scans with the sliding correlator for 11.25 ms. When frequencies are the same, the sliding correlator peaks in all devices that are scanning. To avoid collision, a device detecting the inquiry ID runs a random-number generator and waits for the length of the outcome before it scans the channel again. When the peak appears the second time after a random waiting time, the inquiree terminal sends an FHS packet, allowing the inquirer to learn its ID and timing information. After the process is completed, the inquirer's radio has device IDs and clocks of all radios in its range of coverage. After completion of the first inquiry, the inquired device changes its scan frequency and continues scanning for the next inquiry and follow-up FHS signaling.

10.2.7 Security

Bluetooth specifications provide usage protection and information confidentiality. Bluetooth has three modes of operation: nonsecure, service-level security, and link-level security. Devices also can be classified into trusted and distrusted. It makes use of two secret keys (128 bits for authentication and 8–128 bits long for encryption): a 128-bit long random number and the 48-bit MAC address of devices. Any pair of Bluetooth devices that wish to communicate will create a session key (called the link key) using an initialization key, the device MAC address, and a PIN. This protocol has been shown to have several vulnerabilities [Wet01] by which a malicious entity could obtain the PINs and keys, depending on how the session initialization of the communication protocol is performed.

10.3 INTERFERENCE BETWEEN BLUETOOTH AND 802.11

Obviously, when two wireless networks overlap in their coverage and operate at the same frequency at the same time without any access coordination, they will interfere with one another. The literature on military communication systems offers many detailed analyses of the performance of communication systems in the presence of various intentional interferers or jammers [Sim85]. These jammers are designed to disrupt the operation of a system and they can employ relatively sophisticated techniques, such as multitone jamming and pulsed jamming. In civilian applications, the interference is neither intentional nor sophisticated. Most often, the interferer is simply another system designed to operate in a portion of or the entire band of operation of our system and the users are generally willing to cooperate to minimize the mutual interference. Depending on the level of coordination of the overlapping wireless network, since the early days of the IEEE 802.11 [Hay91, IEE01], the WLAN industry has specified three levels of overlapping: interference, coexistence, and interoperation.

Multiple wireless networks are said to *interfere* with one another if collocation causes significant performance degradation of any of the devices. Multiple wireless networks are said to *coexist* if they can be collocated without significant impact on the performance of any of the devices. Coexistence provides for the ability of one system to perform a task in a shared frequency band with other systems that may or may not be using the same set of rules for operation. *Interoperability* provides for an environment for multiple overlapping wireless systems to perform a given task using a single set of rules. In an interoperable environment, multiple wireless networks exchange and use the information among each other. Interoperability is an important issue for wired and wireless networks. Coexistence and interference are issues mainly consuming the attention of wireless network designers, and it becomes more important for the case of ad hoc networks. This terminology for unlicensed bands was first discussed in the IEEE 802.11 community [Hay91]. Later, when WINForum approached the FCC to obtain unlicensed PCS bands, they came up with *etiquettes* or rules of coexistence in unlicensed PCS bands [Pah97], which was discussed in Chapter 10. More recently, the IEEE 802.15 WPAN group has been engaged in interference analysis in its task group number two (TG-2). They have performed introductory interference analysis between Bluetooth and IEEE 802.11 devices operating in 2.4 GHz ISM bands and, at the time of writing, are working on practical coexistence and interoperability methods [IEE01, ENN98].

Bluetooth is a fast FH (1600 hops/s at 1 Mb/s) wireless system operating in 84 MHz of bandwidth that is available in the 2.4 GHz ISM bands that are also used for DSSS IEEE 802.11 (1 and 2 Mb/s) and CCK IEEE 802.11b (5.5 and 11 Mb/s), as well as in slower FHSS (2.5 hops/s at 1 and 2 Mb/s) IEEE 802.11 systems. Therefore, the interaction between a Bluetooth system and a collocated 802.11 WLAN system needs an analysis of the interference between the FHSS and DSSS, as well as between fast FHSS and slow FHSS systems.

10.3.1 Interference Range

The first issue in interference is the *interference range*, which is the distance between two terminals in order to interfere, in case they operate at the same frequency and at the same time. The range of interference is related to the propagation characteristics of the environment, processing gain of the receivers, and the transmitted power from different devices.

Figure 10.15 illustrates an interference scenario between a Bluetooth (BT-1) device and a receiving FHSS IEEE 802.11 mobile terminal (MT) collocated in an area. The IEEE 802.11 AP is usually located on a wall to provide better coverage; as a result, they are usually less likely to be interfered with by the Bluetooth devices. The interference takes place both when the mobile terminal is receiving information from the AP and BT-1 is transmitting information to BT-2, and when the mobile terminal is transmitting and BT-1 is receiving. For our analysis we assume that interference from the AP to the Bluetooth devices and interference of the BT-2 device to 802.11 devices are negligible. Following the same analysis for interference presented in Chapter 7, when the mobile terminal is

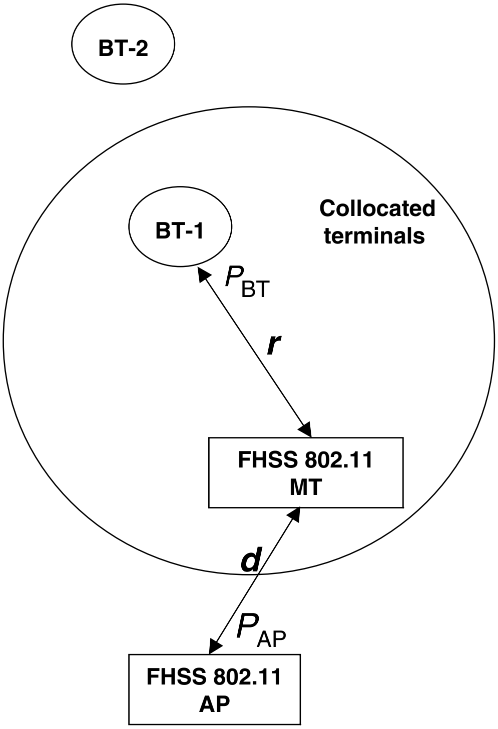


FIGURE 10.15 The basic interference scenario between Bluetooth and IEEE 802.11 FHSS.

receiving and BT-1 is transmitting, the signal-to-interference level at the mobile terminal is given by

$$S_{\mathrm{r}} = \frac{KP_{\mathrm{AP}}d^{-\alpha}}{KP_{\mathrm{BT}}r^{-\alpha}} = \frac{P_{\mathrm{AP}}}{P_{\mathrm{BT}}}\left(\frac{r}{d}\right)^{\alpha} \tag{10.1}$$

where d and r are the distance from the MT to the AP and the distance between the two Bluetooth devices respectively. Also, P_{AP} and P_{BT} represent the transmitted power by the AP and the Bluetooth device respectively, and α is the distance power gradient of the propagation environment. Therefore, the *range of interference* between the Bluetooth device and the mobile terminal is given by

$$r_{\mathrm{int}} = d\sqrt[\alpha]{S_{\min}P_{\mathrm{BT}}/P_{\mathrm{AP}}} \tag{10.2}$$

where r_{int} is the maximum distance at which the two terminals interfere and $S_{\min}$ is the minimum acceptable received SNR needed for proper operation of the mobile terminal. In other words, the range of interference of the BT-1 terminal to the mobile terminal is directly related to the distance to the AP, required SNR for proper operation of the mobile terminal, and transmit power of BT-1, and it is inversely related to the transmit power of the AP. In general, as we discussed in Chapter 2, the value of α may change from <2 in hallways and open areas up to ~ 6 in a building with metal partitioning. Depending on the location of the Bluetooth device, the path-loss gradients may be different as well. In open areas with no walls, which includes a number of scenarios involved with short-range devices, the environment is close to free-space propagation and α is often close to 2 [Pah95]. Although the coverage of the 802.11 devices is estimated to be 100 m (at 20 dBm transmit power), 802.11 APs in practice are installed every 20–40 m, allowing maximum distances of $d = 10$–20 m between an AP and a mobile terminal. The low-power (0 dBm transmit power) Bluetooth devices are used for WPAN applications where r, the distance between the devices, is only a few meters. Bluetooth also allows 20 dBm operation, which can cover up to 100 m.

Example 10.12: Interference Range between Bluetooth and 802.11

1. Assuming an open area with $\alpha = 2$, $S_{\min} = 10$ (10 dB), $P_{\mathrm{AP}} = 100$ mW (20 dBm), $P_{\mathrm{BT}} = 1$ mW (0 dBm), and $d = 20$ m, we have $r_{\mathrm{int}} = 6.4$ m. That means that if the frequencies of the BT-1 and mobile terminal are the same and BT-1 transmits at the same time that the mobile terminal receives, then a BT-1 device that is closer than 6.4 m to the mobile terminal will interfere and destroy the received packets.
2. In a partitioned environment, with $\alpha = 4$ we will have $r_{\mathrm{int}} = 10.2$ m.
3. If the Bluetooth device is at its maximum transmit power of 100 mW (20 dBm) in the same partitioned area, then $r_{\mathrm{int}} = 17.7$ m, which is an order of magnitude larger than the value in the 0 dBm mode.

Figure 10.16 illustrates the simple scenario for the interference of FHSS 802.11 to Bluetooth terminals. In this case the mobile terminal is transmitting to the AP and BT-1 is receiving from BT-2. If we assume that the transmitting mobile terminal is at a distance r from BT-1 and the two Bluetooth devices are a distance d apart, then we have

$$r_{\mathrm{int}} = d \sqrt[\alpha]{S_{\min} P_{\mathrm{MT}} / P_{\mathrm{BT}}} \tag{10.3}$$

Again, the range of interference is directly proportional to the distance between desired terminals, the minimum acceptable SNR of the receiving terminal, and the power of the interfering terminal, and inversely proportional to the power of the desired transmitter.

Example 10.13: Another Example for Calculating r_{int}

1. With typical values of 2 m for the distance between the two Bluetooth devices, $P_{\mathrm{BT}} = 1$ mW (0 dBm), $S_{\min} = 10$ (10 dB), and $P_{\mathrm{MT}} = 100$ mW (20 dBm) we will have $r_{\mathrm{int}} = 63.2$ m. This is because the 802.11 device is radiating 100 times more power.
2. If the Bluetooth device operates at 20 dBm, with the same power as the 802.11, then $r_{\mathrm{int}} = 6.32$ m.

If, instead of FHSS, we use DSSS in the scenario of Fig. 10.15, then, as we discussed in Chapter 3, the minimum required received signal-to-interference ratio at the mobile

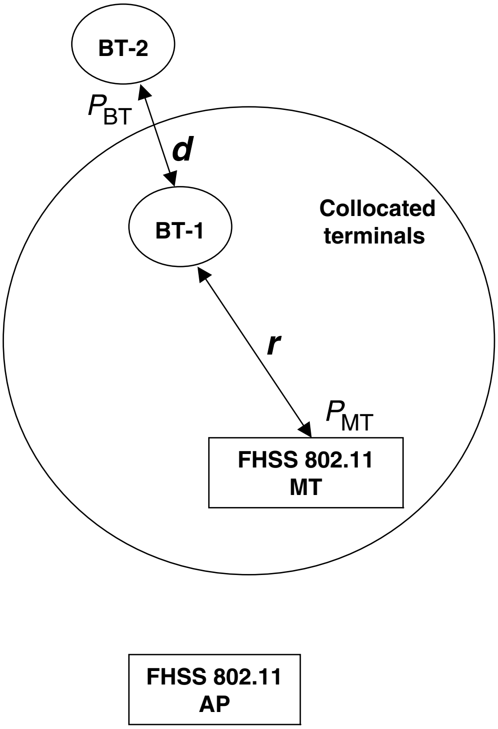


FIGURE 10.16 The basic interference scenario between FHSS IEEE 802.11 and BT.

terminal is reduced by a factor equivalent to the value of the processing gain of the DSSS, N. Then, the interference range (BT-1 interfering with mobile terminal) becomes

$$r_{\text{int}} = d\sqrt[\alpha]{S_{\min} P_{BT}/P_{AP}N} \tag{10.4}$$

For the case of a mobile terminal interfering with Bluetooth, the spectral height of the DSSS is reduced by the value of the processing gain, which results in a similar effect and a range of

$$r_{\text{int}} = d\sqrt[\alpha]{S_{\min} P_{\text{MT}}/P_{BT}N} \tag{10.5}$$

Example 10.14: Interference between Bluetooth and DSSS-based IEEE 802.11

1. Assume an open area with $\alpha = 2$, $S_{\min} = 10$ (10 dB), $P_{\text{AP}} = 100$ mW (20 dBm), $P_{\text{BT}} = 1$ mW (0 dBm), and $d = 20$ m. For a processing gain of $N = 11$, used in IEEE 802.11, the interference range will reduce to around $r_{\text{int}} = 1.9$ m (from 6.4 m) (BT-1 interfering with the mobile terminal).
2. For $P_{\text{BT}} = 100$ mW (20 dBm) we have an interference range of $r_{\text{int}} = 19$ m.

3. With a typical values of 2 m distance between the two Bluetooth devices, $P_{\mathrm{BT}} = 1$ mW (0 dBm), $S_{\min} = 10$ (10 dB), $P_{\mathrm{MT}} = 100$ mW (20 dBm), and $N = 11$ we will have $r_{\mathrm{int}} = 19$ m (the mobile terminal interfering with BT-1).
4. If the Bluetooth device operates at 20 dBm power transmission option, then $r_{\mathrm{int}} = 1.9$ m.

The conclusion from these simple examples is that considering the 10 m range of operation of a Bluetooth piconet and a 100 m range of operation of the 802.11 devices, if a Bluetooth hop coincides with the frequency of an FHSS or DSSS IEEE 802.11 WLAN, then the interference is serious. The DSSS reduces the interference of the narrowband systems and interference to the narrowband system by the value of its processing gain. This results in a $\sqrt[\alpha]{1/N}$ reduction in the range of interference compared with an FHSS system. However, the spectrum of DSSS is much wider and the probability of frequency coincidence of the DSSS and Bluetooth is much higher than the probability of hit frequency coincidence of an FHSS system and Bluetooth. In the next section we quantify this statement further.

10.3.2 Probability of Interference

In the last section we showed that the range of interference of Bluetooth and IEEE 802.11 DSSS is smaller than the range of interference of Bluetooth and FHSS 802.11 systems. However, FHSS is a narrowband signal that changes its frequency of operation randomly, while a DSSS is a true wideband system. A narrowband Bluetooth transmitter will interfere with the reception of a wideband DSSS signal with a greater probability than it will with the reception of an FHSS signal on a different narrowband channel. Therefore, the probability of interference between Bluetooth and 802.11 DSSS or 802.11b CCK devices is much higher than the probability of interference between a Bluetooth device and an FHSS 802.11 system.

To analyze the interference further, we first pay attention to interference between Bluetooth and FHSS 802.11 devices. Both Bluetooth and FHSS 802.11 are FH systems using the 79 carrier frequencies in the 2.4 GHz ISM bands shown in the vertical axis of Fig. 10.17. Bluetooth packets are normally shorter than 802.11 packets and hop at a much slower rate of 2.5 hops per second. When a terminal is in the interference range of the other terminal and the hopping frequencies are the same, then packets collide and get destroyed. To analyze this situation we need to find the probability of collision in time and in frequency.

Since Bluetooth packets are shorter than 802.11 packets, during transmission of one 802.11 packet, the collocated Bluetooth device hops and sends one packet per hop several times. Assuming L_{IE} is the length of the IEEE 802.11 packet and L_{BS} the length of a Bluetooth slot, then the minimum number of Bluetooth hops occurring during transmission of one 802.11 packet is $n = \lceil L_{\mathrm{IE}}/L_{\mathrm{BS}} \rceil$, where $\lceil x \rceil$ represents the smallest integer greater than or equal to x. The maximum number of Bluetooth hops occurring in duration of an 802.11 packet is $\lceil L_{\mathrm{IE}}/L_{\mathrm{BS}} \rceil + 1$. It can be easily shown [ENN98] that the probability of an 802.11 packet overlap with $n = \lceil L_{\mathrm{IE}}/L_{\mathrm{BS}} \rceil$ Bluetooth dwell periods of duration L_{BS} is

$$P_n = L_{\mathrm{IE}}/L_{\mathrm{BS}} - \lceil L_{\mathrm{IE}}/L_{\mathrm{BS}} \rceil$$

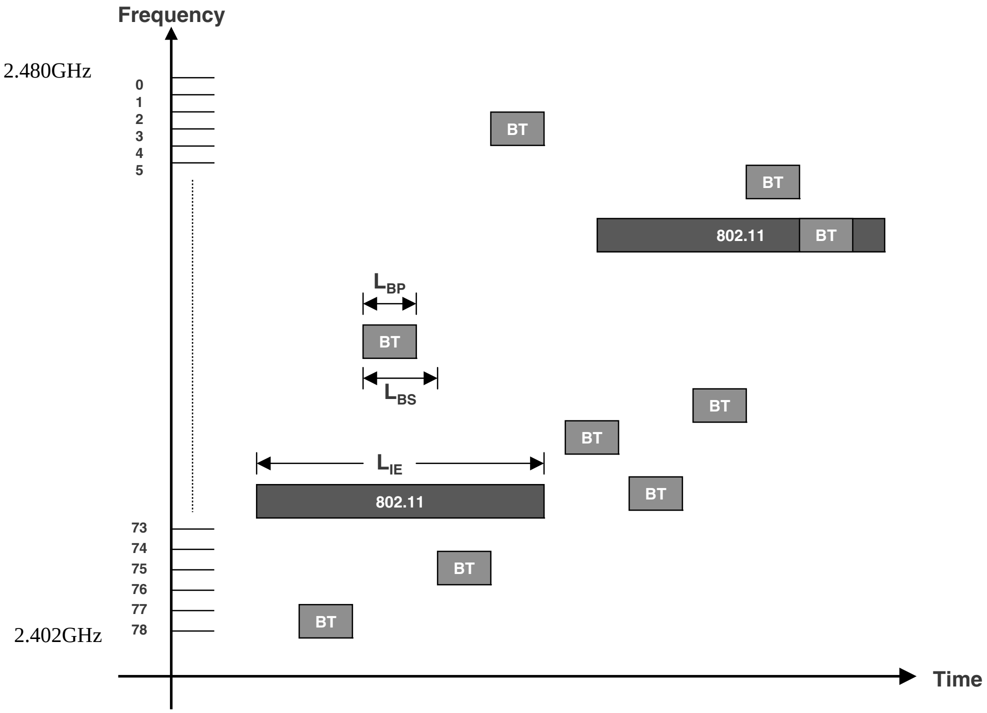


FIGURE 10.17 Time–frequency characteristics of the FHSS IEEE 802.11 and Bluetooth.

The probability that it overlaps with $n+1 = \lceil \mathrm{L}_{\mathrm{IE}}/L_{\mathrm{BS}} \rceil + 1$ dwell periods is

$$P_{n+1} = 1 - L_{\mathrm{IE}}/L_{\mathrm{BS}} + \lceil L_{\mathrm{IE}}/L_{\mathrm{BS}} \rceil$$

Example 10.15: Probability of Interference between Bluetooth and FHSS IEEE 802.11 If $L_{\mathrm{IE}}/L_{\mathrm{BS}} = 4.3$, then the probability of overlap of an 802.11 packet with $n = 4$ Bluetooth dwell periods is 30% and the probability of overlap with $n+1 = 5$ dwell periods is 70%.

Considering these expressions, the probability of an 802.11 packet surviving Bluetooth interference P_{survive} is approximated by

$$P_{\mathrm{survive}} = (1 - P_{\mathrm{hit}})^n P_n + (1 - P_{\mathrm{hit}})^{n+1} P_{n+1}$$

where P_{hit} is the probability of having the same frequency for both 802.11 and Bluetooth. The probability of collision is given by $P_{\mathrm{collision}} = 1 - P_{\mathrm{survive}}$.

Example 10.16: Collision Probability between IEEE 802.11 FHSS and Bluetooth The probability of a Bluetooth hop to occur at the operating frequency of the FHSS system is $P_{\mathrm{hit}} = 1/79 = 0.013$. For a 1000-byte 802.11 packet at 2 Mb/s:

$$L_{\mathrm{IE}} = \frac{1000\,(\text{bytes}) \times 8\,(\text{bits/byte})}{2(\text{Mb/s})} = 4\,\text{ms}$$

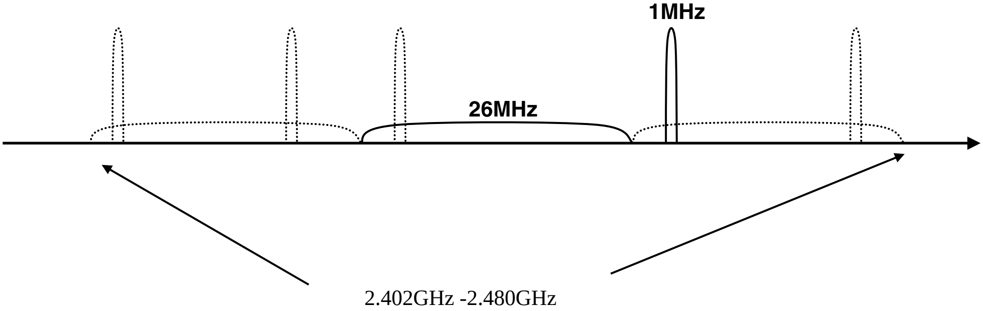


FIGURE 10.18 Overlapping DSSS IEEE 802.11 and FHSS Bluetooth spectrum.

If Bluetooth is sending one-slot packets, $L_{\text{BS}} = 625\,\mu\text{s}$. Therefore:

$$n = \left\lceil \frac{4\text{ ms}}{625\,\mu\text{s}} \right\rceil = 6$$

and $P_n = 0.4$, which results in $P_{n+1} = 0.6$. Therefore:

$$P_{\text{survive}} = (1-0.013)^6 \times 0.4 + (1-0.013)^7 \times 0.6 = 0.92$$

Therefore, the collision probability is 0.08 or 8%.

Example 10.17: Collision Probability between IEEE 802.11 DSSS and Bluetooth Figure 10.18 shows the mechanism with which the FH pattern of Bluetooth and the spectrum of the DSSS 802.11 or CCK 802.11b hit one another. The probability of a Bluetooth hop occurring at the operating frequency of the DSSS system is $P_{\text{hit}} = 26/78 = 0.33$. For a 1000-byte 802.11 packet at 2 Mb/s, all the other parameters remain the same as the last example and we will have

$$P_{\text{survive}} = (1-0.33)^6 \times 0.4 + (1-0.33)^7 \times 0.6 = 0.072$$

The probability of collision is 0.928 or 92.8%, compared with 8% for the FHSS 802.11 example.

Example 10.18: Bluetooth Interference with IEEE 802.11b IEEE 802.11b uses the same band as 802.11 DSSS to transmit at 11 Mb/s. Therefore, again we have $P_{\text{hit}} = 26/79 = 0.33$. However, for a 1000-byte 802.11 packet at 11 Mb/s we have

$$L_{\text{IE}} = \frac{1000(\text{bytes}) \times 8(\text{bits/byte})}{11(\text{Mbps})} = 727\,\mu\text{s}$$

With Bluetooth one-slot packets we have $n = \lceil 727\text{ ms}/625\,\mu\text{s} \rceil = 1$ and $P_n = 0.16$, which results in $P_{n+1} = 0.84$. Therefore:

$$P_{\text{survive}} = (1-0.33)^1 \times 0.16 + (1-0.33)^2 \times 0.84 = 0.49$$

The collision probability is 0.51 or 51%, which is substantially better than 802.11 DSSS and much worse than 802.11 FHSS.

10.3.3 Empirical Results

The analysis in the last section is at the PHY layer, but a more thorough analysis including the effects of all layers should be done experimentally. A group of undergraduate students at WPI developed a testbed for the experimental analysis of the interference between the IEEE802.11b and Bluetooth voice and data channels for their senior undergraduate project [Cha00]. In this project they considered a number of scenarios and measured the overall packet loss, throughput, and delay characteristics of the interfering Bluetooth and 802.11 devices as well as cordless telephones. In this section we provide some of their results and conclusions that are related to the scenarios described in Figs 10.15 and that relates the performance of interfering 802.11b and BT terminals to the distance between the devices.

Example 10.19: PLR in Bluetooth with Interfering IEEE 802.11b Devices Figure 10.19 shows the floor plan and one of the measurement scenarios in which two 20 dBm Bluetooth-equipped laptops (triangles) are separated by 10 m and an 802.11b laptop (circle) is moved from a distance of 1 to 10 m from the Bluetooth laptop. The 802.11b station is communicating with another laptop that is far away and does not interfere significantly with the Bluetooth device. Figure 10.20 shows the PLR of the Bluetooth device. As the distance of the interfering 802.11 device increases, so the packet loss

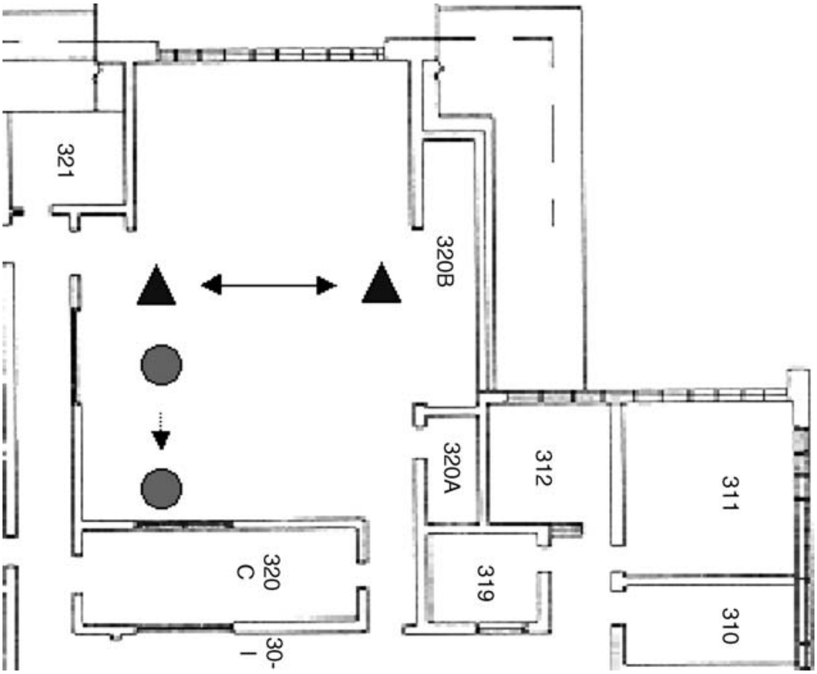


FIGURE 10.19 Bluetooth interfering scenario for the experimental interference analysis.

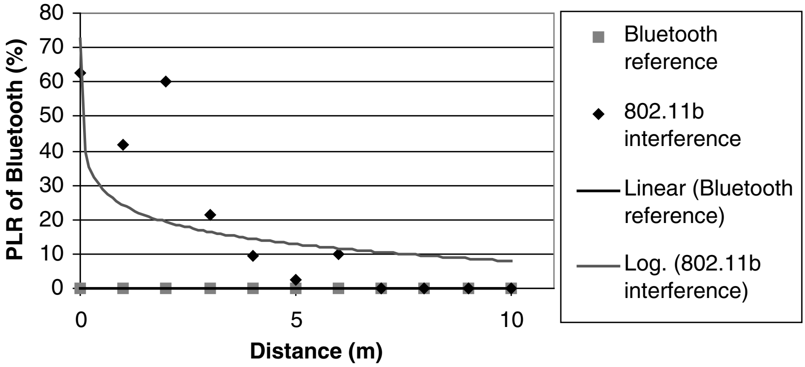


FIGURE 10.20 PLR of Bluetooth with and without 802.11b interfering terminal in an open office environment ad hoc network.

reduces. When the Bluetooth and 802.11b interferers are next to one another the PLR is 70%; as the distance increases to 5 m, there is no interference effect. In this experiment, the lengths of the 802.11 packets are 1000 bytes and Bluetooth data packets are 366 bits long. Figure 10.21 shows the delay characteristics evaluated a using ping message, which measures the round-trip delay.

Example 10.20: PLR in IEEE 802.11b with an Interfering Bluetooth Device Figure 10.22 shows the PLR of 802.11b in a scenario that is the opposite of the last example shown in Fig. 10.18. In this example, two 802.11b devices are located at a distance of 10 m and an interfering Bluetooth terminal is moved from 1 to 10 m from them. At close distances, the PLR is close to 45%; as the distance of the interfering Bluetooth device increases beyond 3 m the effect of interference is negligible.

The general conclusion of these studies is that the interference between FHSS 802.11 and Bluetooth devices is negligible; however, DSSS 802.11 devices will interfere significantly with Bluetooth devices. IEEE 802.15 is currently working on this issue to find remedies for the coexistence of these systems [IEE00].

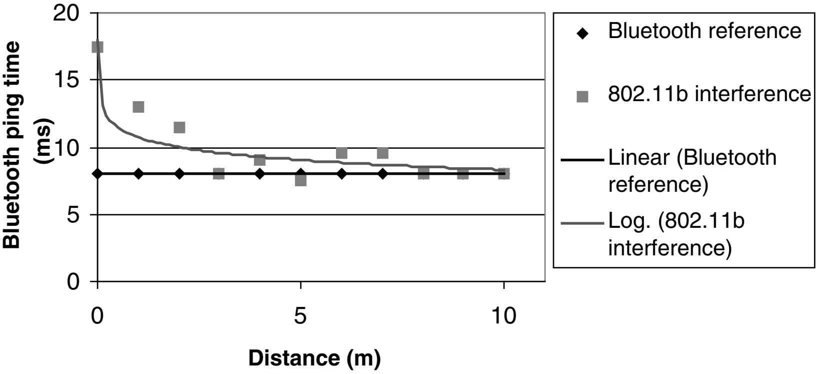


FIGURE 10.21 Bluetooth delay characteristics with and without 802.11b interference in an open office environment ad hoc network.

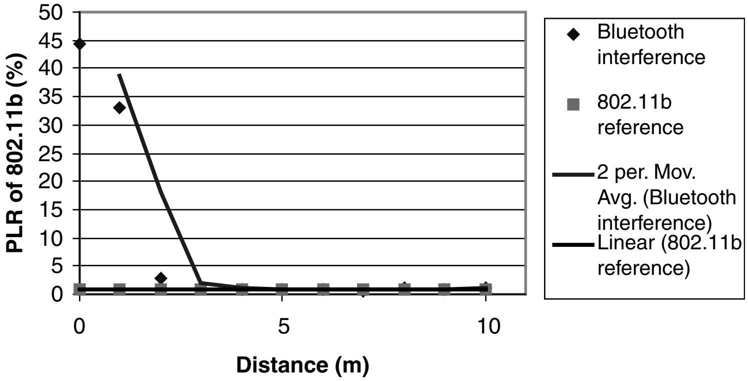


FIGURE 10.22 PLR of 802.11b with and without interfering Bluetooth device in an open office environment ad hoc network.

10.4 IEEE 802.15.3A ULTRA WIDEBAND WIRELESS PERSONAL-AREA NETWORKS

In around 2000, a new wave of interest in UWB communication for military and commercial applications began with the idea of impulse radio transmission [Win00]. The two attractive features of UWB were the ability to provide extremely high wireless data rates in the order of gigabits per second and the potential capability to support precise localization in multipath-rich indoor areas. In response to this interest, in 2002 the IEEE 802 community established a working group to develop standards for UWB WPANs. This group went under IEEE 802.15.3a for the design of PHY and MAC. In actual operation of the impulse radio, the UWB pulses interfered with many existing systems, including cellular systems, WLANs, and GPSs [Ham02]. In particular, GPSs operating with very low signal strength were vulnerable to UWB interference. As a result of these concerns, the FCC decided to mandate the frequency of operation for UWB devices to 3.1–10.6 GHz. This decision in 2003 redirected the attention of the IEEE 802.15.3a UWB standardization activities toward this band; consequently, impulse radio using the lower frequency bands lost its support in the standardization committee. Two leading 802.15 proposals brought forward after the 2003 FCC announcement are known as direct sequence UWB (DS-UWB) and MB-OFDM. The DS-UWB technique is addressed in the remainder of this section and the MB-OFDM technique is described in the next section. The basic coverage cell in the WPAN industry is referred to as a *picocell* having a nominal coverage range of about 10 m. A network operating within that range is referred to as piconet. Different WPAN technologies support different numbers of overlapping piconets. For example, Bluetooth, the first WPAN standard under IEEE 802.15.1, supports seven overlapping piconets. The new UWB proposals consider multiple bands that can be combined with MAC to support substantially larger numbers of overlapping piconets.

10.4.1 Direct Sequence Ultra Wideband

The DS-UWB system uses the DSSS technique, which successfully emerged as the PHY layer of choice in 3G cellular networks. This technique employs BPSK and QPSK

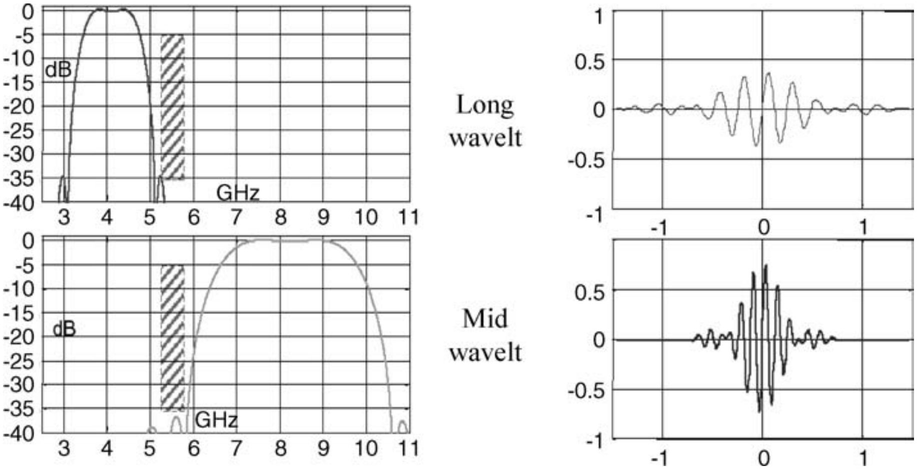


FIGURE 10.23 Frequency and time response of the two basic channels in DS-UWB proposal [Koh04].

modulation and a MAC that combines FDM, TDM, and CDM. In the DS-UWB system, as shown in Fig. 10.23, the 3.1–10.6 GHz band is divided into a low band from 3.1 to 4.9 GHz and an optional high band from 6.2 to 9.7 GHz. The bandwidth of the high band is twice the bandwidth of the low band, resulting in shorter time-domain pulses in the high band. The 4.9–6.1 GHz band is purposely neglected to avoid interference with IEEE 802.11a devices operating in the 5 GHz U-NII bands. Each piconet of the DS-UWB operates in one of the two bands, and piconets in the same band are separated by CDM using ternary multiple bi-orthogonal keying (M-BOK) spreading codes of length 24 or 32.

Example 10.21: Data M-BOK Codes Table 10.2 shows the M-BOK ternary codes of length 24 used in the DS-UWB system with BPSK. Each chip can take three values: $\{-1, 0, 1\}$. For $M = 2$, the first row or its reverse are used to identify two symbols, and each incoming information bit is mapped to one of the two symbols. For $M = 4$, the first two rows and their reverses are used to form four symbols. The incoming information bits are grouped into 2 bits and one of the four symbols is selected for each information di-bit. For $M = 8$, all four rows and their inverses are used to form 8-BOK codes. The incoming information bits are grouped into 3-bit blocks, each selecting one of the 8-BOK codes. For more details on the longer codes one can refer to Welborn *et al.* [Wel03]. Figure 10.24 shows the comparative

TABLE 10.2 M-BOK Ternary Codes of Length 24 for $M = 2$, 4, and 8[a]

Code no.	Code
1	−1 1 −1 −1 1 −1 −1 1 −1 0 −1 0 −1 −1 1 1 1 −1 1 1 1 −1 −1 −1
2	0 −1 −1 0 1 −1 −1 1 −1 −1 1 1 1 1 −1 −1 1 −1 1 −1 1 1 1 1
3	−1 −1 −1 −1 1 −1 1 −1 1 −1 −1 1 −1 −1 1 −1 −1 1 1 0 −1 0 1 1
4	0 −1 1 1 1 −1 −1 −1 −1 −1 −1 −1 1 −1 1 −1 0 1 −1 1 1 −1 −1 1

[a] 2-BOK uses Code 1; 4-BOK uses codes 1 and 2; 8-BOK uses codes 1–4.

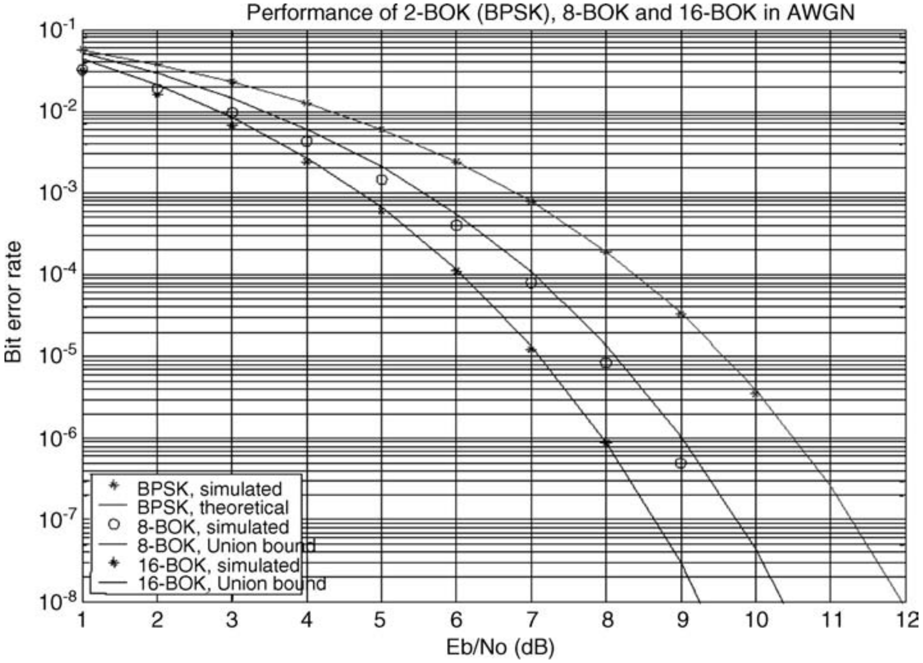


FIGURE 10.24 Theoretical and simulated performance of the M-BOK codes in AWGN [Wel03].

performance of 2-, 8- and 16-BOK codes. At low error rate, the 8-BOK and 16-BOK need approximately 1.0 dB and 2.5 dB respectively more signal power to perform as well as a 2-BOK. With this additional power they can increase the data rate by three and four times respectively.

Variable data rates in DS-UWB are achieved by changing the processing gain, and switching between BPSK and QPSK modulation. Lower data rates with higher processing gains can cover larger areas. The chip rates in the low and high bands are 1.368 Gc/s and 2.736 Gc/s respectively. The basic symbol transmission rate for the low band with spreading factors of 24 and 32 are 1.368 (Gc/s)/24 (chips) = 57 and 1.368 (Gc/s)/32 (chips) = 42.75 MS/s. Similarly, high-band operation supports two basic data rates: 114 and 85.5 MS/s. Different data rates are derived from these basic rates and three different coding schemes. The first coding option is a convolutional code with coding rate $R = 0.50$. The second coding option is a (55, 63) RS code with rate $R = 0.87$. The third option is a concatenated code that has both rate-$^1\!/_2$ convolutional and a (55, 63) RS code, resulting in an overall code rate $R = 0.44$.

Example 10.22: Data Rate Calculation in DS-UWB For the 25 Mb/s data rate in Table 10.3 we have two codes representing bits 0 and 1 and a BPSK modulation (no quadrature phase). Therefore, each point in the constellation is identified by one bit and the uncoded transmission data rate is 57 Mb/s, the same as the symbol transmission rate. The coded data rate is 57 Mb/s × 0.44 = 25 Mb/s. For the 200 Mb/s rate, we have four symbols, resulting in 2 bits per symbol and a QPSK modulation that has another 2 bits per symbol. Therefore, in the overall scheme we have 4 bits per symbol. With a symbol rate of 114 MS/s

TABLE 10.3 Different Data Rates Supported by DS-UWB [Koh04]

Data rate (Mb/s)	Constellation	Symbol rate	Quadrature	FEC rate
25	2-BOK	57	No	0.44
50	2-BOK	114	No	0.44
114	4-BOK	114	No	0.50
112	8-BOK	85.5	No	0.44
200	4-BOK	114	Yes	0.44
224	64-BOK	85.5	No	0.44
450	64-BOK	85.5	Yes	0.44
900	64-BOK	85.5	Yes	0.87

and a coding rate of $R = 0.44$ we have

$$114\,\mathrm{MS/s} \times 4\,\mathrm{bits/S} \times 0.44 = 200.64\,\mathrm{Mb/s}$$

As we noted earlier, the MAC combines FDM, TDM, and CDM techniques. The FDM scheme chooses one of the two operating frequency bands to control the interference. The CDM uses ternary code sets ($\pm 1, 0$) with 2-, 4-, and 8-BOK with length 24 and 64-BOK with length 32. Four CDMA codes within each frequency band are used to separate the piconets further by providing for implementation of logical channels. Within each piconet, TDM separates different users. The following example is provided to aid the understanding of this complex medium-access method.

Example 10.23: MAC in DS-UWB Figure 10.25 illustrates an example using FDM, CDM, and TDM access in the DS-UWB scheme in a multiroom indoor area. Each set of piconets is separated by FDM between the low band (LB) and high band (HB). Within each set using the same band, the piconets are separated by CDM codes into channels A, B, and C. Then, within each cell, users are separated by TDMA using a central scheduling system similar to those used in other TDMA systems, such as GSM and HIPERLAN/2.

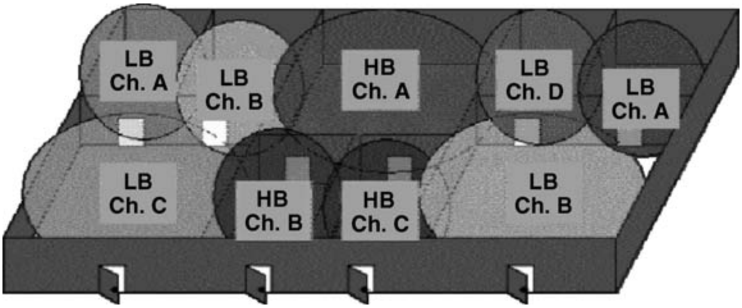


FIGURE 10.25 Distribution of piconets in FDM/CDM/TDM access used in DS-UWB proposal for IEEE 802.15.3a [©IEEE].

10.4.2 Multiband Orthogonal Frequency-Division Multiplexing

At the time of writing, the second and leading proposal for IEEE 802.15.3a is MB-OFDM, the technology developed by the Multiband OFDM Alliance (MBOA). The MB-OFDM system [Sho03] uses the OFDM technique, which emerged as the technology of choice for IEEE 802.11 WLAN standards operating in the U-NII 2.4 and 5 GHz unlicensed bands, and in the UWB 3.1–10.6 GHz unlicensed bands. Following this approach, the spectrum is divided into 15 bands each of width 528 MHz. In each band, a 128-point OFDM system using QPSK modulation is implemented to limit the required precision of mathematical operations and make digital implementation at ultrahigh sampling rates feasible. The MAC is time–frequency multiple access (TFMA), which combines the time- and frequency-diversity benefits of FHSS and DSSS into one MAC technique.

Example 10.24: TFMA in MB-OFDM Figure 10.26 [Sho03] illustrates the basic operation of TFMA for a spread-spectrum code of length 7. The chip duration is 3.79 ns, resulting in a code duration of $3.79 \times 7 = 26.53$ ns. Similar to DSSS, each symbol is divided into 7-chip periods to provide time diversity. Similar to FHSS, each chip is transmitted on a different frequency channel to provide frequency diversity.

Figure 10.27 gives an overview of the MB-OFDM proposal. The 15 bands in the 3.1–10.6 GHz unlicensed UWB spectrum are divided into five groups of 528 MHz bands. Group 1 is the most desirable, because Group 2 interferes with U-NII bands and IEEE 802.11a devices and higher groups have smaller coverage areas. Each physical piconet is implemented in a band group and several logical piconets share a band group using different TFMA codes. Groups 1–4 have 4-TF (time–frequency) codes and Group 5 has 2-TF codes for logical channel separation. In this manner, this proposal can accommodate 18 piconets in the entire UWB spectrum, and four of these piconets implemented in band Group 1 are the most popular of all.

Table 10.4 provides the four patterns of TF codes used in Groups 1, 2, 3, and 4 and the two patterns for Group 5. Groups 1–4 have three different frequencies and the length of the time sequence is six, in which each band is used twice per symbol. Group 5 has two bands, a code

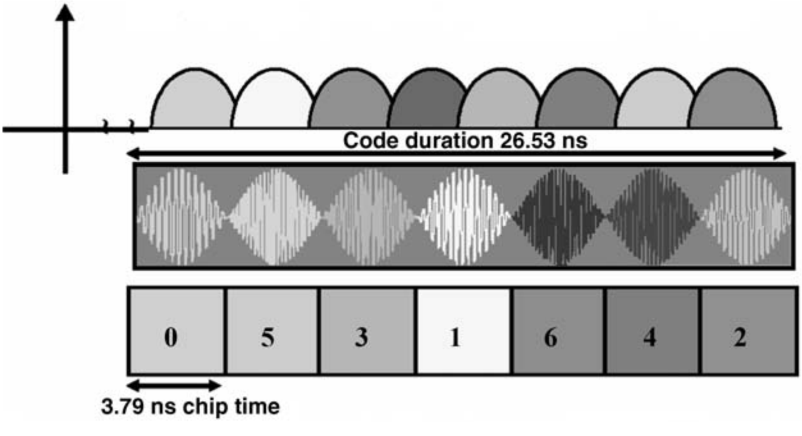


FIGURE 10.26 Explanation of the TFMA operation using a length 7 TF code. Each chip of a spread-spectrum code is transmitted in a different frequency channel [Sho03].

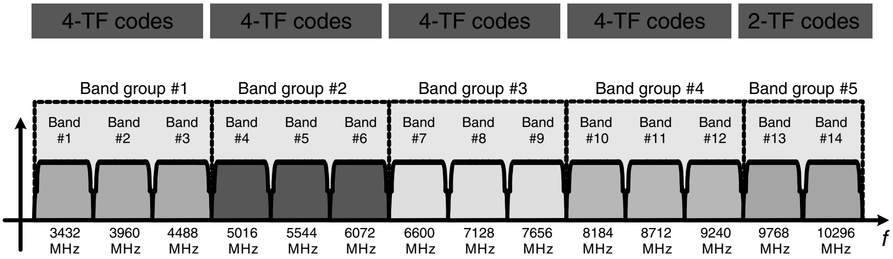


FIGURE 10.27 Frequency bands, groups of frequencies, and TFMA codes within each group of the MB-OFDM approach to UWB communications in 3.1–10.6 GHz unlicensed UWB bands proposed to the IEEE 802.15.3a WPAN standard.

length of 4, and each band is used twice during each symbol transmission. With this technique, as shown in Table 10.4, neighboring piconets have two collisions per code length.

Multirate data communication in this system is supported by adjusting the spread of the pulses in time. The same time sequence is spread by a factor of two or four to increase the symbol transmission rate by two or four. The following example illustrates this technique.

Example 10.25: Variable Data Rate in TFMA Figure 10.28 shows the TFMA system with seven bands, discussed in Example 10.8. The top of the figure is the same as in Fig. 10.25 with a full pulse-rate frame (PRF), in which pulses are transmitted in sequence one after another. The lower part of the figure shows a transmission with a half PRF, where pulses are transmitted every other chip interval. The data rate of the scheme shown in the lower part of the figure is half of the data rate of the scheme shown in the upper part of the figure.

Figure 10.29 shows the general block diagram of the OFDM system proposed by the MBOA. This system uses a 242.4 ns information length with a 60.6 ns prefix for multipath protection, and a 9.5 ns guard interval to provide time for switching between bands, for a total symbol duration of 310.5 ns. From the 128 tones or carriers, 100 are data tones used to transmit information, 12 are pilot tones used for carrier and phase tracking, 10 are guard tones (previously called dummy tones), and 6 are NULL tones. Table 10.5 provides the specification of all mandatory and optional data rates supported by this system. A simple example illustrates how this table can be read.

TABLE 10.4 TF Codes Recommended for the MC-OFDM System by the MBOA [Wel03]

Band groups	Preamble pattern	TF code length	TF code					
1, 2, 3, 4	1	6	1	2	3	1	2	3
	2	6	1	3	2	1	3	2
	3	6	1	1	2	2	3	3
	4	6	1	1	3	3	2	2
5	1	4	1	2	1	2	–	–
	2	4	1	1	2	2	–	–

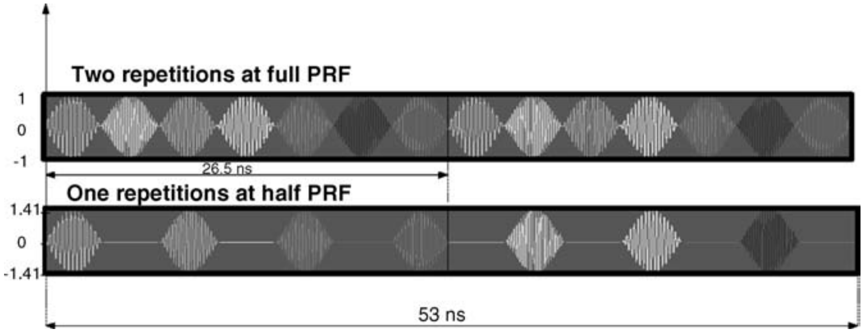


FIGURE 10.28 Control of data rate using spreading time. The data rate of the scheme shown in the lower part of the figure is half of the data rate of the scheme shown in the upper part of the figure.

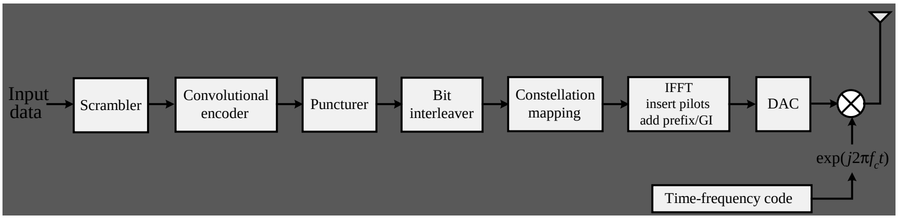


FIGURE 10.29 General block diagram of the OFDM modulation system recommended for MC-OFDM proposal [Wel03].

TABLE 10.5 Specification of Different TF Codes Recommended for the MC-OFDM System by the MBOA [Wel03]

Info. data rate	55 Mb/s[a]	80 Mb/s[b]	110 Mb/s[a]	160 Mb/s[b]	200 Mb/s[a]	320 Mb/s[b]	480 Mb/s[b]
Modulation/ constellation	OFDM/ QPSK	OFDM/ QPSK	OFDM/ QPSK	OFDM/ QPSK	OFDM/ QPSK	OFDM/ QPSK	OFDM/ QPSK
FFT size	128	128	128	128	128	128	128
Coding rate R (K = 7)	11/32	1/2	11/32	1/2	5/8	1/2	3/4
Spreading rate	4	4	2	2	2	1	1
Data tones	100	100	100	100	100	100	100
Info. length (ns)	242.4	242.4	242.4	242.4	242.4	242.4	242.4
Cyclic prefix (ns)	60.6	60.6	60.6	60.6	60.6	60.6	60.6
Guard interval (ns)	9.5	9.5	9.5	9.5	9.5	9.5	9.5
Symbol length (ns)	312.5	312.5	312.5	312.5	312.5	312.5	312.5
Channel bit rate (Mb/s)	640	640	640	640	640	640	640

[a] Mandatory.
[b] Optional.

TABLE 10.6 Simulated Performance of the MB-OFDM System in Four Different Band Groups. The Coverage is Compared with the Coverage in an AWGN [Wel03]

	Range (m)				
	AWGN	CM1	CM2	CM3	CM4
110 Mb/s	20.5	11.4	10.7	11.5	10.9
200 Mb/s	14.1	6.9	6.3	6.8	4.7
480 Mb/s	7.8	2.9	2.6	N/A	N/A

Example 10.26: Variable Data Rate in TFMA Since the symbol duration is 310.5 ns and modulation is QPSK with 2-bits/symbol, the basic transmission rate all data rates is

$$1/310.5(\text{ns/S}) \times 2(\text{bits/S}) \times 100(\text{carriers}) = 640\,\text{Mb/s}$$

The lowest data rate in the table is 55 Mb/s, which is obtained by a coding rate of 11/32 with a spreading rate (repetition for variable data rate support) of 4. Therefore, the data rate is

$$640\,\text{Mb/s} \times 11/32\ (\text{bits/coded bits})/4(\text{spreading rate}) = 55\,\text{Mb/s}$$

Table 10.6 [Wel03] shows the range of coverage for the 100, 200, and 480 Mb/s systems in additive white Gaussian noise (AWGN) and the first four groups of bands for the MB-OFDM proposal. The 480 Mb/s option in band Groups 3 and 4 has limited coverage, and in the first two bands has a range of less than 3 m. This distance is suitable for WPAN applications, such as connecting a device to a USB port of a computer or a laptop. The 110 Mb/s option covers about 10 m, which is the desirable coverage range for IEEE 802.15 devices. The 200 Mb/s option covers more than 6 m in the first three groups and less than 5 m in the fourth group. The performance criterion for this simulation was packet error rate of 8% in 90% of locations. Simulations include several practical factors, such as losses due to front-end filtering and conversions, multipath degradation, channel estimation errors, carrier tracking, and packet acquisition. More detailed analysis of the performance, the effects of interference with IEEE 802.11 and other devices, and specification for implementation of MB-OFDM systems can be found in [MBO04].

10.5 IEEE 802.15.4 ZIGBEE

The IEEE 802.15.4 standard specifies PHY and MAC layers for low-rate, low-power, low-cost WPANs. The first specification was made in 2003 and the latest revision was done in 2006. Compared with IEEE 802.11 WiFi devices, ZigBee is designed for very low cost communications among scattered devices with minimal infrastructure. Many applications and the coverage for ZigBee are similar to those of the Bluetooth. However, ZigBee intends to provide faster formation of the piconet, a larger number of active users, longer battery life, and lower data rates of 20–250 kb/s. ZigBee is a specification for high-level communication protocols capable of mesh networking using the lower layer IEEE 802.15.4 standard. The first ZigBee specification was ratified in 2004 and the latest in 2007. The relation between IEEE 802.4 and ZigBee is similar to the relation between IEEE 802.11 and WiFi. So much as the IEEE 802.1 Bluetooth was a low-complexity and low-power ad hoc network design affected by the IEEE 802.11 FHSS standard, the design of the IEEE 802.15.4 is affected by the IEEE 802.11 DSSS/CCK standards.

10.5.1 Overall Architecture

One of the differences of the IEEE 802.15.4 standard is that it defines two types of node in the network. The two types of node are referred to as *full-function devices* (FFDs) and *reduced-function devices* (RFDs). An FFD is similar to a Bluetooth device and it can serve as the coordinator or master of a piconet or as a common node or a full-function slave. An FFD can communicate with any other device and it can help routing messages throughout the network. The RFD nodes are defined to be extremely simple with very modest resource and communication capabilities, only used as a slave node to communicate with an FFD. This provides flexibility for implementation of a variety of topologies addressing more diversified applications. A typical example application would be a wireless light lamp switch. The node at the lamp can be an FFD, since it is connected to the mains supply, while the battery-powered light switch would be an RFD.

In a manner similar to Bluetooth and IEEE 802.11, IEEE 802.15.4 and ZigBee support both peer-to-peer and star network topologies. A new topology for the IEEE 802.4 ZigBee is the cluster tree topology. Figure 10.30 shows all three topologies for IEEE 802.15.4 ZigBee networks. The peer-to-peer networks, shown in Fig. 10.30*a*, form arbitrary patterns of connections. How far the extension of these connections is possible depends on the distance between each pair of nodes. These nodes are all FFDs and meant to serve as the basis for on-the-fly networks that are capable of performing self-management and organization. Figure 10.30*b* shows a simple star topology with master–slave operation that is similar to that of Bluetooth. In the ZigBee network, however, we have two classes of nodes which provide more flexibility in forming a network to support an application. The network coordinator must be an FFD and other nodes can be either an RFD or an FFD, depending on the application. Figure 10.30*c* shows a more complex clustered tree topology with three star

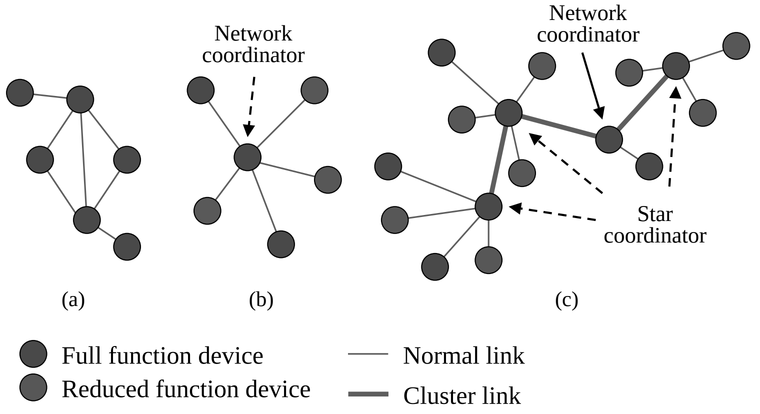


FIGURE 10.30 IEEE 802.4 ZigBee topologies: (*a*) peer-to-peer; (*b*) star for master–slave operation; (*c*) clustered stars.

coordinators and one network coordinator connecting stars together. The coordinators in this figure play the role of a router in the ad hoc network, which means the network coordinator manages the entire network. This type of topology is unique to ZigBee and it allows formation of a hierarchical ad hoc network with simple end nodes which are in sleeping mode most of the time, allowing extremely long battery lives.

In a cluster tree topology there are three distinct classes of node operation. The first class is the dumb end-node devices, which are sleeping most of the time to save on battery consumption. Each end device allows up to 240 end-point separate applications sharing the same radio. For example, a three-gang light switch would have three distinct end points sharing the same radio electronics and battery. The second class of devices is the mid-level routers, which have the ability to stack up communication messages and respond to general enquiries about sleeping end devices in their vicinity. These routers are also responsible to find out the best way to pass on a message to a node that is not in the range. The third class of nodes at the top of the hierarchy is the network coordinator, which is always on and relies on connection to a good power source. In addition to being a router the network coordinator sets the rules for basic network operation, such as finding an appropriate frequency channel for the network.

10.5.2 Protocol Stack

Though WPANs are simple ad hoc sensor networks designed to operate over short distances of up to 10 m, the network has several layers designed to enable communication within the network, connection to a network of higher level, and ultimately an uplink to the Web. Like all other communications networks, IEEE 802.15.4 ZigBee divides up the communications tasks into layers on a stack of protocols, as shown in Fig. 10.31. The lower PHY and MAC layers are defined by IEEE 802.4 and the higher layer by the ZigBee Alliance. In addition, application developers can define their own application layers. The top of the stack is the application and the bottom the physical radio. The middle layers are used to glue the application to the actual transmission so that nodes can communicate reliably, efficiently, and securely and so designers can develop their application more easily.

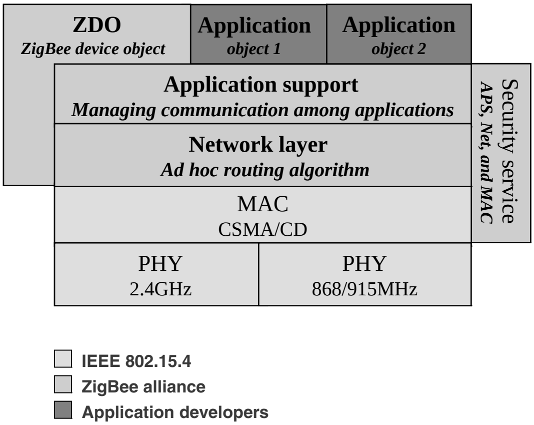


FIGURE 10.31 The IEEE 802.15.4 ZigBee protocol stack.

A *ZigBee device object* (ZDO) is a special application object that is resident on all ZigBee nodes. The address of the ZDO application is always end point zero, and other *application object* end points running on the ZigBee module are numbered 1 to 240. This node has its own profile, which other user application end points and other ZigBee nodes can access. This program is responsible for overall device management and security keys and policies. Each profile in a ZigBee module has the table of other ZigBee modules on the network and the services they offer. All other applications' end points use this information to discover other devices, to manage binding, and to specify security and network settings. The *application support layer* routes messages on the network to different application end points running on a ZigBee module by maintaining a binding table and forwarding messages to the appropriate application. The *network layer* provides the routing and multi-hop capability required to turn MAC-level communications into a full star, tree, or mesh network. ZigBee employs a distance–vector routing algorithm suitable for ad hoc networks called the ad hoc on-demand distance vector (AODV) protocol. This AODV algorithm automatically constructs a low-speed ad hoc network of nodes by forwarding messages, discovering neighboring devices, and building up a map of the routes to other nodes. In the coordinator nodes, the network layer assigns network addresses to new devices when they join the network for the first time. The *security service* provides for establishing and exchanging security keys, and using these keys to secure the communications. The security services work across three layers to provide security at each level. Like all other security services in wireless networks, this layer is responsible for encrypting the data when it is generated and authenticating it when it is received. The ZDO layer dictates the security policies and configurations implemented by the security services.

To follow the same format of presentation that we used to present Bluetooth details, here we treat MAC and PHY layer in separate sections.

10.5.3 Medium Access Control Layer

The MAC layer provides reliable communications between a node and its immediate neighbors. One of its main tasks, particularly on a shared frequency channel, is to listen for when the channel is clear before transmitting. This is known as CSMA-CA communication. In addition, MAC can provide beacons and synchronization to improve communications efficiency. The MAC layer also manages packing data into frames prior to transmission, and then unpacking received packets and checking them for errors.

The basic channel access mechanism for IEEE 802.15.4 is CSMA/CA (see Chapter 13 for more details on the specifics of channel access in IEEE 802.15.4). In addition, the MAC layer provides beacons and synchronization to improve communications efficiency. Beacons do not follow the carrier sensing and they are sent on a fixed timing schedule. Networks are formed either based on beacon or without a beacon. Acknowledgement packets are also sent without carrier sensing after packet arrival. The MAC layer also manages packing data into frames prior to transmission, and then unpacking received packets and checking them for errors. Two other important features of 802.15.4 MAC are (a) the allowance for guaranteed time slots and (b) integrated MAC support for secure communications. Devices that have low latency real-time requirements can use the so-called guaranteed time slots allocated to them by a coordinator without resorting to carrier sensing. As we mentioned in the previous section, security is implemented across several layers, including the MAC. Again, more details are available in Chapter 13.

In general, ZigBee protocols minimize the time the radio is on to reduce unnecessary power consumption. In non-beacon-enabled networks, power consumption is decidedly asymmetrical: some devices are always active, while others spend most of their time sleeping. In these networks, an unslotted CSMA/CA channel access mechanism is used and routers typically have their receivers continuously active, requiring a more robust power supply. These networks allow a bipolar consumption of energy. The router consumes substantially, while end nodes only transmit when an external stimulus is detected. The typical example for such an operation is the lamp operation which we discussed earlier. In networks using a beacon, nodes are only activated after a beacon is transmitted. The special ZigBee router node transmits a beacon periodically to confirm its presence to other users. Other nodes may sleep between beacons to save in energy consumption. For different options of the PHY the beacon interval ranges from 15 to 24 ms. This option is more suited for higher traffic loads and it is very similar to the IEEE 802.11 operation. In general, the CSMA/CA in IEEE 802.15.4 is a simpler version of the CSMA/CA used by the IEEE 802.11, which includes different options for conservative power consumption. The PCF option, variety of interframe delays (PIFS, DIFS, and SIFS), and the RTS and CTS mechanisms which were included in the IEEE 802.11 MAC are not included in 802.15.4. What RTS/CTS and PCF mechanisms were providing here is provided by simpler and more practical guaranteed time-slot transmission.

10.5.4 Physical Layer

Similar to IEEE 802.11, IEEE 802.15.4 operates in the unlicensed ISM radio bands. In addition to the 2.4 GHz band used in IEEE 802.11, which is available worldwide, IEEE 802.15.4 also has options for unlicensed 868 MHz in Europe and unlicensed 915 MHz in countries such as the USA and Australia. Figure 10.32 shows the 26 different channel numbers specified by the standard in each band. The basic radios use DSSS with different chip rates and modulation techniques. BPSK modulation and 15-chip LFSR codes are used

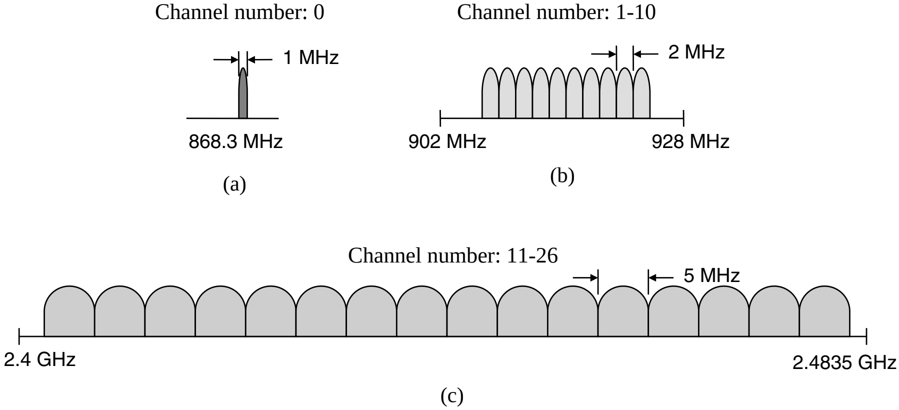


FIGURE 10.32 Frequency bands used by different PHY options in IEEE 802.15.4 ZigBee: (*a*) one channel at 868 MHz; (*b*) 10 channels in 915 MHz; (*c*) 16 channels in 2.4 GHz.

for the 868 MHz and 915 MHz bands with chip rates of 0.3 Mc/s and 0.6 Mc/s respectively. This results in

$$0.3\ (\mathrm{Mc/s})/15\ (\mathrm{c/bit}) = 20\ (\mathrm{kb/s})$$

for the single channel in 868 MHz and

$$0.6\ (\mathrm{Mc/s})/15\ (\mathrm{c/bit}) = 40\ (\mathrm{kb/s})$$

for each of the 10 channels at 9.15 MHz.

For 2.4 GHz operation, 16-ary orthogonal coding with code words of length 32 chips is used. The modulation technique used for transmission of data is the offset QPSK. This general structure is similar to the reverse channel of the IS-95 CDMA system, also using *M*-ary orthogonal coding and offset QPSK modulation. The offset QPSK provides a constant envelope which reduces the power consumption of the last-stage amplifiers in the radio. Modulation is also implemented noncoherently, which is simpler and avoids power consumption and the complexity of the phase-lock loops needed for coherent modulations. These measures are taken to satisfy low-cost and low-power implementation of the radio. To map the data bits to the transmitted symbols, the arriving raw data stream at rate of 250 kb/s is used to create 4-bit blocks of data at a rate of

$$250\ (\mathrm{kb/s})/4\ (\mathrm{bits/S}) = 62.5\ (\mathrm{kS/s})$$

Each 4-bit block is then used to address one of the 16 orthogonal symbols selected from a 32-bit length block of chips. The chips are then transmitted at a rate of 2 Mc/s using offset QPSK modulation. Therefore, the net processing gain of this DSSS used in the IEEE 802.15.4 is

$$2\ (\mathrm{Mb/s})/250\ (\mathrm{kb/s}) = 16$$

It is useful to remind the reader that the processing gain of the IEEE 802.11 DSSS was 11, which is reasonably close to what is used in IEEE 802.15.4. It is also useful for the reader to remember that the CCK coding used in IEEE 802.11b was also using a different sort of *M*-ary orthogonal coding. These observations show the path of evolution of experimentally successful local radio networks. The maximum transmission power of IEEE 802.15.4 radios is usually 0 dBm (1 mW), compared with 20 dBm (100 mW) used in IEEE 802.11. This 20 dB difference, assuming free-space propagation with a distance–power gradient of 2, accounts for a one order of magnitude higher coverage for IEEE 802.11 devices. Although, as we explained in Chapter 2, the coverage depends on the environment, it is customary to assume that the coverage of WPANs is roughly 10 m and that of WLANs is approximately 100 m. The lower transmission power and power-conscious design of the radio is one of the major difference between the design of IEEE 802.15.4 and IEEE 802.11. The sensitivities of the radio for 2.4 GHz and 868/915 MHz are specified by the standard as −85 dBm and −92 dBm respectively. This allows 7 dB more for operation at 868/915 MHz. In addition, using Eq. (2.8), the path loss in the first meter for the 2.4 GHz and 915 MHz we have another

$$20\log\left(\frac{2.4\ (\mathrm{GHz})}{915\ (\mathrm{MHz})}\right) = 8.3\ (\mathrm{dB})$$

difference between the two bands that adds up to a 15.3 dB edge for operation at lower frequencies. In free space this accounts for approximately up to $10^{15.3/20} < 6$ times longer coverage for operation in lower frequency bands. In indoor environments, lower frequencies penetrate better through the walls, which increases the coverage at lower frequencies to even higher values.

10.5.5 Frame Format

IEEE 802.15.4 uses four different types of frame, for data, acknowledgment, beacon, and MAC commands. Figure 10.33 shows the frame format for all four types of packet. Every frame has a 4-byte preamble, a 1-byte start of the packet delimiter, and a 1-byte start of the frame which uses 7 bits to identify the length of the packet in bytes and 1 bit to identify the addressing scheme because the device address is either 2 bytes (16 bits) or 8 bytes (64 bits). In addition to payload, the physical service data unit (PSDU) for all packets has a 2-byte frame control to carry control messages, a 1-byte sequence number to provide for tracking the sequence of the packets, up to 20 bytes of addressing, and 2 bytes of CRC-16 as a frame check sequence. The acknowledgement packet does not have any address field and the other packets have source and destination addresses with 2 bytes for coordinator node of the PAN address plus 2 bytes or 8 bytes for the device address. The payload for the data frame is the information with a length that insures that the overall length of the PSDU stays under 127 bytes. The payload for the beacon provides other devices with the necessary bits for synchronization and configuration. The MAC command control field carries different MAC control messages.

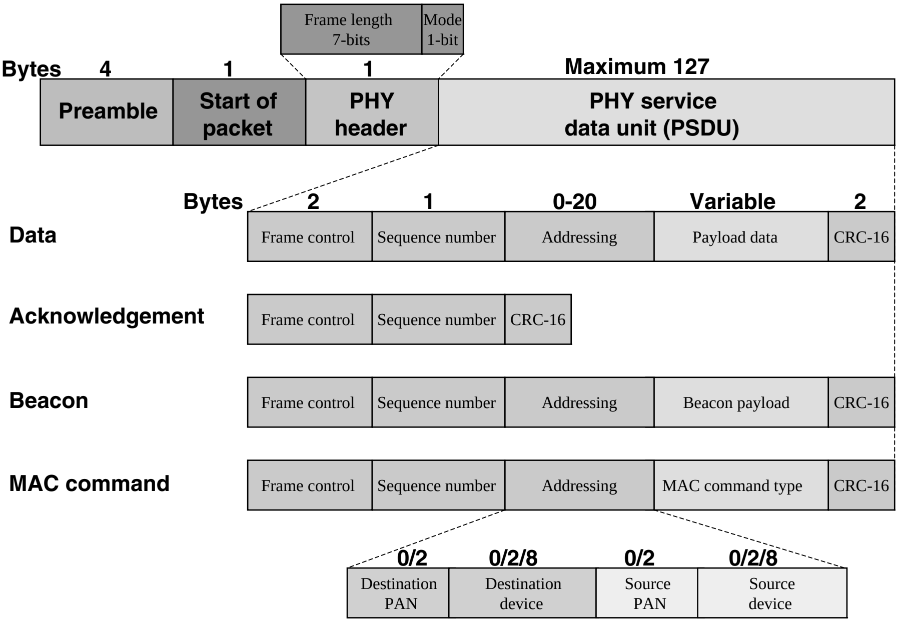


FIGURE 10.33 Frame format for IEEE 802.15.4 packets.

TABLE 10.7 Comparison of 802.11, 802.15.1, and 802.15.4

Technology/feature	802.11	802.15.1	802.15.4
Frequency	2.4 GHz	2.4 GHz	2.4 GHz–868/915 MHz
Modulation	OFDM	FHSS/BPSK	DSSS/QPSK
MAC	CSMA/CA	TDMA/TDD	CSMA/CA
Max. data rate	54 Mb/s	1 Mb/s	20/40/250 kb/s
Device types	One	One	FFD/RFD
No. of channels	11–14	79	26
No. of devices	32	8	Up to 65 535
Battery life	Days	Weeks	Months
Coverage (m)	100	10	10/50
Topology	Star	Star	Star and cluster-tree

10.5.6 Comparison of ZigBee with Bluetooth and WiFi

Table 10.7 provides a summary comparison among WiFi (802.11g), Bluetooth, and ZigBee technologies. All devices use 2.4 GHz unlicensed ISM bands, but ZigBee also supports 868/915 MHz, which has a better coverage for the same level of power. The modulation techniques for Bluetooth and ZigBee both use spread-spectrum technology, which is a power-efficient transmission technique. Both devices also implement noncoherent modulations to save in the complexity and electronic design power consumption. The MACs of WiFi and ZigBee are both CSMA/CD, which is a distributed MAC protocol suitable for data applications. The ZigBee implementation is a light version of WiFi, because it has fewer features. Each piconet of Bluetooth uses the centralized TDMA/TDD technique, which is more appropriate for telephone applications. Different piconets are separated using the FHSS CDMA technique. The data rate of ZigBee is in the range of the data rates for Bluetooth which are suitable for ad hoc sensor network applications. WLANs generally provide higher data rates (54 Mb/s for 802.11g), which are more suited to modern computer networking applications for home and small office networking and pervasive access in public buildings.

ZigBee allows two different types of terminal which allow the design of simple end nodes that may sleep most of the time to stretch the battery lifetime. Although the number of channels for WiFi and ZigBee in 2.4 GHz look similar, the ZigBee channels are narrower (5 MHz), allowing implementation of 16 nonoverlapping channels. The IEEE 802.11 DSSS has only three nonoverlapping channels. Bluetooth allows up to 79 piconets, which is far beyond the other two, but the number of nodes per terminal has a more practically important role because, usually, we have a few piconets but a large number of nodes. In ad hoc sensor networking, having only seven simultaneously connected nodes is rather limited, and so ZigBee's higher number of nodes is very desirable. All the features of ZigBee are designed to minimize power consumption, which has resulted in a longer battery lifetime. Coverage of WiFi is better, and that is important for computer networking; Bluetooth and ZigBee are designed for ad hoc sensor networking and the coverage can be extended by the selected topology. The ability of ZigBee to form cluster-tree networks increases the coverage of the ad hoc network significantly, making it a better choice for applications where we have a number of sensor networks spread over a large geographical area.

As a result of these differences, WiFi and Bluetooth have found their specific applications. WiFi is dominating the WLAN market for the home, small office, and ad hoc public building access. Bluetooth has found a number of ad hoc wire replacements for telephone connections inside cars and for audio devices. It is expected that ZigBee will develop a similar market for sensor networks for medical and other data applications.

QUESTIONS

1. What is IEEE 802.15 and what is its relation to Bluetooth and ZigBee?
2. What are the differences between IEEE 802.15 device specifications and the device specifications of the IEEE 802.11 devices?
3. Name the four states that a Bluetooth terminal can take and explain the differences among these states.
4. Name the three classes of applications that are considered for Bluetooth technology and identify those which can also be considered for 802.11 WLAN technologies.
5. What are the similarities and differences between FHSS used in the IEEE 802.11 and Bluetooth in terms of data rate, modulation technique, available frequencies for hopping, speed of the hop, and the number and pattern of the hops?
6. What are the differences between ad-hoc solutions offered by 802.11 and Bluetooth?
7. What is the difference between the MAC protocol of Bluetooth and IEEE 802.11 FHSS?
8. Which IEEE 802.11 standards interfere with Bluetooth and which of these standards have more serious interference conditions with it?
9. How many different voice services does Bluetooth support and how are they differentiated from one another?
10. How many different symmetric and asymmetric data services does Bluetooth support?
11. What is the maximum supported asymmetric packet data rate by Bluetooth? How many slots per hop does it use? What is its associated data rate in the reverse channel?
12. Compare the header and access code of Bluetooth with the PLCP header of the FHSS IEEE 802.11.
13. What is the maximum data rate of an overlay Bluetooth network? How does it compare with the maximum data rate of the overlay FHSS IEEE 802.11?
14. What are the differences between the implementation of paging and inquiry algorithms in Bluetooth?
15. Which is the IEEE standardization group involved in UWB communications and what other popular wireless standard has been developed by that group?
16. Compare Bluetooth and UWB as two solutions for WPAN applications.
17. Name two popular antenna systems used for UWB communications and compare them for practical applications in personal communications.
18. Which is the ISM band that overlaps with the UWB bands and which is the IEEE WLAN standard that uses this band?
19. What are the data rates, frequency bands and MAC/PHY layers of the DS-UWB proposal for 802.15.3a?
20. What are the data rates, frequency bands, and MAC/PHY layers of the MB-OFDM proposal for 802.15.3a?

21. What is the difference between DSSS and UWB time-hopping?
22. What is the difference between the bandwidth and signal spreading techniques of DSSS and UWB time-hopping techniques?
23. Give the frequency band, modulation technique, chip rate, code length, and number of code words in the constellation for the 900Mbps DS-UWB proposal for 802.15.3a.
24. Why in the DS-UWB system is the band between 4.9–6.2GHz not used?
25. Give the range of the frequency bands, number of channels, and the range of data rates that are supported by the IEEE 802.15.3a DS-UWB proposal.
26. Give the range of the frequency bands, number of channels, and the range of data rates that are supported by IEEE 802.15.3a MB-OFDM proposal.
27. What is the MB-OFDM proposal for the IEEE 802.15.3a, how does it relate to UWB, and what is its medium access control scheme?

PROBLEMS

Problem 1:

Give the complete protocol stack for the implementation of email application over Bluetooth.

Problem 2:

Consider that the encoded voice in Bluetooth is at 64Kbps in each direction.

(a) Using the packet format for the HV1 channels show that these packets are sent every six slots.

(b) Using the packet format for the HV2 channels find how often these packets are sent?

(c) Repeat (b) for HV3 packets.

Problem 3:

(a) What is the hopping rate in Bluetooth and how many bits are transmitted in each one slot packet transmission?

(b) If each frame of the HV3 voice packets in Bluetooth carries 80 bits of the samples of speech. what is the efficiency of the packet transmission (ratio of the overhead to overall packet length)?

(c) Determine how often HV3 packet have to be sent to support 64Kbps in each direction.

(d) The DH5 packets carry 2712 bits per five slots packet. Determine its effective data rate in each direction.

Problem 4:

Repeat examples 10.7 and 10.8 for all other data rates supported by Bluetooth, see Table 10.1.

Problem 5:

Consider the Bluetooth and FHSS IEEE 802.11 interference scenario of Fig. 10.15.

(a) If the acceptable error rate for the MT is 10^{-5}, determine the $S_{\min}$ that supports this error rate. Assume that the probability of error as a function of γ_b (say the same as $S_{\min}$) is given by this approximation: $P_e \approx 0.5e^{-\gamma_b/2}$.

(b) Using $S_{\min}$ of part (a) and Eq. 10.2 calculate r_{in} for $d = 10m$, $\alpha = 2$, $P_{BT} = 0$ dBm and $P_{AP} = 20$ dBm.

(c) Produce a computer plot to illustrate the relation between r_{in} and acceptable error rates between 10^{-2} and 10^{-7} (in logarithmic form). Using the computer plot discuss the impact of error rate requirement on the range of interference between Bluetooth and FHSS IEEE 802.11. Assume the rest of parameters are the same as part (b).

(d) Produce a computer plot to illustrate the relation between r_{in} and distance-power gradient of the medium for values of α between 1.5 and 6. Using the computer plot, discuss the impact of medium on the range of interference between Bluetooth and FHSS IEEE 802.11. Assume the rest of parameters are the same as part (b).

(e) Repeat (c) and (d) for $P_{BT} = 10$ dBm. Compare the results with associated results in the previous parts and discuss the effects of power level in the interference.

Problem 6:

Consider the FHSS IEEE 802.11 and Bluetooth interference scenario of Fig. 10.16.

(a) Assuming that the acceptable error rate for the Bluetooth is 10^{-5} determine the $S_{\min}$ that supports this error rate. Hint: use equations from Chapter 3 for calculation of γ_b (the same as $S_{\min}$).

(b) Using $S_{\min}$ of part (a) and Eq. 10.3 calculate r_{in} for $d = 10m$, $\alpha = 2$, $P_{BT} = 0$ dBm and $P_{AP} = 20$ dBm.

(c) Produce a computer plot to illustrate the relation between r_{in} and acceptable error rates between 10^{-2} and 10^{-7}(in logarithmic form). Using the computer plot discuss the impact of error rate requirement on the range of interference between FHSS IEEE 802.11 and Bluetooth. Assume the rest of parameters are the same as part (b).

(d) Produce a computer plot to illustrate the relation between r_{in} and distance-power gradient of the medium for values of α between 1.5 and 6. Using the computer plot discuss the impact of medium on the range of interference between FHSS IEEE 802.11 and Bluetooth. Assume the rest of parameters are the same as part (b).

Problem 7:

Repeat problem 5 if the FHSS IEEE 802.11 device is replaced by a DSSS IEEE 802.11 device.

Problem 8:

A FHSS IEEE 802.11 and a Bluetooth device are operating in close vicinity of each other. Generate a computer plot illustrating the probability of collision of their packets versus the size of the FHSS packet. Using the results of computer plots explain the impact of packet length on the probability of collision between FHSS IEEE 802.11 and Bluetooth. Note that the maximum length of the 802.11 packets is specified by the standard.

Problem 9:

Repeat Problem 8 for interference analysis between the DSSS IEEE 802.11 and Bluetooth.

Problem 10:

Repeat Problem 8 for interference analysis between the CCK IEEE 802.11b and Bluetooth.

Problem 11:

Assume an UWB device with a bandwidth of 1 GHz and power of 0.1 mW using DSSS transmission with precessing gain of 128 operates in a band that includes the entire IEEE 802.11a bands. If both devices attempt to communicate with an access point supporting both technologies and they are located at the same close distance from the access point for which the distance-power gradient is two,

(a) What is the signal to interference ratio for the received IEEE 802.11a signal at the access point?

(b) What is the signal to interference ratio for the received UWB signal at the access point?

Problem 12:

(a) Give all 8-BOK codes of length 24. Show that the first code is orthogonal to the second and third code.

(b) Calculate the two basic data rates of the DS-UWB for 24 and 32 spreading factors in the high-band.

Problem 13:

(a) Show that all TF hopping patterns for the MC-OFDM, shown in Table 10.4, have 2 collisions per code.

(b) Repeat example 10.25 for all data rates given in Table 10.5.

PART FOUR

SYSTEM ASPECTS

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

11

NETWORK SECURITY

11.1 INTRODUCTION

As the reliance of people, critical infrastructure, and national security on information flow and information networks has increased, so has the importance of maintaining the assurance of information stored or in flow over a network. The Internet is a global network and security attacks can originate from any corner of the world. The Akamai "State of the Internet" first-quarter report of 2008 [Aka08] lists China, the USA, and Taiwan as the top three origins of Internet attack traffic (16.8%, 14.3%, and 11.8% respectively) targeting a variety of services (database, remote procedure calls, etc.). The same report cites one estimate that 2% of all interdomain Internet traffic being raw attack traffic intended to deny services. Website hacks, spams, phishing attacks, and identity thefts have all made headlines in recent years. In this chapter, we provide an overview of issues, terminology, and techniques related to the security of the network. This is primarily at a broad level, although some topics are treated in greater depth than others.

Securing a network requires several *ongoing activities*. Security is not a one-shot solution, but something that needs to be considered on a continual basis. It is common

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

to classify these ongoing activities into the following:

1. *Assessment* of the network for its current state of security. This includes verifying whether or not known vulnerabilities have been patched and security policies are being implemented correctly.
2. Employing adequate *protection and prevention* mechanisms against security threats. These include the deployment of perimeter security mechanisms such as firewalls and the use of cryptographic protocols to secure communications to and from a network.
3. Active *detection* mechanisms to rapidly identify security attacks that may be occurring or have already been successful.
4. Incorporating policies, procedures, and techniques in place to have an effective *response* to security attacks.

We discuss assessment, protection/prevention, detection, and response as they relate to network security in a succinct manner in this chapter. We do not consider security in cellular networks and WLANs. They are considered in Chapters 7 and 9 respectively and should be read in conjunction with this chapter. We start in Section 11.2 by describing network communications and how they are vulnerable to security attacks. There, we also consider a brief overview of security services. Section 11.3 is devoted to mechanisms that are used to protect networks from security threats or prevent successful attacks; we discuss firewalls and cryptographic protocols. Intrusion detection is examined in Section 11.4 and response mechanisms are considered in Section 11.5. A summary is provided in Section 11.6.

11.2 NETWORK ATTACKS AND SECURITY ISSUES

We start with a very brief overview of how two or more devices communicate across networks. This is useful in understanding how security attacks are launched. We discuss some specific attacks that are instructive in how security is impacted in networks. Finally, we define the various security services commonly considered in the literature.

11.2.1 Network Communications

It is instructive to examine, at a high level, how two hosts on the Internet typically make connections to one another. This is useful to understand how attacks occur over the network. Our goal in this section, however, is not to explain protocols from a communications perspective (such as performance, reliability, and so on) or carefully explore their details. We note here that what is described below corresponds only to a typical scenario and at a very simple and high level. There are exceptions and many different possible variations for communications across the Internet, many of them dependent upon the applications, the network infrastructure, architecture, and so on. For example, the FTP behaves quite differently from the following description (it makes use of something like a control channel and a data channel making it different from typical client–server communications). Another example of communications that differ from what is presented below is with applications like Skype.

Let us suppose that a client application on host A on a network P wishes to connect to a server application on host B on a network Q. The client and server applications run as processes on the respective hosts. The client application creates data that is sent down the protocol stack to the transport layer. The transport layer adds information to this data in a structured manner, creating a *segment* that is passed down to the network layer. The TCP and the UDP are two common transport-layer protocols. The transport-layer segment forms the payload of a network layer *packet* or *datagram* usually carried by the IP. The IP datagram is further carried by a link or MAC-layer protocol in a *frame* on each link between the host A and host B (examples are Ethernet and WiFi – see Chapters 8 and 9). Each link may have its own physical-layer-dependent transmission mechanisms (e.g. DSSS or CCK with WiFi).

At the transport layer, a *port number* will identify the process in host A. Let us denote this port number as P_{A}. Host A will have an IP address that belongs to network P. Let us denote this IP_{A}. The tuple $\langle P_{\mathrm{A}}, \mathrm{IP}_{\mathrm{A}}\rangle$, which is sometimes called a socket, is a globally unique identifier of the client process that intends to communicate with the server process. Similarly, the server process will be associated with a port number P_{B} and IP address IP_{B}. A connection between the client and server can thus be uniquely identified through the tuple $\langle P_{\mathrm{A}}, \mathrm{IP}_{\mathrm{A}}, P_{\mathrm{B}}, \mathrm{IP}_{\mathrm{B}}\rangle$. The transport-layer segment consists of a header containing the source port P_{A} and the destination port P_{B}. The IP datagram has a header that contains the source IP address IP_{A} and the destination IP address IP_{B}. Routers on the Internet use these addresses to route packets appropriately (see Chapter 6).

NICs only recognize the MAC address. When the NIC in host A creates a MAC frame on the physical medium of network P, it typically uses a 48-bit source MAC address and a 48-bit destination MAC address (see Chapter 6 for more details on addressing). In the general case, host B is on a different network, possibly using a different link and physical layer. Thus, the destination MAC address does not belong to host B, but instead to a gateway or router that connects network P to other networks or the Internet. The IP address of the gateway is either manually installed in host A or host A finds this information using the *dynamic host configuration protocol* (DHCP). The DHCP is also used to assign IP addresses dynamically to hosts in a network. However, knowledge of simply the IP address of the gateway does not suffice, since the MAC address is necessary for the frame to be received by the gateway. A mapping of the IP address to the MAC address can be obtained using the ARP. Similarly, when a frame arrives at the gateway from the Internet to some host on network P, the gateway will have to use the ARP to determine the MAC address of the destination host. The gateway is responsible for routing the IP datagram in the received MAC frame to another router in the Internet, which forms a node on one of the available paths to the destination network Q. Such paths are determined using routing information through routing protocols like the RIP, OSPF and BGP.

How does the application process on host A know the IP address of host B? Usually, the IP address is not known. Instead, a domain name such as “www.cnn.com” which is human friendly is used in the application. It is necessary for host A to use the *domain name service* (DNS) to determine the IP address of host B. This has to happen *prior* to the actual data being sent in an IP packet to host B. Each network has a local name server that is known to every host in that network (possibly through the DHCP). Host A contacts the local name server when the application process in host A desires to send a packet to host B with information about host B (say "www.cnn.com"). If the local name server has cached information about the IP address of host B, it provides that information to host A immediately. If not, it contacts a root name server (there are only 13 of these worldwide). The root name servers have information about authoritative name servers that have information related to hosts on their networks. In the above example, the root name server may provide the IP address of the authoritative name server for network Q to the local name server of network P. The local name server of network P then contacts the authoritative name server of network Q to obtain the IP address of host B. Then the IP address is forwarded to host A. As we will see later, the DNS provides an avenue for malicious users to launch attacks by diverting packets to malicious IP addresses and also for malicious users to hide themselves.

Now suppose that host A was successful in finding the IP address of host B using the DNS. The application process in host A with port number P_A sends data to a process in host B with port number P_B. How did the process in host A know the port number P_B? Standard applications have standard port numbers. For example, a web server usually employs the port number 80, a Telnet server uses 23, a web server running the SSL uses 443, the simple mail transport protocol uses 25, and so on. Port numbers may also be changed after initial contact, as in the case of protocols like FTP or applications like Skype. Although port numbers for standard services are well known, this does not automatically imply that such services are not available at other port numbers. For instance, it is quite possible to run a web server at a port number other than 80. It is also possible to run some other service at port 80.

Services on servers "listen" for initial contact from clients at the standard port numbers. These are what we call "open" ports. When a packet from host A arrives at host B, it is sent up the protocol stack to the transport layer where the server which is listening at port number P_B receives the application data in the transport-layer segment. The server processes the data appropriately and responds to the client at port number P_A, which is known because of the initial packet received from the client host A. Some of the standard server port numbers are as follows: 80 for http (web), 25 for smtp (sending e-mail), 23 for Telnet, and 53 for DNS.

Figure 11.1 shows a very simplified view of some of the many protocols and applications that are common in networked communications today. It is to be noted that this is just a very small fraction of the protocols and applications in use today. Each of these protocols could perhaps create security problems, because they are capable of being abused by malicious entities in ways in which they were not anticipated to be used.

11.2.2 Why Security Attacks are Possible

The way communications usually occur between a client and server gives us a high-level idea as to why security attacks are possible. For instance, a malicious entity has simply to craft a packet (or set of packets) and send it to a certain port on a certain host on a certain network. If the port is open (i.e. the server is listening) and the software has some bugs, then the server will pick up the packet and behave abnormally. At a minimum, the server may simply crash. In the worst case, it may provide the attacker openings to take complete control of the host. Security problems occur for a variety of reasons, but the open nature of communications over the Internet, as previously described, assists malicious entities in launching attacks because of the inherent vulnerabilities that exist. The emergence of very large cyber-crime operations has moved network security attacks from the realm of

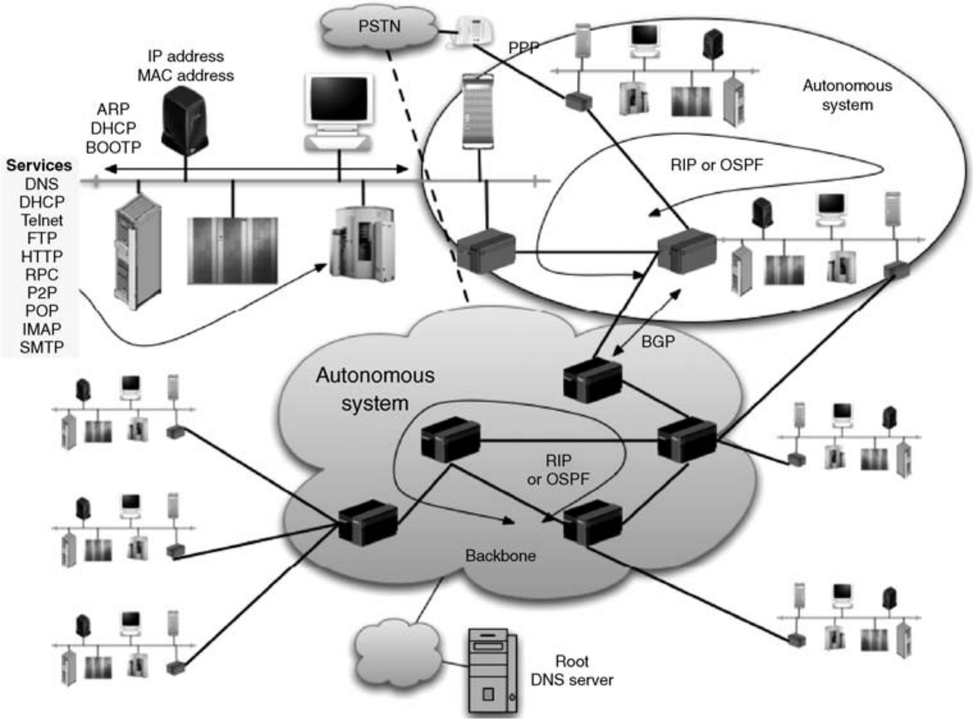


FIGURE 11.1 Simplified view of the many protocols that impact network communications.

hobbyists to criminal organizations, making them more dangerous, with the potential for great economic harm.

Figure 11.2 summarizes some of the technical reasons why security attacks are possible. We discuss these in more detail below. We will refer to a general malicious entity – a human, a criminal organization, or software – as Oscar in this chapter.

Information Leaks. Just as a robbers would stake out their physical target (e.g. a bank) first before attempting the actual robbery to determine how many security guards are there, what exits they could use, and so on, it is common for cyberattackers to stake out their victims' networks. Many protocols aid attackers by "leaking" information about networks and

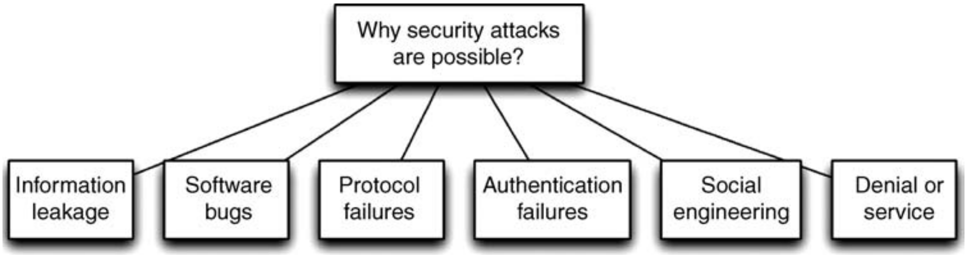


FIGURE 11.2 Summary of some of the reasons why security attacks are possible.

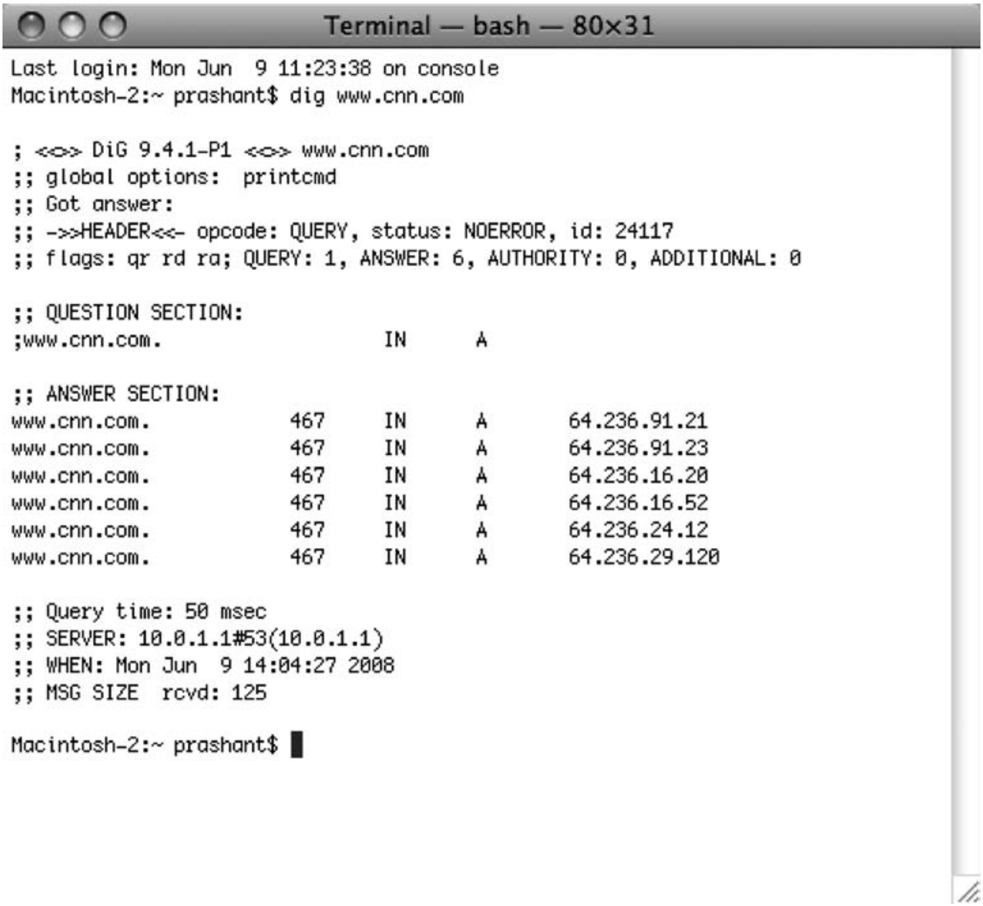
services running in these networks. Information leaks can be used to launch social engineering attacks (discussed later), for cracking passwords, mapping network topologies, and determining open services provided by servers. Information leaks are part of the normal operation of some protocols and it may be difficult to prevent such leaks. Other leaks can and should be plugged, especially those that occur when Oscar actively probes the victim networks for information. Some examples of information leaks are listed below.

Example 11.1: Wiretapping and Tempest The best source for obtaining information about a network is to tap the physical medium. If Oscar can gain access to the signals flowing on the wire leading into or out of the victim's network, then he can gain knowledge of a lot of information. Access to wiring cabinets must, therefore, be restricted. Unfortunately, physical security only goes so far. If an organization deploys WLANs that are not protected, then Oscar could potentially tap into such networks using powerful antennas and sensitive receivers from a remote location. Making use of security loopholes in routing, Oscar could attract traffic to a router he controls by falsely advertising shorter routes to destinations. *Tempest* is another form of leakage where radiation from wires and displays can be detected and the information that they carry be captured. In the military, it is common to separate equipment carrying classified information physically and employ careful design of circuits and grounding to prevent such leaks.

Example 11.2: Using DNS for Information Leaks The DNS maintains resource records that can contain a lot of information about hosts, such as the operating systems that they run, IP addresses, mail and web server aliases, and so on. A zone transfer with DNS involves copying all resource records from a primary name server to a back-up name server. If Oscar can perform a zone transfer, then he will have access to all of the information about all of the hosts in a network that the resource records may carry. Tools such as *nslookup* and *dig* (see Fig. 11.3 for an example of *dig*) can be used to obtain such information easily. It is increasingly common for network administrators to block unnecessary information in DNS resource records from being revealed.

Example 11.3: The Web as a Source of Information Leaks Search engines can be employed by Oscar profitably to obtain information about e-mail addresses, employee information, and sometimes even about host names, IP addresses, operating systems, etc. Older implementations of web servers allowed a complete listing of a directory's contents, thereby revealing some sensitive files. HTTP reveals quite a bit of information as well. The browser's user agent specifies what browser it is and what operating system it is running on. This information can be used by Oscar to determine if there are any software bugs on the system that can be exploited. HTTP may also carry information about the page where someone clicked the link that is being accessed and the data formats accepted by the browser (which may include formats with associated bugs).

Software Bugs. One common reason for successful security attacks is that servers listening at known ports have bugs in their implementation (e.g. buffer overflows). In such cases, it is possible for a malicious entity like Oscar to craft packets that can be sent to buggy services. When the service is compromised, it can enable Oscar to take control over the host. This means that Oscar can perhaps install malicious software on the host, use the host to launch



```
Last login: Mon Jun  9 11:23:38 on console
Macintosh-2:~ prashant$ dig www.cnn.com

; <<>> DiG 9.4.1-P1 <<>> www.cnn.com
;; global options:  printcmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 24117
;; flags: qr rd ra; QUERY: 1, ANSWER: 6, AUTHORITY: 0, ADDITIONAL: 0

;; QUESTION SECTION:
;www.cnn.com.                   IN      A

;; ANSWER SECTION:
www.cnn.com.            467     IN      A       64.236.91.21
www.cnn.com.            467     IN      A       64.236.91.23
www.cnn.com.            467     IN      A       64.236.16.20
www.cnn.com.            467     IN      A       64.236.16.52
www.cnn.com.            467     IN      A       64.236.24.12
www.cnn.com.            467     IN      A       64.236.29.120

;; Query time: 50 msec
;; SERVER: 10.0.1.1#53(10.0.1.1)
;; WHEN: Mon Jun  9 14:04:27 2008
;; MSG SIZE  rcvd: 125

Macintosh-2:~ prashant$ 
```

FIGURE 11.3 Example of using *dig* to get a list of cnn.com's web servers.

malicious packets targeted towards other vulnerable hosts, steal files containing valuable information that may be stored on the compromised host (or on other hosts on the network that trust the compromised host), and so on. Many services have seen bugs that have been exploited by cyberattackers. These include remote procedure calls (RPCs), FTP, sendmail, and web servers like Microsoft's Internet Information Server (IIS). IBM's Internet Security Systems X-Force 2007 Trend Statistics report indicates that the number of reported vulnerabilities (defined in the report as "any computer-related vulnerability, exposure, or configuration setting that may result in a weakening or breakdown of the confidentiality, integrity, or accessibility of the computing system") has increased almost every year since 2000. It has gone up from about 1300 in the year 2000 to more than 6000 in 2006 and 2007. Such vulnerabilities make it possible for attackers to develop a suite of exploits, sometimes in rapid succession.

Authentication Failures. A question that arises is why services do not authenticate the origin (and content) of packets such that maliciously crafted packets do not make it and, thus, they do not exploit software bugs. In fact, many services do try to implement mechanisms to verify the legitimacy of the source of a command or request. Unfortunately, not all such authentication mechanisms always work. You cannot trust the “From” field in an e-mail address, simply because it is too easy to forge. Similarly, many services implicitly trust the source IP address. Again, the source IP address is easy to spoof. We will discuss cryptographically secure authentication and integrity mechanisms later in the chapter.

Social Engineering. Over the centuries, malicious entities have exploited the lack of experience, wisdom, or judgment in people for their own benefits. One example is Oscar walking into a building wearing an official-looking uniform so that no one questions him and getting whatever it is he can lay his hands on. Social engineering is this approach that Oscar uses to exploit the naïveté or carelessness of users. Crafted e-mails that make users click on links or open attachments, fabricated websites that look like their original counterparts, and malicious wireless APs that are named like the trusted one are some examples of how Oscar may lure victims. Social engineering is not restricted to the cyberdomain. It is quite possible that Oscar makes use of some information he has obtained to call up a system administrator and ask them to reset a password. Social engineering attacks are easier in many cases than breaking encryption or overcoming other technological hurdles.

Example 11.4: Stealing Passwords Passwords can provide Oscar a wealth of information and access. Some protocols like *Telnet* send passwords in cleartext. If Oscar has access to the medium, then he can sniff the passwords. Sometimes, Oscar can get access to so-called “encrypted” password files, but which in reality are “hashed” passwords. Users typically pick passwords that are easy to remember, and some passwords are quite common. Since the hash function is public, Oscar can try a “dictionary” of passwords to see whether there is a match with the entries in the file he has obtained. Oscar can also fool victims into installing keylogger software (when they visit a fraudulent website by mistake) that records all keystrokes and reports them back to Oscar. When a victim types a password, the keylogger records it and transmits it to Oscar.

Denial of Service. Denial of service, sometimes called DoS, is primarily an attack against the availability of resources. Resources could mean the bandwidth in the network, information flow or access to stored information, computing resources, or software at the client/server side. Thus, a network DoS typically involves flooding a target with packets at the link or network layers; application-level service denial could include providing false information, interrupting information, or crashing a server. DoS on the client side could mean crashing the client software. It is impossible to prevent DoS, but it may be possible to mitigate its effects. We discuss some specific attacks that cause DoS in the next section.

11.2.3 Some Example Security Attacks

In this section we discuss some specific security attacks that will lead us to a general discussion of security attacks and security services in the next section. We do not provide an exhaustive list of attacks, but pick a few for illustration. The website of US-CERT (United States Computer Emergency Readiness Team) [Cert] is a good source for past and recent vulnerabilities and security incidents.

Address Spoofing and Sequence-Number-Guessing Attacks. Several services use the IP address or host name to provide access to the service. As discussed previously, it is very easy for Oscar to craft packets. Similarly, Oscar can easily spoof IP addresses and also host names. There have been instances of attacks where root access to certain hosts has been obtained by sending crafted packets with spoofed IP addresses. Oscar can accomplish a DNS zone transfer, discussed in Example 11.2, if the only authentication used by the primary name server is the IP address of the secondary name server.

Although IP address spoofing is easy, in many of the attacks it is not sufficient to spoof IP addresses. It is also necessary to guess sequence numbers of other protocols carried in the IP packet as payload (such as TCP or DNS). Consider TCP, for example. As mentioned earlier, TCP is the most common transport-layer protocol. It is used by many application-layer protocols like HTTP and FTP. TCP was designed to provide reliable service on top of the unreliable network layer provided by IP. So, among other things, TCP is connection oriented and it carefully maintains buffers, windows, and other resources to count segments and track lost segments. When host A wants to connect to host B, a "three-way" handshake occurs to set up the connection (see Fig. 11.4). First, host A sends a TCP segment with a SYN flag set (this is one of six flags, i.e. bits, in TCP for indicating information). Host B acknowledges the SYN segment with its own TCP segment with the SYN flag and ACK flag set. Host A completes the handshake with a TCP segment with the ACK flag set. Then data transfer begins. The TCP connection can be torn down using segments with either the FIN flag set or with the RST flag set.

As part of the three-way handshake, both the client and the server use initial sequence numbers (client_isn and server_isn in Fig. 11.4), which are incremented in the corresponding acknowledgments. If the IP address is spoofed and Oscar wishes to fool the server into

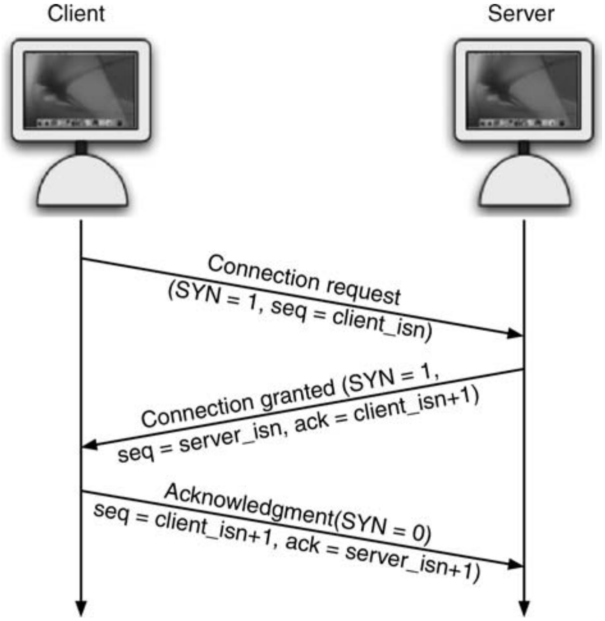

Client
Server
Connection request
(SYN = 1, seq = client_isn)
Connection granted (SYN = 1,
seq = server_isn, ack = client_isn+1)
Acknowledgment(SYN = 0)
seq = client_isn+1, ack = server_isn+1)

FIGURE 11.4 Three-way handshake in TCP.

believing that a legitimate client has connected with it, Oscar needs to “guess” the sequence number generated by the server. This is because the server’s SYN ACK segment is delivered to an IP address that does not belong to Oscar (and, hence, Oscar may not receive the response from the server). Oscar may be able to get the server sequence number only if he is on the same LAN and it is a broadcast medium.

The server sequence number is supposed to be random and difficult to guess. However, poor implementations of TCP have allowed malicious entities to guess the sequence number generated by the server easily. Many old TCP implementations use predictable ways of generating sequence numbers. For instance, the Berkeley implementation incremented the sequence number 128 times a second. Today, it is generally recommended that the sequence number be incremented 250 000 times a second. The idea behind this is that the round-trip time measured or predicted by Oscar will be random enough to prevent him from guessing the right sequence number. Similarly, spoofed DNS responses that can poison the DNS cache (see pharming below) can be generated if the sequence numbers associated with DNS requests can be guessed. The most secure way of generating sequence numbers such that they cannot be guessed is to use a cryptographic random number generator. We consider cryptography in more detail in Section 11.3.

BotNets and Distributed Denial of Service Attacks. As discussed previously, DoS impacts some resources of the victim, such as bandwidth or computing cycles. The “features” of many communication protocols can be exploited to deny services. TCP is a good example of such exploitation. Whenever a server receives a SYN segment from a client, it sets aside some resources (e.g. memory) anticipating a completed handshake and subsequent data transfer. As there are limited resources at a server, only a set number of connections can be accepted. Other requests are dropped. Oscar can make use of this “feature” to deny services to legitimate hosts by sending a flood of crafted SYN segments to a server with possibly spoofed source IP addresses. The server responds with SYN–ACK segments and waits for completion of the handshake (which never happens). Meanwhile, legitimate requests for connection are dropped. Such an attack is called a SYN flood attack and was the cause of DoS to popular web servers in recent years. Note that Oscar primarily makes use of a feature in a communications protocol to launch a DoS. The absence of authentication of the source IP address makes it difficult to block such attacks, since it is hard to separate legitimate requests from malicious requests.

Similarly, ICMP, UDP, and other protocols can be used to launch floods that result in DoS. ICMP echo requests and replies (commonly called pings) are used to check whether or not certain hosts are alive. An ICMP broadcast echo request is possible where *all hosts* in a network are requested to respond to the echo request. If the responses are sent to a victim that has limited bandwidth, then the victim will be swamped and have no access to the network. Such an attack is called the “Smurf attack,” where an ICMP broadcast echo request is sent with a source IP address that belongs to the victim. Obviously, the ICMP broadcast echo request is sent not by the victim, but by a malicious node that desires to deny services to the victim using the victim’s source IP address.

Distributed DoS (DDoS) attacks involve several hosts on the Internet near simultaneously launching one or more DoS attacks on a victim. Such DoS attacks can be SYN floods, Smurf attacks or other DoS attacks, but launched by a distributed set of nodes. DDoS attacks have recently made headlines by bringing down several popular websites in recent years, as well as launching attacks on root DNS servers. How do these multiple hosts launch attacks on the same victim? Security attacks are seldom one-shot events. Over time, Oscar, the attacker, installs his agents on some hosts by exploiting software bugs in them. These hosts, in turn, may help Oscar to automate the search for other vulnerable hosts to take over. Eventually, a large set of hosts that have backdoors and malicious software installed on them is created. The hosts could span multiple networks, domains, and even countries. The hosts are under Oscar's control whenever they are connected to the Internet. They are often called *zombies* or *bots.* A network of bots is called a Botnet. Botnets have been known to be used for a variety of malicious activities. DDoS is the obvious one, but Botnets have been used to generate millions of spam e-mails (distributing and hiding the origins of spam), for distributed vulnerability scanning to detect other potential zombies, and in some cases for distributed cryptanalysis.

Oscar could create a hierarchy of bots to hide his identity or location. Some of the bots have a *master* tool installed in them that controls other *agent* zombies. Communications between the agents and their master(s) could involve Oscar's manual intervention, which provides him the flexibility to change attacks, or they may be automated, which reduces Oscar's direct involvement but prevents him from adapting the attack to any defenses that may have been enhanced by the victims. Such communications can be encrypted or hidden to prevent easy detection. Oscar uses the masters to let the agents know when to launch the attack, who the victim is, how long to launch the attack, and what attack to launch. Attacks could be of multiple types, at different rates, involving time-varying subsets of agents. The agents could use their own IP addresses or spoof the IP addresses. Sometimes the spoofed IP address space could be designed to point the blame of launching the attacks on a particular victim. All of these make it extremely hard to detect or block DDoS attacks. A complete taxonomy of DDoS attacks is available in Mirkovic and Reiher [Mir04].

Worm Attacks. Worms are self-replicating malicious software programs that can crash hosts or services or open trapdoors for installing keyboard loggers or perform other malicious activity. Worms do not need active human intervention to propagate after the first instance of a worm being launched. Once a worm is installed on a host, it probes other networked hosts for bugs or vulnerabilities in services that can be exploited. This essentially means that the worm sends crafted packets to certain port numbers at IP addresses. If the services listening to such port numbers are vulnerable, then the worm can exploit such vulnerabilities to install itself on such hosts. For example, in July 2001, web servers running Microsoft's IIS software were discovered to have a buffer overflow bug. Although a patch was issued for this bug, not every host running IIS was patched. The Code Red (two versions) and Code Red II worms exploited this bug and spread rapidly across the Internet [Moo02]. It is estimated that Code Red infected at least 350 000 hosts.

The speed with which a worm spreads depends on the design of the worm (e.g. the rate at which it scans for other vulnerable hosts), whether patches exist for the vulnerability exploited by the worm, the number of hosts running the vulnerable software, and the clean-up rate [Che03]. The manner in which worms find other hosts to exploit can also influence their spread [Li08]. Finally, worms that use TCP as the transport protocol are limited by TCP's flow control scheme, while worms that employ UDP often compete with one another for bandwidth.

There are several ways in which a worm can find targets for propagation. Many early worms would randomly pick IP addresses to probe for vulnerabilities. However, this meant that many IP addresses would either not belong to hosts that existed or to hosts that did not run the vulnerable service or operating system, thereby limiting the spread of the worm. Others had a hard-coded sequence of IP addresses that would be probed. This meant that infected hosts would likely probe other infected hosts first. Recent worms are intelligent: they look for ‘neighboring’ IP addresses first. Some worms make use of “hit-lists” that have been previously generated by an attacker. Other worms use Internet search engines to discover vulnerable hosts. However, most search engines present the same set of results for a query, thereby reducing the set of hosts scanned for vulnerabilities. The most rapidly spreading worms use e-mail and entries in the address books of infected hosts to reach a variety of legitimate and potentially vulnerable hosts. In the past, exploits for vulnerabilities would not appear quickly, but it is common to see so-called “zero-day” exploits today. A zero-day exploit, for instance, can result in a worm that can be released on the same day that a vulnerability is discovered in a service. This makes it almost impossible to patch the exploit in time, enabling the worm to spread extremely rapidly. Botnets can also be used to launch worm attacks. In this case, the infected host becomes a member of the Botnet and has access to e-mail addresses or IP addresses that it can target as its victims.

Table 11.1 provides a list of some Internet worms and different properties of such worms. In-memory-resident worms do not write themselves to the hard drive. Simply rebooting the host will clear the host of the worm, but not the vulnerability. Thus, the host could be reinfected. Known worms can be detected by intrusion detection systems (IDSs, discussed in Section 11.4) using worm signatures.

Phishing Attacks. Phishing is an example of a social engineering security attack where legitimate users are fooled into revealing information such as logins, passwords, credit card numbers, and so on by making them visit websites that look like legitimate sites, but which are actually fakes run by criminal organizations. Legitimate users can visit such sites, for instance, by clicking on links that appear in e-mails that look legitimate.

Most phishing attacks target financial organizations like banks or e-commerce sites like Paypal or eBay. According to the January 2008 report of the Anti-phishing Working Group [APWG], a consortium of more than 3000 members, including nine of the top ten US banks, more than 90% of phishing attacks targeted financial organizations, with ISPs coming second at 3.8%. According to the same report, the USA, the Russian Federation, and China are the top three countries where most of the phishing websites are hosted.

Recently, a special form of phishing attack called “evil twins” has appeared, whereby WiFi APs are placed in areas (e.g. hot spots like coffee shops or hotels) close to where a legitimate service is being provided by some service provider. When a legitimate user tries to connect to such APs placed by Oscar, a web page similar to the one displayed by a legitimate service provider is displayed. It is common for subscribers to enter credit card and other sensitive information on these web pages, enabling Oscar to steal such information.

Pharming is a more dangerous security attack. As described previously, DNS is used to discover IP addresses associated with domain names. In the case of pharming, DNS caches can be poisoned with fake entries so that a user sees a fake website even if a legitimate URL is typed in the browser. DNS cache poisoning is possible when name servers use vulnerable versions of software that can be exploited with unsolicited DNS responses. Once again, the impact is similar to phishing attacks, where a legitimate user will reveal sensitive information to the bad guys. We discuss methods that have been proposed to combat phishing in Section 11.3.3.

TABLE 11.1 **Some Internet Worms**

Name	Launch date	Vulnerability exploited	Propagation mechanism	Remarks
Morris Worm	1988	Buffer overflow in sendmail and finger services	Random scanning	Earliest known worm
Code Red I	2001	Buffer overflow in Microsoft's IIS	Random scanning	In-memory resident
Code Red II	2001	Buffer overflow in Microsoft's IIS	Local subnet scan	Unrelated to Code Red I
Nimda	2001	Unicode vulnerability in Microsoft's IIS; bugs in Internet Explorer	Automatic rendering of HTML in e-mail, browsing of compromised websites; 50% of IP addresses with same first two octets, 25% with same first octet, 25% random	Affected both server and client
Slammer	2003	Buffer overflow in Microsoft SQL server	Random IP addresses	In-memory resident
Blaster	2003	Buffer overflow in Microsoft Windows RPC Interface	Primarily random IP addresses	Date-triggered payload that launches a TCP SYN flood DoS attack
Sobig	2003	E-mail attachment (needs user action)	E-mail addresses on host	Widespread
MyDoom	2004	E-mail and Kazaa P2P networks	Search engine for e-mail addresses	Also performs DDoS attack
Witty	2004	Buffer overflow in Internet Security Systems products	Botnets sending UDP packets to random IP addresses	Affected a small number of hosts, but considered nasty
Sasser	2004	Buffer overflow in Local Security Authority Subsystem Service in Windows XP or 2000	52% random, 25% last two octets are random, 23% last three octets are random	Impacted several systems by swamping their networks
Santy	2004	Vulnerability in phpBB used in discussion forums	Google search for hosts	Written in Perl
Storm	2007	E-mail attachment	Botnet and e-mail	Creates a powerful P2P Botnet that is hard to disable; used for spam

11.2.4 Defining Security Attacks, Services, and Architecture

In the previous subsection we saw some examples of security attacks, such as DoS, session hijacking, worms, and social engineering. One way of classifying security attacks is to consider their nature: whether they are passive or active. In the case of passive attacks, Oscar does not interfere with the information flow or storage (e.g. eavesdropping), making such attacks hard to discover. It is important to prevent such attacks. Active attacks (such as masquerading) involve interference and participation by Oscar. As they are hard to prevent, they must be detected and stopped as rapidly as possible. Security attacks can be of many types: eavesdropping (interception) on information and revealing such information, interrupting the flow or availability of information, masquerading as a legitimate entity to access services, information, or resources, and fabricating information with the aim of causing damage are all different security attacks. Security attacks usually do not occur in one shot. Oscar typically first engages in mapping out the victims network, resources, IP addresses, open services, and so on. This is sometimes called reconnaissance, and Oscar may try to get information that appears to be harmless if revealed, but may impact security later. This is followed up by exploitation of vulnerabilities, theft of information, taking over of hosts, etc. Bejtlich provides an excellent treatment of the security attack process [Bej04].

The common security services to protect against security attacks as defined in the literature are confidentiality, authentication, integrity, nonrepudiation, and availability [Sta03]. *Confidentiality* implies that information or data is kept secret from unauthorized entities, specifically Oscar. This is the most intuitive form of security service, where two communicating parties do not wish to reveal the contents of their transactions to a third party. In more rigid cases, the existence of the communication itself must not be revealed to unauthorized entities. Encrypting the messages and the identities of the two communicating parties is the most popular method of providing confidentiality. In the case of *authentication*, it is necessary for communicating parties to: (a) ensure at the start of communications that they are communicating with who they are thinking they are communicating with – that is, Oscar should not fool an honest Alice into thinking that she is communicating with an honest Bob; (b) ensure that, after communications have been established and verified to be between legitimate parties, Oscar does not hijack the communications session and interpose himself as one of the legitimate parties. The first part of authentication is often called *entity authentication* or *identification.* Identification is also used in transactions such as obtaining cash from an automatic teller machine. The second part of authentication is often called *message authentication* and it is combined with *integrity*. In such a case, once a legitimate communication has been established, it is necessary to ensure that any messages exchanged have not been modified, fabricated, reordered, replayed, or deleted. *Nonrepudiation* refers to a security service where a person, once they have sent a message, cannot deny having created the message. This service is similar in nature to a signature by the creator of a document, and *digital signatures* based on public key encryption schemes, discussed below, are employed to provide this service. *Availability* ensures that resources or communications are not prevented from access or transmission by malicious entities. DoS is the attack corresponding to the security service of availability. While the above security services are the most prominent ones, other security services also play a role in certain applications. *Authorization* is sufficient in certain cases instead of a sender authentication. Authorization allows an authenticated user or communication command to execute certain operations.

Such a security service would be sufficient, for example, to instruct a coffee machine to execute certain operations.

Note that all security services may not be present all the time, and different protocols and applications support different subsets of security services. Sometimes, architectural methods (using firewalls, screened subnets, and demilitarized zones) are necessary for ensuring some of the security services (e.g. confidentiality or availability).

11.3 PROTECTION AND PREVENTION

In this section we consider security mechanisms for protection against and prevention of security attacks. The common methods of protecting against security attacks are blocking unwanted or suspicious packets using firewalls and to encrypt and authenticate communications traversing the network. We consider firewalls and perimeter security in Section 11.3.1 and cryptographic protocols in Section 11.3.2. The interested reader is referred to the literature [Nor05, Ches03] for more details on firewalls. Stinson is a good reference that considers cryptography and cryptographic protocols [Sti05].

11.3.1 Firewalls and Perimeter Security

To block malicious packets from entering a network, it is common to employ firewalls. Firewalls in olden days referred to thick walls of brick constructed especially for preventing the spread of fires from one building to another. Firewalls today refer to hardware, software, and policies to prevent the spread of security attacks into an organization's (or individual's) network or host. As discussed in Section 11.2, attacks of many kinds occur due to maliciously crafted packets that arrive at the target network. If such packets can be identified and discarded, then they will no longer be a threat to the security of the network. This, in essence, is the idea behind firewalls. However, it is not trivial to identify such packets efficiently and correctly all the time. One advantage of using firewalls is that they provide a "point of access" to the network, which can be monitored instead of monitoring multiple hosts. Thus, firewalls are often said to provide "perimeter security" or security from the outside. Firewalls can be built on systems running a minimal number of software programs providing a controlled environment and reducing the chance of the device having exploitable bugs. However, firewalls often create a false sense of security and overconfidence: they are not foolproof and caution must be exercised even with firewalls protecting the network.

As shown in Fig. 11.5, the firewall sits between the "inside" and the "outside". The inside is usually what needs to be protected. The firewall controls what packets enter and what leave the network. How this may be accomplished differs depending on the type of firewall. The term "firewall" can mean many things today, all the way from a simple packet filter to a complex intrusion prevention system that is capable of examining a series of packets and reconstructing sessions for comparison with known attack signatures. We discuss some of these categories below. We note here that actual products sold today can be configured to do the simplest filtering or perform quite complicated attack prevention schemes.

Packet Filters. A *packet filter* is the simplest type of firewall. It filters incoming or outgoing packets based on *rules* created manually by the administrator of a network. Rules are

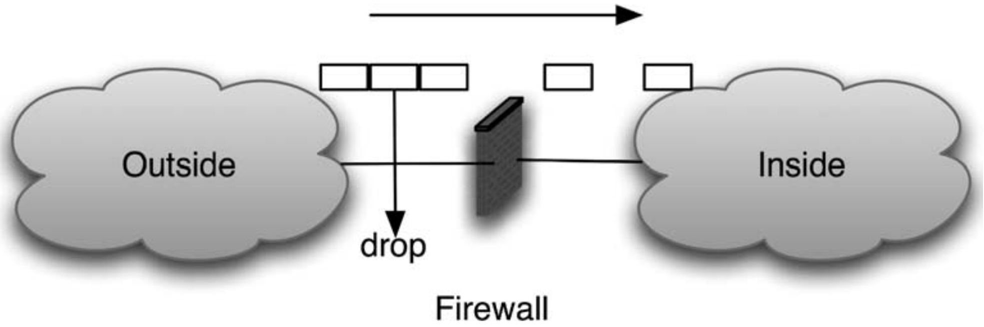


FIGURE 11.5 Schematic of a firewall.

considered sequentially. Packets are tested to see whether they satisfy the rules one by one. Packet filters usually have a default “drop” policy. This means that if a packet does not satisfy any of the rules that allow it into the “inside,” then it is dropped. Each packet is considered independently without consideration of previous or future packets, making packet filters fast and capable of handling high data rates.

The simpler the rules are, the faster is the filtering and the smaller is the performance hit. Cisco’s standard access control lists (ACLs) filter packets based solely on source IP addresses. In this case, it is easy to filter packets with source IP addresses that are obviously spoofed or other packets from sources that are not expected to communicate with the inside. Examples are IP packets that arrive from the outside with nonroutable source IP addresses, loopback IP addresses, multicast IP addresses, or IP addresses that belong to hosts in the inside. Nonroutable RFC 1918 IP addresses include the IP address ranges 10.0.0.0–10.255.255.255, 172.16.0.0–172.31.255.255, and 198.168.0.0–192.168.255.255. The loopback IP address is 127.0.0.1 and the multicast IP address range is 224.0.0.0–239.255.255.255. If certain malicious domains can be identified or if there is no need for receiving packets from certain geographical areas, then such packets can also be blocked using source IP addresses. If certain source IP address can be identified as temporarily being spoofed to launch attacks, then they can be blocked as well. Standard ACLs can also be employed for egress filtering, i.e. packets leaving an organization’s network, in a similar manner to prevent suspicious packets leaving the network or unauthorized communications from being completed.

However, standard ACLs cannot block packets to specific hosts on the inside or packets that correspond to specific protocols. The extended ACL from Cisco allows a packet filter to look at source and destination IP addresses, TCP or UDP port numbers, and TCP flags and make decisions on whether or not a packet should be allowed into the inside. For example, a rule that allows all hosts to connect to a host with IP address 136.142.117.13 if the source port number is larger than 1023 and they are connecting to port number 80 only is possible. This allows packets to a web server on the host with IP address 136.142.117.13, but not to other services that it may be running. Other firewall software (e.g. IPTables in Linux) and hardware have equivalent ACLs for filtering packets.

It should be noted here that it is quite hard to set packet filtering rules correctly. As the number of rules and the protocols involved increase, it becomes an extremely error-prone process, especially since the order in which the rules are enumerated matters. Since port 80 (used for web servers) is typically open, it is known that people abusing it by tunneling other applications within HTTP using protocols like SOAP (the simple object access protocol). Sometimes, system administrators open holes in the packet filter rulesets to accommodate certain protocols temporarily. Such holes must be plugged at the earliest possible time. Also, care must be exercised to restrict access through such holes to a limited number of hosts in the network.

Stateful Firewalls. As previously mentioned, the rules in the packet filter are considered in strict order, creating potential for configuration errors as the list of rules grows in size. One way of overcoming this problem is to use so-called dynamic packet filters or stateful firewalls. Dynamic packet filters build rules on the fly. The assumption here is that hosts on the inside are to be trusted. When they send packets to open connections with hosts on the outside, a stateful firewall builds a rule on the fly that allows packets from the specific external host (and port number at that host) to the specific internal host (and the port number at this host). The rule is deleted when the connection is terminated. This reduces the number of hard-coded rules and makes it difficult for Oscar to guess what packets may make it through a firewall.

For example, consider a host from the "internal network," say 136.142.117.221, that connects to a Telnet server in the outside network with IP address 130.215.17.13. Let us suppose that the port number on the client side is 1091. The server port number is 23 (which is the common port number used for Telnet). A new ruleset would be created in the dynamic packet filter as follows: "Allow packets from host 130.215.17.13 with port number 23 to host 136.142.117.221 at port number 1091." The dynamic packet filter will examine all packets to make sure that the TCP SYN, SYN–ACK and ACK segments were successfully sent or received. When the dynamic packet filter observes TCP FIN segments or RST segments that are used for tearing down the connection, it deletes the ruleset, thereby disallowing further communication from 130.215.17.13. In Cisco devices, this is called a "reflexive" access list. Reflexive ACLs can be a burden on routers in terms of performance.

Proxy Firewalls. Packet filters can still be fooled through a variety of loopholes that exist (e.g. by sending fragmented packets). In order to determine whether or not packets are legitimate, it is often necessary to look at the application payload. Sometimes it is even necessary to reconstruct the application data. This is possible if proxy firewalls are used. Proxy firewalls consist of hardened hosts (usually dual homed) that run reduced modules of certain applications. When an internal host makes a connection to the outside, it really makes a (say TCP) connection with the proxy firewall. The proxy then makes a connection to the external host. Thus, there are two connections that exist. External hosts only see the proxy firewall. They are not even aware of the existence of other internal hosts. When packets are returned, they make their way up the protocol stack, where the application (with reduced features) reconstructs the data. If the data is legitimate, then it is forwarded to the internal host. Moreover, Oscar can gain very little knowledge during reconnaissance because internal hosts are not visible to the outside world. However, proxy firewalls create performance bottlenecks. They also do not support a variety of applications, often frustrating legitimate network communications.

Architectural Approaches. Architectural approaches can approximate the benefits of proxy firewalls and yet keep performance levels reasonable. One common approach is to screen

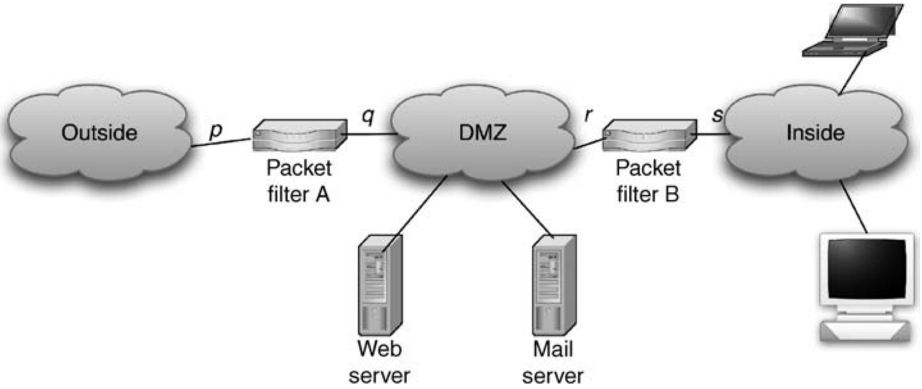


FIGURE 11.6 Schematic of a screened subnet and DMZ.

the inside from the outside by using one or more packet filters. In Fig. 11.6, for example, packet filter A allows packets (from most legitimate hosts on the outside) through interface *p* to reach either the web server or the mail server. As almost anyone can reach these servers, this is called a demilitarized zone (DMZ). If it is also a router, then it does not advertise the existence of the inside network to the outside world. Similarly, packet filter B allows packets from either the web server or the mail server to the inside through interface *r*. Thus, the inside network is screened from the outside.

Note that packet filters can also be used to stop packets from the inside from going out (for example, through interfaces *s* and *q* in Fig. 11.6). This may be necessary if hosts on the inside have been compromised and are launching attacks or hosts are trying to access services not allowed by corporate policy.

Nowadays, firewalls are more than simple packet filters. They can maintain state, do load balancing (if multiple firewalls are used), do some inspection of application payloads, detect attacks based on known signatures, maintain logs useful for forensics or analysis, and also act as end points for connectivity to mobile users who need to connect to the inside from the outside. For example, firewalls can now be the terminating points for VPN connections using IPSec or SSL, which make use of cryptography to prevent outsiders from connecting to the inside or monitoring connections made by mobile employees. We discuss cryptographic protocols next.

11.3.2 Cryptography and Cryptographic Protocols

Security services such as confidentiality, entity and message authentication, integrity, and nonrepudiation can be provided to communication protocols using cryptography. In this section, we provide a brief overview of the important topics in cryptography and cryptographic protocols. More details can be found in the literature [Sta03, Kau02, Ches03].

Cryptographic Primitives. Cryptographic protocols make use of cryptographic *primitives* that are used to provide the required security services. A classification of such primitives is shown in Fig. 11.7. Cryptology is the broad discipline that includes the science of designing

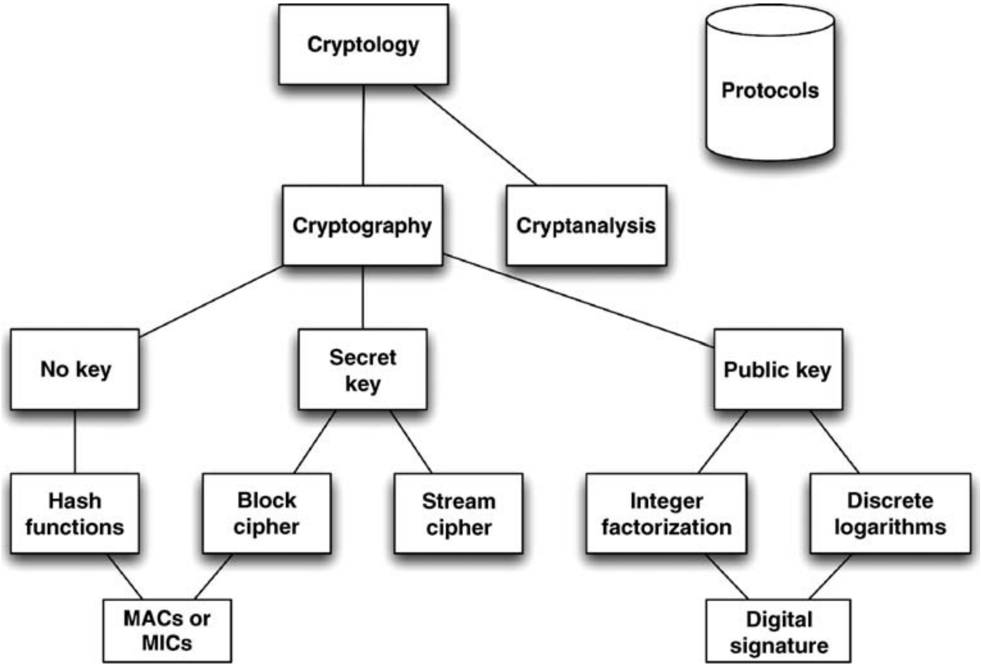


FIGURE 11.7 Classification of cryptographic primitives.

ciphers (cryptography) and that of breaking ciphers (cryptanalysis). Data that is to be encrypted is called "plaintext" and the result of encryption is called "ciphertext."

It is common to denote two communicating parties as Alice and Bob and the adversary or opponent as Oscar. Mathematically, the encryption of a plaintext x into a ciphertext y using a key k is written as

$$y = e_k(x) \tag{11.1}$$

The corresponding decryption is written as

$$x = d_k(y) \tag{11.2}$$

Ideally, we would like the encryption scheme to be such that it cannot be broken at all. Since there are no practical methods of achieving such an unconditional security, encryption schemes are designed to be computationally secure. The encryption scheme has to be powerful, in that, given significant computational resources, an adversary must not be able either to find the key or decrypt the message in a reasonable time. Alternatively, if either the key or the plaintext can be determined in a short time, then it should cost the adversary much more than what the value of the secret information would be to him. Usually, it is assumed that Oscar has knowledge of how the algorithm works, but not the key. Also, because of the standard formats of data packets and control messages in voice networks, Oscar usually has access to a limited number of plaintext–ciphertext pairs that he can use to perform a known plaintext attack to recover the key. Once the key is recovered, all subsequent ciphertext can be decrypted easily.

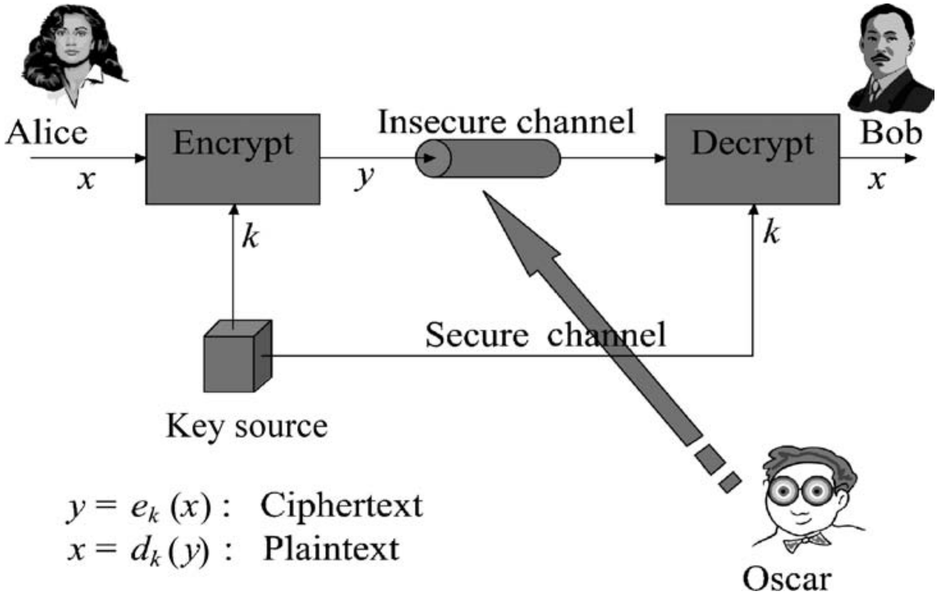


FIGURE 11.8 Conventional encryption model.

So far, we have discussed security services and we have said that encryption can provide some of these services. What we have not discussed are the encryption algorithms that are employed or can be employed within these mechanisms. The details of these algorithms are beyond the scope of this book, and the subject matter forms a wide area of interest in itself. In this section, we will briefly mention some of the algorithms that are in wide usage today. We will also discuss the key sizes that are required to make these algorithms secure.

Ciphers or encryption algorithms can be classified into *secret key* and *public key* categories. Encryption schemes have been available through the ages and have all been what are known as secret-key algorithms. Here, the communicating parties (Alice and Bob in Fig. 11.8) share a secret key that they use to encrypt any communication between themselves. Usually, the encryption and decryption algorithms use the same key; hence, such algorithms are also called symmetric key algorithms. Block ciphers such as the AES also fall under this category. Figure 11.8 illustrates a schematic of a conventional encryption scheme. The opponent Oscar has access to the insecure channel and, thus, the ciphertext. However, he has no knowledge of the secret key k shared by Alice and Bob.

Secret-key algorithms such as AES are based on two principles: confusion and diffusion. The former introduces a layer of scrambling that creates confusion as to what exactly might be the transmitted message. The latter creates a randomness whereby the effect of changing a small part of the plaintext message will result in changing half of the encrypted ciphertext. This eliminates matching patterns or frequencies of occurrence of messages. Most secret-key algorithms are thus unbreakable except by brute force [Sil00]. If the length of the key of a secret-key algorithm is n bits, at least 2^{n-1} steps are required to break the encryption. Today, a key length of 80 bits is considered to be sufficiently safe from brute-force attacks, even though a key size of 128 bits is usually recommended.

Example 11.5: Security of the Data Encryption Standard against Brute-Force Attacks The data encryption standard (DES) is a block cipher that encrypts plaintext messages in blocks of 64 bits using keys that are 56 bits long. The total number of keys is 2^{56}. On average, half of them will have to be examined to determine the right key if a known plaintext–ciphertext pair is available. If a 500 MHz chip is employed for this attack, and one decryption (or encryption) can be performed in one clock cycle, to test 2^{55} keys, it will take $2^{55}/(500 \times 10^6)$ s = 834 days to break the encryption. This is not very secure if 834 chips are used in parallel, since the key can be obtained in a single day. The total cost will be about $16,680 if each chip costs $20!

Example 11.6: Security versus Advances in Chip Speeds DES was broken in less than a day in January 1999 at a cost of $500,000. Today, it is virtually impossible to break a well-designed block cipher with key sizes of more than 80 bits (which translates into examining around 280 keys by brute force). However, a common assumption (called Moore's law) is that processor or chip speeds double every 18 months, thereby weakening any encryption scheme with time. For example, using a speed of 500 MHz for today, then an encryption scheme that employs key sizes twice that of DES (i.e. 112 bits) would be broken in a day 100 years from now.

The primary advantage of secret-key algorithms is that they are fast, and at the huge data rates that are being supported by today's networks it is virtually impossible to employ *public-key* algorithms (discussed below). However, since every pair of users has to have a key, for a communication system with N users, at least $N(N-1)/2$ keys need to be created and distributed. This is not a trivial exercise and has its own weaknesses.

Example 11.7: Number of Keys with Symmetric Key Encryption Algorithms Assume a small corporate network with 500 computers. A total of 124 750 keys are required (one between each pair of computers). Each computer needs to store 499 keys associated with the remaining computers. Suppose an employee gets a new hand-held personal computing device. Not only will this hand-held device need to load 500 new keys, but the remaining 500 old computers also need to be each updated with a key for the handheld computer.

There are techniques of key distribution for symmetric-key algorithms, such as the Needham–Schroeder key distribution scheme and Kerberos [Sta98]. All of these schemes need several handshaking steps and also an initial configuration of computers with *master keys*. Such master keys can be distributed physically in a secure manner. However, key distribution is still a potential weakness in the system. Another way of generating fresh keys for each communicating session is to use the master key and a one-time random number (called nonce – or number used once) as inputs to a one-way hash function to generate a key. Alternatively, the nonce can be encrypted using the master key.

Block ciphers such as AES encrypt blocks of data at a time. Stream ciphers encrypt bits or bytes of data [Sti95]. The advantage of stream ciphers is that there is no need for buffering data up to the block size or for padding. Stream ciphers may also be more suitable for jitter-sensitive voice conversations. The disadvantage is that these have to be used carefully, because encryption with stream ciphers uses simple XOR operation. Thus, it becomes necessary to use different key streams for encryption each time (because a simple XOR can be inverted to obtain a previously used key stream).

DES was the secret-key encryption standard for over 20 years. NIST examined proposals for the AESin 1998. Of the five candidate algorithms, NIST selected Rijndael as the algorithm for AES in October 2000. A variety of factors were considered by NIST to determine the suitability of the algorithm for a standard. Security (resistance to cryptanalysis, mathematical soundness, randomness of the algorithm output), cost (licensing requirements, computational efficiency on various platforms, memory requirements), and algorithm implementation characteristics (ability to handle variable key sizes and block lengths, implementation as stream ciphers and hash functions, hardware and software implementations, and algorithm simplicity) are three categories used for evaluation of these algorithms. In addition to these standards, several freeware and other secret-key algorithms are available, such as IDEA, RC4 and Blowfish [Sta98]. RC4, in particular, has been widely employed in web browsers, as well as in wireless networks like IEEE 802.11.

Public-key encryption is a radical shift in the way data is encrypted. Diffie and Hellman (DH) introduced the concept in 1977. With secret-key algorithms, we have a situation that is similar to having a locked mailbox for *each pair of users.* Both users associated with a mailbox share a key that can unlock or lock the mailbox. Consider Fig. 11.9. Here, if Alice desires to communicate with Dan, she unlocks the mailbox shared between her and Dan, deposits the message, and locks the mailbox again. The message is now accessible only to Alice and Dan, who also has an identical key.

Clearly, the number of mailboxes required for N users like Alice and Dan is $N(N-1)/2$. For example, we have six mailboxes for four users, as shown in Fig. 11.9. The situation described above is not the natural way in which we employ mailboxes. Mailboxes are

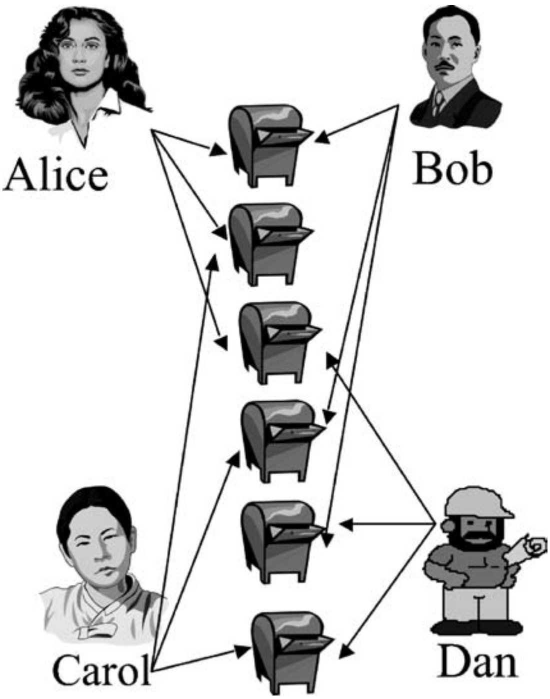


FIGURE 11.9 Multiple mailboxes with secret-key encryption.

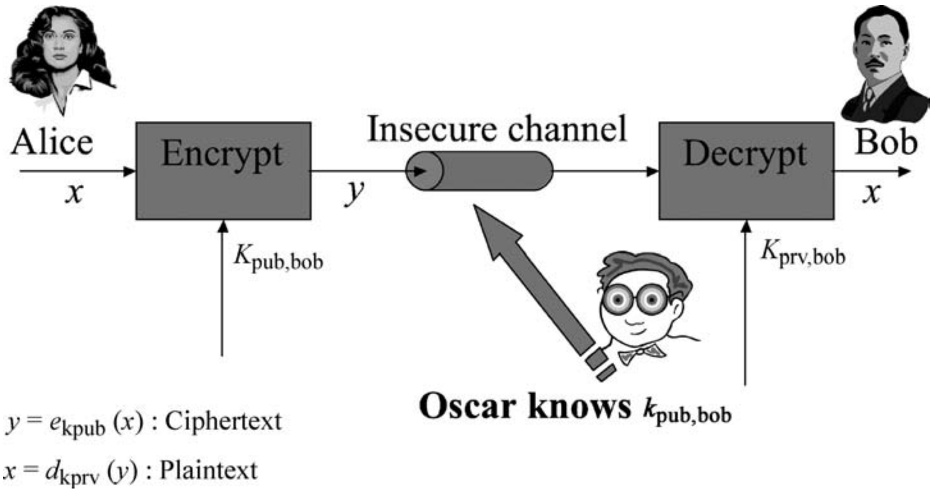

Alice
Encrypt
Insecure channel
Decrypt
Bob
x
y
x
$K_{\text{pub,bob}}$
$K_{\text{prv,bob}}$
Oscar knows $k_{\text{pub,bob}}$
$y = e_{\text{kpub}}(x)$: Ciphertext
$x = d_{\text{kprv}}(y)$: Plaintext

FIGURE 11.10 Public-key encryption scheme.

associated with individuals, not pairs of communicating parties. The natural way to employ a mailbox is as described in the following example. Alice owns a mailbox. Only she has a key to lock or unlock the mailbox (i.e. only Alice has complete control over the mailbox). *Any other person* who wishes to communicate with Alice will deposit the message through a *slot* in the mailbox. Once the message is deposited in the slot, *only Alice has access to it*. Even the originator of the message cannot retrieve it, although they may regenerate the message from knowledge of the contents.

Public-key algorithms are similar to this example. Each individual has a pair of keys: the public key and the private key. As the name suggests, everyone knows the public key. So, anyone can employ the public key to encrypt a message intended for the owner of the key. The public key is like the slot in the mailbox. Only the owner knows the private key. As a result, once the message is encrypted using the public key of the owner, only they can decrypt the message. Not even the originator of the message can decrypt it once the message has been encrypted.

Figure 11.10 shows the schematic of a public-key encryption scheme. Note that there is no longer a need for secure transfer of the key. Alice encrypts a message intended for Bob with Bob's public key $K_{\text{pub,bob}}$. The ciphertext is decrypted by Bob via his private key. The design criterion for public-key algorithms is as follows. Given a function $f(k, x)$, the following properties always hold:

- it is extremely easy to compute $y = f(k_{\text{pub}}, x)$
- given k_{pub} and y, it is computationally not feasible to determine $x = f^{-1}(k_{\text{pub}}, y)$
- with a knowledge of k_{prv} that is related to k_{pub}, it is easy to determine $x = f^{-1}(k_{\text{pub}}, y)$.

Example 11.8: Trapdoor One-Way Functions Functions that have the above properties are called trapdoor one-way functions. Examples are the factorization problem and the discrete logarithm (DL) problem. The former is based on the fact that it is easy to multiply prime factors to arrive at a composite number (e.g. it is easy to find 7 × 17 × 109 × 151 = 195 821, but it is quite a hard task to split 30 616 693 into its prime number factors). The latter is based on the fact that it is easy to determine what 2^{23} mod 109 is (the answer is 77). It is quite hard to find out what u is given $2^u \bmod 109 = 68$. Note that with real number arithmetic, it would have been trivial to determine u, as $u = \log_2 68$. The modulo function, which reduces the operations to be on set of numbers that are nonnegative integers less than 109, makes this problem very hard to solve. Integer factorization is employed in RSA (from Rivest, Shamir, and Adleman) and the DL problem is used in the DH key exchange protocol and digital signatures.

Example 11.9: The DH Key Exchange Protocol The DH key exchange protocol is based on the DL problem discussed in Example 11.8. Let us suppose that Alice wishes to exchange a session key with Bob without sharing any secret with him. Alice chooses a base α and a large prime number p that are publicly known. She only chooses a random private number a. She computes $k_{\text{pubA}} = \alpha^a \bmod p$, which she sends to Bob. Note that given k_{pubA}, α, and p, it is computationally impossible to determine a. Similarly, Bob chooses a private random number b and computes $k_{\text{pubB}} = \alpha^b \bmod p$, which he transmits to Alice. Once again, it is extremely difficult to determine b. After obtaining the public keys of each other, Alice and Bob raise these public keys to the exponent corresponding to their own private numbers a and b respectively. That is, Alice will compute

$$k_{\text{s}} = k_{\text{pubB}}^{a} \bmod p = \alpha^{ab} \bmod p$$

Bob computes

$$k_{\text{s}} = k_{\text{pubA}}^{b} \bmod p = \alpha^{ab} \bmod p$$

This way, both Alice and Bob have generated a common session key. An adversary Oscar cannot determine this key without solving the DL problem. At least, there is no known solution for obtaining the session key other than by solving the DL problem.

RSA has been the most popular public-key algorithm. It employs integer factorization. The DH key exchange protocol based on DLs is also very popular in wireless networks. This protocol is described in Example 11.9 and is commonly employed for key exchange for web transactions, e-commerce, and IP security. The digital signature standard (DSS) is also based on DLs. Signature schemes based on RSA are also widely employed.

However, in the case of public-key algorithms, Oscar, the opponent, is aware of Bob's public key and this adds an additional parameter to the problem. Since public-key algorithms are based on mathematical structures, for small key sizes there are well-known results or tables that can be employed to break the encryption. As such, the key sizes are extremely large compared with secret-key algorithms. Today, for good security, public-key algorithms need keys that are 3 to 15 times larger than their secret-key counterparts. Because the mathematical bases on which public-key algorithms work are well known, they are susceptible to analytical attacks and require much larger key sizes than secret-key algorithms. The mathematics of elliptic curves is also being employed in encryption schemes, since they need smaller key lengths than RSA.

Table 11.2 presents the key lengths and the time required to break some of the well-known public and secret-key algorithms [Sil00]. The values in this table are based on the assumption that $10 million is available for computer hardware. The key sizes in each row are equivalent.

TABLE 11.2 Cost-Equivalent Key Lengths (in Bits) of Various Encryption Schemes

Secret-key algorithm	Elliptic curve	RSA	Time to break	Memory
56	112	430	<5 min	Trivial
80	160	760	600 months	4 Gb
96	192	1020	3 million years	170 Gb
128	256	1620	10^{16} years	120 Tb

The mathematical operations for public-key algorithms are quite computationally intensive. Consequently, the encryption rates are quite small and public-key algorithms are rarely used for bulk data transfer. Instead, they are employed to exchange a *session key* between a pair of communicating entities who will then use the session key with secret-key algorithms for the duration of that communication (bulk data encryption). This ensures that a new session key is employed each time a communication is initiated, thereby reducing the possibility of an adversary breaking the encryption scheme.

Although the public key of an honest party like Alice can be made public, its authenticity needs to be verified, since Oscar can claim to be Alice and publish his key as hers. It is common to use *digital certificates* signed (see digital signatures below) by one of a few trusted certification authorities to verify the authenticity of the public key. This approach is used in modern web browsers for e-commerce applications. We discuss this in more detail in Section 11.3.3.

We include hash functions in the classification in Fig. 11.7. Hash functions are not strictly encryption schemes. They map any-sized data to a fixed-size digest. Given the digest, it is considered infeasible to obtain any data that maps to the digest if the size of the digest is at least 160 bits. Popular hash functions in use today are MD-5 and SHA.

Providing Security Services Using Encryption. Encryption schemes and hash functions are widely employed in password protection schemes and ACLs that are used for access control – the ability to allow or deny people access to certain resources based on their identification. Identification or entity authentication by itself is an important security service that needs to be provided for a variety of applications. Access to an automatic teller machine, logging on to a computer, identifying the user of a cellular telephone to the network, etc. involve identification schemes. Note that there is a difference between identification and message authentication. When we talk about message authentication (discussed next), there is usually some information-containing message that is exchanged between the parties and one or both parties need to be authenticated. Identification schemes (sometimes referred to as *entity authentication*) involve real-time verification of a party's identity but *need not* involve exchanging information-bearing messages.

Weak identification schemes are based on passwords or PINs that are time invariant. Usually the password or PIN value is compared with a securely stored hash value. Such schemes are easily susceptible to replay attacks especially if the password or PIN is transmitted over the air in an insecure manner. *Challenge–response* identification or *strong identification* schemes are usually employed in wireless networks. Here, Alice proves her identity to Bob by demonstrating knowledge of a secret, rather than presenting the secret itself. For this purpose, a quantity called the "nonce" is used. A nonce is a value employed no more than once for the same purpose and eliminates replay attacks. Random numbers, time stamps, sequence numbers, etc. are used as a nonce in practice. One example of a challenge–response protocol is as follows:

1. Alice is registered with Bob via a password and user name;
2. Bob sends Alice a random number (challenge);
3. Alice replies with an encrypted value of the random number where the encryption is done by using her password as the key (response);
4. Bob verifies that Alice indeed possesses the key (the password).

An eavesdropping Oscar cannot replay the response because the challenge is different if he tries to contact Bob. Oscar also cannot determine the password, because the encryption scheme is sufficiently strong and the password is *never* revealed.

Message authentication is a security service that provides two functions: sender authentication and message integrity. By sender authentication, what we mean is that the receiver can be assured that the message has been originated from the person who claims to have sent the message. Message integrity assures a receiver that no one has modified the message in transit. Both these functions can be accomplished by adding a *keyedmessage digest* (MD), *message authentication code* (MAC), or *message integrity checks* (MICs) to a message. The MAC here should not be confused with the MAC layer discussed in Chapter 4. Block ciphers and hash functions can be used to create MACs. These are checksums on data created using block ciphers or hash functions with a shared secret key between the communicating parties. MACs or MICs provide message authentication and integrity. If Oscar were to fabricate a message or modify a legitimate message, then the checksum would always fail, alerting the receiver of a problem with the received data. The CBC-MAC, which uses block ciphers, and HMAC, which employs hash functions, are popular standard implementations of MACs.

The way a MAC is used to provide message authentication is as follows. It creates a fixed-length sequence of bits that depend on the message itself and a secret key shared between the communicating parties. Irrespective of whether the message is a few kilobytes long or hundreds of megabytes, the MAC creates a sequence of bits of fixed length that directly depends on the message and the key. This sequence of bits is appended to the message and then the result is transmitted over the insecure channel. Note that the message could be sent in plaintext form if confidentiality is not an issue. It is computationally infeasible to create a replica of the MAC without the message and key. If the message is modified in transit, then the receiver can discover this fact by creating a MAC from the received message and comparing it with the transmitted MAC. Since the secret key is shared only between the communicating parties, it also assures the receiver of the origin of the sender.

A keyed MD operates in a slightly different manner. The MD depends only on the message, not the key. Hash functions are used to create MDs. The message is appended with the MD and the result is encrypted using a session key shared between the communicating parties. This way, both the message and the MD, which verifies it, are kept secure. The MD has to be sufficiently long to prevent what is known as the "birthday attack." Given an MD of length b bits, with a good probability, a fake message with the same MD can be generated in $2^{b/2}$ trials. This result is due to the fact that good probabilities of finding two people with the same date of birth exist in a group with roughly the square root of the number of days in a

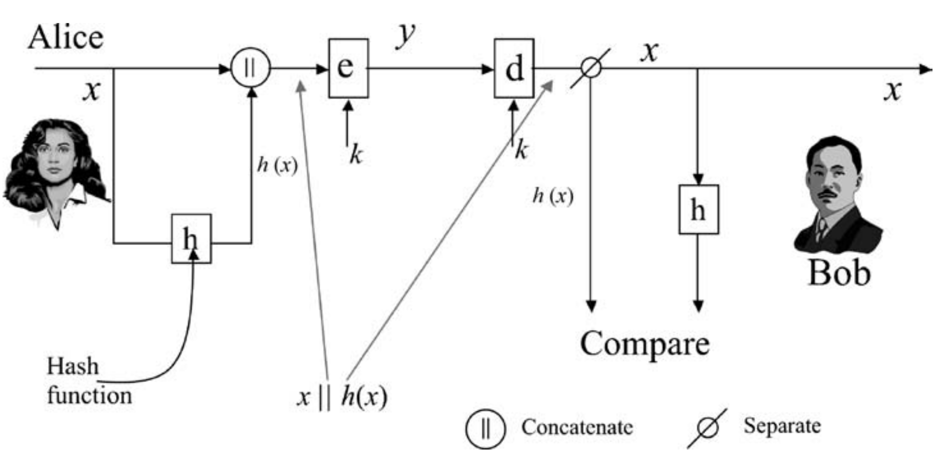


FIGURE 11.11 Message authentication with hash functions.

year. That is, in a group of 20 people, it is quite likely that any two would have the same birthday.

Figure 11.11 shows a schematic of message authentication with hash functions. On the left-hand side, Alice concatenates the message x and its hash value $h(x)$ together before encrypting the result with the secret key k. The ciphertext $y = e_k(x\|h(x))$ is transmitted over an insecure channel. Bob decrypts the ciphertext y and expects to find a message and its hash value concatenated together. He separates the message x from the hash value, computes a new hash value and compares the two together. If the ciphertext is modified or replaced in between, then Bob is able discover this fact easily. No one can impersonate Alice, since it is computationally impossible to create a ciphertext that decrypts into a message and its hash value without knowledge of the key k. Thus, both sender authentication and message integrity are assured. The interested reader is referred to [Sta98] for other schemes for message authentication. Using the hash function is generally preferred because of its speed.

Digital signatures are like a physical signature. They attest some information and are bound to that information. Typically this involves encrypting the hash value of some information with the private key of a public key/private key pair. Suppose Alice generated some data and created a digital signature of the data. Anyone can verify the signature, because decrypting the signature requires the public key, which is available to everyone. No one except Alice can generate the signature, because she is the only one in possession of the private key. Recall that knowledge of the public key does not help Oscar or others deduce the private key.

Example 11.10: Nonrepudiation and Digital Signatures We considered sender authentication and message integrity in this chapter. This does not, however, assure nonrepudiation. For instance, let us suppose that Alice is a consumer and Bob an e-commerce service provider. Bob claims that Alice placed an order with him for purchasing books worth \$350 and Alice denies the transaction. Alice claims that she had requested books worth only \$100. Both of them are able to produce ciphertexts and messages purportedly used in the transaction. Since both parties know the shared session key, it is impossible to verify who is being truthful and who is not. Public-key algorithms and digital signatures can be employed to resolve such situations.

We know that *only* the owner of the key knows the private key part of a public-key algorithm. Consequently, this information can be used to bind the owner to a message transmitted by them. Popular public-key algorithms operate such that it is possible to encrypt a message using a private key as well. We can compare this to the following scenario. Only the owner of a mailbox can slip a message through the slot because only they have access to the private key that opens the mailbox. No one other than Alice can encrypt the message using her private key (or produce a meaningful ciphertext that can be decrypted with her public key). The problem with this encryption is that *anyone* will be able to decrypt the message because the public key dual is available to everyone. But this is exactly the concept of a signature. If Alice were to sign a document, then this means that anyone should be able to verify her signature. However, no one should be able to forge her signature. This is indeed the case here when a message is encrypted using the private key.

Digital signatures take the concept a step further. The entire document *need* not be encrypted. As already discussed, this process would be extremely slow. Instead, an MD of the message is "signed" or encrypted with the private key. The encrypted "signature" is appended to the message. Once again, since it is computationally impossible to derive a message from the hash, the signature and the message are bound together. If the document needs to be confidential, then the usual encryption procedures can be employed after the signature is applied. Fig. 11.12 shows how digital signatures can be applied. Here, k_{AB} is a session key that is used to keep the document confidential.

Cryptographic Protocols. The cryptographic primitives discussed above are used in cryptographic protocols, which are designed with specific security objectives in mind. Cryptographic protocols are notoriously hard to design, since they will likely have pitfalls that are hard to detect [Kau02]. A good example of a cryptographic protocol that fails to meet most of its security objectives is the WEP protocol used in legacy IEEE 802.11 WLANs [Edn04]. Moreover, cryptographic primitives make use of keys shared between communicating parties. Establishing secret keys between legitimate parties interested in communicating, such that Oscar does not obtain any knowledge of the keys, is not trivial

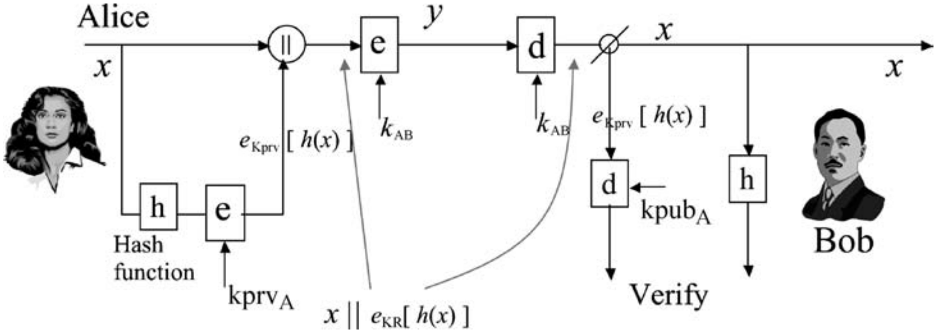


FIGURE 11.12 Digital signatures.

and itself requires cryptographic protocols. Key establishment is usually based upon master keys established with trusted third parties or public-key cryptography.

Most well-designed cryptographic protocols have three phases. In the first phase, the communicating entities *identify* or *authenticate* themselves to one another. In some cases the entity authentication is unilateral (i.e. Alice authenticates herself to Bob, but not vice versa). Entity authentication makes use of passwords, PINs, pass phrases, biometrics, security tokens, and the like. Challenge–response protocols that do not require an entity to reveal the password, but only demonstrate knowledge of the password, are commonly used for entity authentication. In the second phase, or as part of the first phase, the communicating entities also establish keys for security services to be provided next. Establishment of keys can be in two ways: key transport or distribution, where one party generates the keys (or a master key) and transports it securely to the other party, or key agreement, where both parties exchange information used in the secure creation of the same key at both ends. It is common for both parties to exchange random numbers, sequence numbers, or time stamps (called nonces) which are used as input in key generation. In the third phase, the established keys are used to provide confidentiality (through encryption with a block or stream cipher) and integrity (through MACs or MICs). We briefly describe some examples here.

Kerberos. The term Kerberos is originally from Greek mythology and refers to a dog with multiple heads that guards the gates to Hell. The security protocol Kerberos was developed at the Massachusetts Institute of Technology. Kerberos is used for authenticating users and restricting access to services only to authorized users when they access services from workstations, typically on a LAN. Services supported by Kerberos are said to be Kerberized. Kerberized versions of services like Telnet, remote file copy, and network file systems are available.

At a very high level, Kerberos works as follows. An *authentication server* shares a password with all users and a key with a ticket-granting server. When a user logs on to a workstation, the workstation contacts the authentication server. The authentication server issues a ticket to the user and also sends a key that the user will share with the ticket-granting server. This key is encrypted with the user's password and, therefore, can be recovered only by the legitimate user. The workstation will not be able to retrieve the key if the user is not legitimate. Thus, recovery of the key to be shared with the ticket-granting server indirectly authenticates the user to the system without the user's password being revealed on the medium. Note that, in this phase, a key has been transported to the user as well. Of course, this assumes that a password has been manually shared between the user and the authentication server.

The ticket itself is encrypted with a key shared between the authentication server and the ticket-granting server. It includes, among other things, the key that has been transported to the user, time stamps that indicate the lifetime of the ticket, and the address of the host where the user has logged on. When the user desires to access a service, the workstation presents the ticket to the ticket-granting server along with a MAC created using the key that was initially received from the authentication server. This verifies the user's legitimacy to the ticket-granting server without any need for using the user's password. The ticket-granting server then issues a key and a ticket to the workstation for use with the requested service. A similar authentication mechanism is used with the server providing the service.

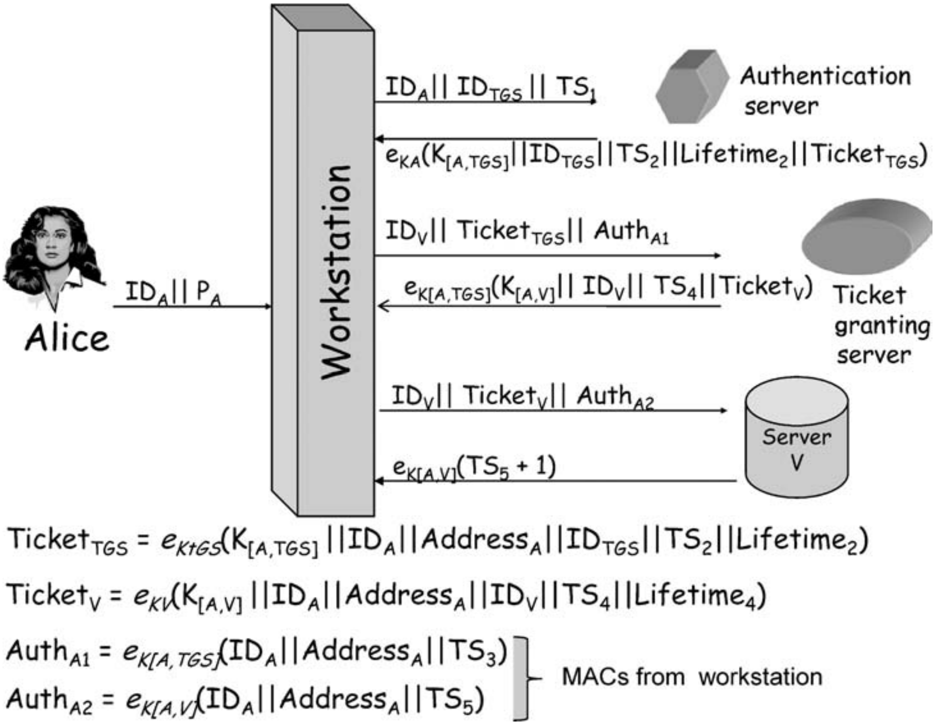

Authentication server
Alice
$ID_A || P_A$
Workstation
$ID_A || ID_{TGS} || TS_1$
$e_{KA}(K_{[A,TGS]} || ID_{TGS} || TS_2 || Lifetime_2 || Ticket_{TGS})$
$ID_V || Ticket_{TGS} || Auth_{A1}$
$e_{K[A,TGS]}(K_{[A,V]} || ID_V || TS_4 || Ticket_V)$
Ticket granting server
$ID_V || Ticket_V || Auth_{A2}$
Server V
$e_{K[A,V]}(TS_5 + 1)$
$Ticket_{TGS} = e_{KtGS}(K_{[A,TGS]} || ID_A || Address_A || ID_{TGS} || TS_2 || Lifetime_2)$
$Ticket_V = e_{KV}(K_{[A,V]} || ID_A || Address_A || ID_V || TS_4 || Lifetime_4)$
$Auth_{A1} = e_{K[A,TGS]}(ID_A || Address_A || TS_3)$
$Auth_{A2} = e_{K[A,V]}(ID_A || Address_A || TS_5)$
MACs from workstation

FIGURE 11.13 Message exchange in Kerberos.

Figure 11.13 shows a summary of the messages exchanged in Kerberos. In this figure, ID refers to the identity, P to the password, K to a key with its subscript indicating the entities sharing the key (if there is a single subscript, such a key is shared by the corresponding entity and the authentication server or ticket-granting server), "e" to encryption, with its subscript being the key used for encryption, and TS to a timestamp. The use of timestamps and lifetimes makes it necessary for hosts on the network employing Kerberos to have at least loosely synchronized clocks.

Kerberos is more complicated than what has been described here. Versions in current use are versions 4 and 5, with version 5 supporting multiple administrative domains (called realms). Version 5 supports a variety of encryption schemes and employs the standard CBC mode for encryption. Note that Kerberos employs only secret-key encryption. Kerberos does not protect against password dictionary attacks, nor does it set the access control privileges for accessing services. More details of Kerberos are available in [Sta03, Kau02].

IPSec. IPSec encrypts all IP traffic between two hosts, or two networks, or combinations of hosts with possibly different terminating points for different security services. By performing encryption and authentication at the IP layer, no applications need be modified and security is transparent to applications and the user with IPSec. However, this approach makes it more difficult to authenticate users and provide them privileges appropriately. Keys may be manually established or a very complex protocol called *Internet key exchange*

(IKE) can be used for authenticating entities to one another and establish keys. Keys are established as part of a unidirectional "security association" that specifies the destination IP address, keys, encryption algorithms, and "protocol" to be used. Mechanisms are included to prevent DoS attacks similar to TCP SYN floods, where one side of the security association would otherwise waste resources for key establishment.

"Protocol" in the above paragraph corresponds to one of two specific security protocols provided by IPSec – *authentication header* (AH) and *encapsulated security payload* (ESP). In AH, a MAC is created on the entire IP packet minus the fields in the IP header that change in transit. This enables the receiver to detect spoofed or modified IP packets. However, the payload is in plaintext and visible to anyone who may be capable of capturing the IP packet. ESP provides confidentiality and integrity to the payload of the IP packet, but not the header.

Use of the two protocols in a simple manner where a security association is set up between two hosts that directly apply either AH or ESP is called "transport mode" in IPSec. It is also possible to use a "tunnel mode," where the original IP packet is tunneled in another IP packet with a new IP header. This makes the original IP packet the payload, thereby protecting it completely with either a MAC in the case of AH or encryption and integrity in the case of ESP. Multiple security associations called "bundles" are also possible where both AH and ESP can be applied to the same IP packet. A security association database and a security policy database are used to decide what to do with an IP packet that is inbound or outbound from a host. IPSec is fairly complicated in all its details. We refer the interested reader to Kaufmann *et al.* [Kau02] for more information.

Secure Sockets Layer. The SSL (the latest version is called transport-layer security or TLS) is used in web browsers to secure data transfer, especially for e-commerce applications, banking, and other confidential transactions. At a high level, the browser is not required to be authenticated by the server (although this is possible and optional in SSL). In such a case, the user employing the web browser is authenticated using passwords or other techniques proprietary to the organization using the server.

The server, however, is authenticated by the browser through its digital certificate. Digital certificates, mentioned previously, are essentially public keys associated with the server authenticated by a trusted third party. The public key is bound to the e-mail address, web URL, or some other identification of the owner of the public key through a digital signature of the trusted third party. The certificate also specifies other kinds of information, such as the issuer, the algorithms for which the public key may be used, and so on. A schematic of a digital certificate is shown in Fig. 11.14.

Many trusted third parties that issue certificates are possible (sometimes, this is called the oligarchy model of a public-key infrastructure) leading to potential security issues (discussed in the next section). These trusted third parties, sometimes called trust anchors, are preconfigured in most modern browsers. Browsers also allow users to install their own trust anchors.

The digital certificate authenticating the public key of the server provides the user some assurance that the transaction is taking place with a legitimate bank or e-commerce site. Note that simply use of SSL is not the assurance of authenticity of the server, since any site or any server could use SSL. It is the information contained in the digital certificate that authenticates the server.

Upon receiving the digital certificate from the server, the browser creates a random secret, encrypts it with the server's public key and sends it to the server. This random secret, along with previously exchanged nonces are used to generate session keys (at both the server

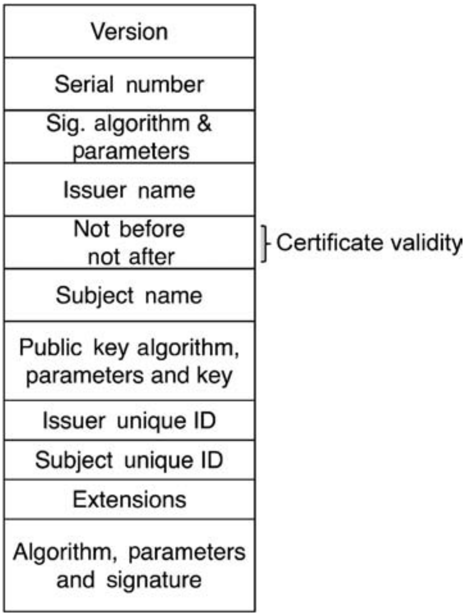


FIGURE 11.14 Schematic of a digital certificate.

and the browser) that are used for encryption with block or stream ciphers (RC4 is commonly used) and integrity with MACs. Figure 11.15 shows a high-level view of the messages exchanged between the client browser and server [Kau02].

The actual process involves four different protocols. The SSL handshake protocol is used to establish the session keys at the client and the server. The SSL record protocol carries all SSL messages as well as HTTP messages in SSL records. SSL records are units (like segments or packets) that are encrypted and authenticated eventually, and support compression as well. The change cipher spec protocol in SSL indicates the successful completion of the SSL handshake and the SSL alert protocol is used to indicate problems

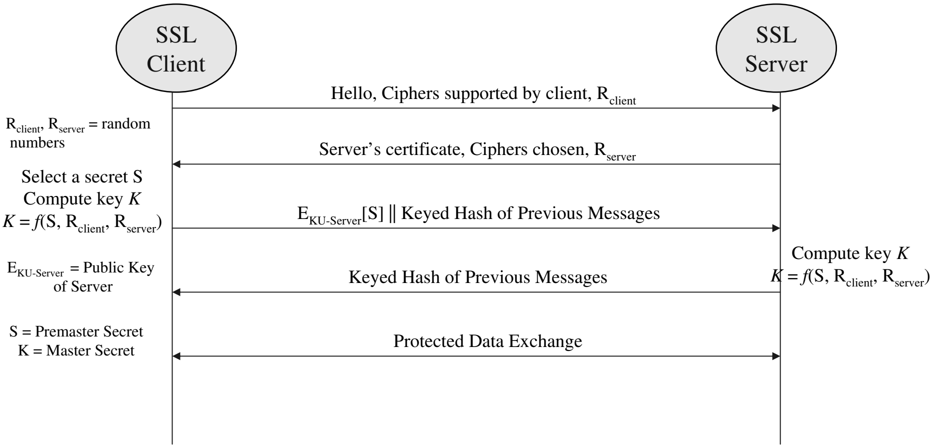

SSL Client
SSL Server
Hello, Ciphers supported by client, R_{client}
R_{client}, R_{server} = random numbers
Server's certificate, Ciphers chosen, R_{server}
Select a secret S
Compute key K
$K = f(S, R_{client}, R_{server})$
$E_{KU\text{-}Server}$[S] || Keyed Hash of Previous Messages
$E_{KU\text{-}Server}$ = Public Key of Server
Keyed Hash of Previous Messages
Compute key K
$K = f(S, R_{client}, R_{server})$
S = Premaster Secret
K = Master Secret
Protected Data Exchange

FIGURE 11.15 High-level view of message exchange and operation of SSL.

during the SSL message exchanges. More details of the message formats and actual message exchanges can be found in Kaufmann *et al.* [Kau02].

11.3.3 Preventing Successful Phishing Attacks

In the previous section we discussed the SSL, which is primarily used as the security mechanism on the web for e-commerce, banking, and other financial operations. In fact, SSL is implicitly used for any kind of secure communications on the web. However, there are potential dangers in the way SSL is employed. Most browsers show a yellow lock that is unlocked or locked depending on whether or not SSL is being used to communicate with the specific site. A lock does not automatically mean that communications are secure and trustworthy because it is possible for phishing sites to deploy SSL as well. As previously discussed, all a site needs to deploy SSL is a digital certificate issued by a trust anchor that is embedded in common browsers. Even if the browser pops up a warning saying that the certificate has not been signed by a valid trust anchor, gullible users may simply accept the certificate and proceed. It is believed that up to 5% of people receiving phishing e-mails fall prey to the attacks.

The APWG [APWG] along with the Financial Services Technology Consortium (FSTC) is heading initiatives to overcome this problem and increase trust in using the web for financial services and e-commerce. Some methods that have been proposed at large include the use of another third-party trust site to check if a URL is legitimate. This trusted third party will prove a site is trustworthy in a more visible fashion. For example, sites that want to be trusted must include an image tag in their web pages that directs the page to the third-party's site. The third party looks at the calling page's URL, verifies if it is legitimate, and dynamically generates an image using a servlet on its own site. Recently, newer browsers are adopting techniques that warn users if a site is suspected to be a phishing site, e.g. with a pop-up message or by showing the URL bar in a red color. A green-colored URL bar indicates a safe site that has a legitimate" long-term" digital certificate. The use of safe domains has also been suggested as an alternative for preventing phishing attacks.

11.4 DETECTION

Irrespective of the protection and prevention mechanisms in place, it is possible that security attacks succeed and proceed in an organizations network. It is extremely important to detect such attacks at the earliest possible time, so that action can be taken to stop further damage. More details of detection mechanisms and processes can be found in [Nor01,Bej04,Amo99].

Detecting who is the originator of security attacks is not trivial, especially when facing an intelligent adversary. As an example, let us consider DDoS attacks. Sources of such attacks are hard to detect because of a lack of detailed attack information, such as analyses of specific attacks or data related to frequency, distribution, number of agents, and effectiveness of response. There are no benchmarks for DDoS attacks, and they are hard to create because of the difficulty of large-scale testing of DDoS attacks. It is possible to determine the fact that an attack is ongoing against a system or network using IDSs.

Intrusion detection is the broad term used to describe the process for identifying the fact that a security attack has occurred (or is occurring). There is no single method for identifying attacks. Typically, three methods are used. In host-based intrusion detection, audit trails, logs, deployment of suspicious code, logins, etc. are monitored to detect the occurrence of a security attack. In network-based intrusion detection, the packets entering the network are examined to see whether they correspond to signatures of known security attacks. Anomaly-based intrusion detection looks for abnormal usage of network or system resources and flags potential problems.

Audit trail processing, used with host-based intrusion detection, is usually done offline. Care has to be taken to ensure that logs in hosts have not been tampered with. Logs from many hosts and systems may have to be correlated to detect attacks. Network-based intrusion detection is in real time as packets are captured. This can be problematic if the amount of data flowing into the network is humongous, as the buffering capacity may be limited and packets may be dropped by an IDS. Using signatures of known attacks is a common technique used for intrusion detection. However, this may miss new, unidentified attacks. If signatures are made too specific, then security attacks may be missed, resulting in false negatives. If signatures are made too general, then it is likely that some normal traffic and activity is flagged as a security attack, resulting in false positives. Thus, careful tuning is often necessary to detect intrusions with low false positives or negatives. The algorithms used for intrusion detection can be fairly complex, making use of data mining, pattern matching, decision making, etc.

Often, IDSs deploy *sensors* to probe or monitor the network or systems in question. It is necessary to deploy sensors on either side of a firewall to get an idea of the attacks that are being blocked. Multiple redundant sensors may be necessary, depending on the network topology. Sensors themselves may have to be networked to correlate the collected data. Such a network may or may not be separate from the network that is being monitored. The Internet Engineering Task Force is working on formats for exchange of intrusion detection information.

It is possible that IDSs may themselves be subject to security attacks. There are techniques that Oscar may employ to thwart detection by IDSs (such as fragmentation, flooding, launching unrelated attacks for distraction, etc.). Recent trends in intrusion detection include *distributed intrusion detection*, where system administrators from all over the world submit their monitored information to a service that then performs correlations to detect and identify attacks.

There are several kinds of IDSs available today, including specialized appliances from vendors. SNORT is an open source IDS that is available for free. While evaluating an IDS, it is necessary to consider the types of attack that an IDS can detect, the operating systems it supports, whether it can handle huge amounts of traffic, if it is capable of displaying large amounts of data in an easily understandable manner, the management framework that it provides, and its complexity.

Today, combinations of IDSs and firewalls, called intrusion prevention systems (IPSs), are also available. Rate-based IPSs block traffic flows if they are seen to exceed normal rates. Signature-based IPSs block traffic when signatures of known security attacks are detected.

Honeypots or Internet traps are systems used to detect and divert security attacks. Such systems look like real resources, perhaps with known vulnerabilities. Their value lies in the fact that Oscar may probe them, launch attacks against them, and perhaps compromise some of the systems. Monitoring Oscar's activities using honeypots can help detect other attacks against real systems or design methods of prevention.

11.5 ASSESSMENT AND RESPONSE

It is important to *assess* the security of the network and systems in an organization periodically. Additionally, assessment becomes important after a security incident has been detected and a *response* to the attack has been put in place. In this section, we briefly consider elements of assessment and response. More details can be found in the literature [Whi05, Nor05, McN04].

Assessment of a network can be done using external auditors who can perform penetration tests (act essentially like Oscar, but not damage systems), enumerate the entities in the network, discover potential vulnerabilities, and verify whether the protection and prevention mechanisms (like firewalls, access control schemes, password management) are working as they are supposed to. Vulnerability assessment tries to identify the presence of known vulnerabilities that can be and must be patched if patches are available. Since vulnerabilities are often operating system specific, vulnerability scanners may not pick up all vulnerabilities present on hosts in a network. Nessus is a popular open source vulnerability scanner. Commercial options also exist.

Responding to security attacks when detected is also an important aspect of security. The person in charge of the network needs to be immediately notified if an attack is detected (possibly through redundant means of communication). The security incident must be documented clearly. There must be processes in place to contact vendors and other external help if necessary. Actions to mitigate the impact of the security attack must be taken, followed by eradication of the vulnerability that caused the attack. An assessment of reasons as to why the attack was successful and steps to prevent recurrence must be taken.

Again, as an example, let us consider DDoS attacks. There are several proposed techniques for mitigation of such attacks. If there is a possibility of cooperation across organizations and ISPs, then ingress filtering at the edge of networks is possible. It is here that it is easiest to identify if packets are using spoofed IP addresses and block them from entering other parts of the Internet. The attacked system can also improve its resilience to DoS by increasing data processing speeds at its servers (e.g. by adding additional hardware) and increasing its link capacity to handle both the normal load and attack packets. Finally, if possible, the criminal justice system should be used to hunt and shut down attackers.

We mention two specific methods: one for identifying the source systems launching attacks and the other for analyzing (assessing) attack traffic. *IP traceback* is an approach that tries to determine where the DoS-related packets are coming from (especially when they have spoofed IP source addresses). There are several ways of tracing the sources. One method is to query routers actively and use a developed signature of the attack to proceed hop by hop to see where the packets actually originated. Another approach is to create a virtual overlay network for selectively monitoring flows and logging packets. Such logs can be used to identify the path taken by DoS-related packets by reconstruction using probabilistic "packet marking." The use of *backscatter* for analyzing DoS attacks has been tried by the research community. This process is useful in terms of evaluating how prevalent DoS attacks are. The idea behind backscatter analysis is as follows. Zombie agents that spoof IP addresses are likely to employ some unused address space. If hosts can be deployed at these IP addresses, with the sole aim of capturing packets that are sent in response to spoofed DoS packets, then the number, type, and frequency of received packets can provide information about DoS attacks. The hope is that, if Oscar chooses IP addresses at random and packets are captured in a sufficiently large address space, they can provide a good sample for analysis. Researchers at AT&T and at the San Diego Supercomputer Center have performed backscatter analysis [Ches03].

QUESTIONS

1. Differentiate between local name servers, authoritative name servers, and root name servers.
2. Name three security services.
3. What is social engineering?
4. How is a vulnerability defined by IBM's Internet Security Systems task force report?
5. How can search engines be used for malicious purposes?
6. What is non-repudiation? Can it be provided using only secret key cryptography?
7. What is the difference between a block cipher and a stream cipher?
8. What is tempest?
9. Differentiate between active and passive attacks.
10. What is a smurf attack?
11. Name a worm that did not use buffer overflows as the primary method of infecting a host.
12. Describe ways in which a worm propagates from a host to other hosts on the Internet.
13. What is a Botnet? Are members of a Botnet always active?
14. What is an evil-twin? How is it different from other phishing attacks?
15. Describe three attacks that can help a malicious entity steal passwords.
16. Describe how the ability to predict sequence numbers can be abused for attacks against TCP and attacks against DNS.
17. How is a packet filter different from a stateful firewall?
18. Differentiate between reflexive access control lists and standard access control lists. Why would you prefer one to the other?
19. How are public-key and secret key algorithms different?
20. Explain the importance of key sizes in the security of an encryption algorithm.
21. What is entity authentication? How is it different from message authentication?
22. What two methods are used to establish keys between two communicating parties?
23. What is a challenge-response scheme?
24. Explain how Kerberos works and how it is used to provide security in a local area network.
25. What is the difference between AH and ESP in IPSec?
26. What is the potential weakness of SSL?
27. Name two techniques being used to prevent phishing attacks.
28. What is intrusion detection?
29. Why is assessment of security important?
30. How should a response to a security attack be undertaken?
31. Differentiate between IP Traceback and backscatter.

PROBLEMS

Problem 1:

Consider a situation where a host X is connecting to a host Y using TCP. The initial sequence number of the SYN packet sent by X is 1789.

(a) What information in the three-way handshake can you predict with this information?

(b) What information can you NOT predict?

(c) If you are not host Y, can you fool host X into thinking that it has connected with host Y? How? State all assumptions.

(d) If you are not host X, can you fool host Y into thinking that it has connected with host X? How? State all assumptions.

Problem 2:

Do an nslookup and dig on www.msn.com from paradox.sis.pitt.edu or from any other computer you may be using. Attach a copy of the results. Based on the outputs, (a) identify the nameserver that is being used to determine the IP address (b) the canonical (real name) of www.msn.com and (c) the number of servers that correspond to the name www.msn.com. Repeat nslookup after a few minutes. Does the IP address change?

Problem 3:

A packet filter has two interfaces *e1* that accepts packets from outside a company's network and an interface *e2* that accepts packets from inside the same company's network. The company's IP address space is 136.142.117.0/24. The web server has an IP address 136.142.117.41.

(a) A packet with a source IP address 136.142.117.132 arrives at *e1*. Should it be accepted or dropped? Why?

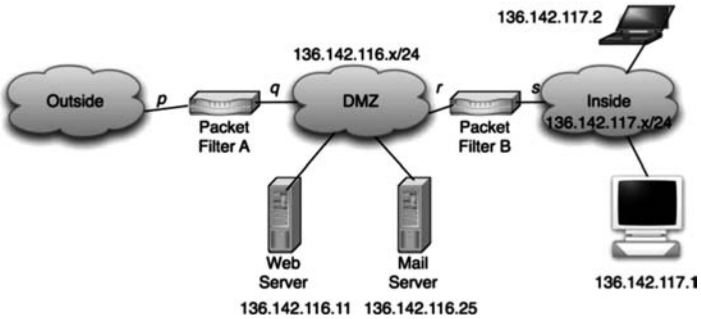
(b) A packet with source address 130.16.11.3 arrives at *e1* destined for 136.142.117.132 with port number 80. What should the packet filter do? Why?

If a packet with source IP address 10.0.1.1 arrives at either interface, what should the packet filter do? Why?

Problem 4:

Let us suppose that the packet filter in problem 3 is a dynamic packet filter. It receives a packet with source IP address 136.142.117.132 and source port number 2111 with destination IP address 64.236.24.4 and destination port number 80 at *e2*. The SYN flag is set in the TCP segment. What rule is automatically created for packets entering *e1*? Why?

Problem 5:



Consider the network shown above. Design rules for packets *entering* interfaces *p*, *q*, *r*, and *s* for static packet filters A and B for the following security policy: Clients from the outside can only connect to the web server and mail server in the DMZ. Most hosts from the inside can connect only to the web server and mail server in the DMZ. Only the host 136.142.117.1 can connect to any web server on the outside. You can use a mix of standard and extended ACLs. Include details and add explanations as necessary.

Problem 6:

A not-so-rich hacker uses an old computer and brute force to break into some wireless systems. It takes him 1 ms on average to test a key to see if it is the right one for an encryption independent of the algorithm employed. How long will it take him to break into an IEEE 802.11 system in the worst case? How long will it take him to break into an IS-136 system on average? Assume that the former uses 128 bit keys and the latter 64 bit keys. Also assume that the encryption scheme and crypto protocols are otherwise secure.

Problem 7:

In Problem 6, the hacker realizes that the last six bits of the keys used in a private 802.11 LAN are always zeros. In what time can he break into the system in the worst case?

Problem 8:

In Problem 6, the hacker manages to (a) buy a second old computer that can test a key in 1.5 ms. With the two computers, in what time can he break into the system in the worst case? (b) upgrades his computer so that he can test a key in 1 microsecond. In what time can he break into the system in the worst case?

PROJECTS

Project 1:

Go to the following URLs. Summarize the attacks described there in your own words. Classify the attack(s) into the categories described in this chapter.

(a) http://www.kb.cert.org/vuls/id/222750

(b) http://www.kb.cert.org/vuls/id/637934

Go to the site http://www.cert.org/nav/index_red.html. Search for incident note IN-2002–03 and read it. Summarize the attack(s) listed there in your own words.

12

WIRELESS LOCALIZATION

12.1 INTRODUCTION

Geolocation, position location, localization, and radiolocation are terms that are widely used today to indicate the ability to determine the location of an MS in different environments. Location usually implies the coordinates of the MS, which may be in two or three dimensions, and usually includes information such as the latitude and longitude where the mobile terminal is located. In indoor areas and within buildings, alternative coordinates and visualization techniques may be employed to indicate the location of an MS. Geolocation technologies are gaining prominence in the wireless market for several reasons, primarily the US FCC mandate requiring all wireless cellular carriers to be able to provide the location of emergency 911 callers to a *public safety answering point* (PSAP). However, geolocation technology has proved to be significant for both military and commercial applications in general beyond emergency location. The

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

use of wireless devices such as cell phones, PDAs, and laptops has become the enabler of viable location-based services and applications that need position location information [Bar03,Dru01,War03]. Examples of commercial location-based services include locating patients and equipment in a timely fashion in hospitals, locating children and pets for personal and residential applications, and concierge and location-aware services (e.g. locating the nearest coffee shop or providing information about exhibits in museums based on the customer's location). The monetary benefits of such location-based services for service providers are expected to increase in the next few years. In the military and public sectors, enabling soldiers, policemen, and fire fighters with knowledge of their location and the location of other personnel, victims, exits, dangers, the enemy, etc. proves to be invaluable. The GPS has been the most successful positioning technique in outdoor areas and we now see the GPS receiver as an inexpensive commonplace gadget. While GPS has been hugely successful, it also has several drawbacks for use in the applications that we have discussed so far, especially in indoor areas. In this chapter, we discuss position location issues in today's wireless networks, alternative technologies that are being investigated and standardized for position location in outdoor and indoor areas, and trends in this area.

The position location information is usually specified with a certain accuracy and precision. Accuracy refers to the error in distance from the determined position and the actual position. Precision refers to the fraction of time that the error is smaller than a given value. Any system that locates an MS needs to *sense* some characteristics related to the MS to determine its location. The position is determined using some algorithms that use these sensed characteristics as input. In addition, there are protocols that are required to transport the sensed information to entities that determine the location or provide other services. In this chapter, we will describe these aspects related to locating the position of an MS. In Section 12.2 we consider example applications and regulatory issues related to position location. In Section 12.3, the sensing process and the location algorithms are discussed. These influence (a) where the position location is determined (at the MS or in the wireless network), (b) the number and nature of reference locations necessary, and (c) the accuracy and precision of the estimated location. This section will also consider the standard specifications for cellular telephone systems that make use of some of these algorithms. Section 12.4 describes the location services (LCS) architecture specified for use in cellular networks. Section 12.5 provides a brief overview of localization in ad hoc and sensor networks.

12.2 WHAT IS WIRELESS GEOLOCATION?

The term "location-based service" is used to denote services provided to mobile users based on their geographic location, position, or known presence. These are primarily based on a geolocation infrastructure and system put in place to obtain location information of users. As mentioned in the introduction, positioning systems have found a variety of applications both in the civilian and military environments. There are numerous such applications that are already available today, such as mapping services (which provide driving directions or locations of businesses in an area), information services (which provide local news, weather, traffic, etc.), and concierge services (for making dinner reservations, movie tickets, directory services, etc.). Commercially, content, advertising, and personalization services that are location dependent are being deployed today. Each of these applications requires accuracy and precision of position that is specific to its needs. In indoor areas, for applications such as inventory and asset management (such as locating wheelchairs in hospitals), the accuracy has to be within a few meters on a given floor of the building. For enhanced-911, or E-911, emergency response, the required accuracy is much lower, as described later. We discuss example indoor and outdoor applications below that are becoming increasingly important.

Indoor geolocation applications traditionally have been directed towards locating people and assets within buildings. Finding mentally impaired patients in hospitals and portable equipment such as projectors, wheelchairs, etc. that are often moved and never returned to a traceable location are two common examples. The so-called *personal locator services* [Kos00], which could also operate outdoors, employ a *locator device* that resides with a person whose location is to be determined. There are two possible scenarios. In the first case, someone requests a service to provide them with the location of the individual and appropriate steps are taken to determine the person's location. In the second case, the person is lost or in some other dire straits and can employ a panic button to request help. Here, the locator service will determine the location and provide the requested assistance. For locating equipment or assets, only the former scenario applies.

Existing communications and computing environments, both in residential areas and offices, typically have been statically configured, making the task of reconfiguration extremely complex and cumbersome, requiring manual intervention. To overcome this inconvenience, smart spaces and smart office environments are being considered for deployment that can automatically change their functionality depending on the context [War97]. Such *context-aware* networks are based on awareness of who or what is present around them. With location awareness, computing devices ranging from small PDAs to desktops and Internet appliances could personalize and adapt themselves to their current set of users, each requiring their own services from the smart environment. For this purpose, not only should the smart space be aware of who is present, it should also be aware of where the user is located and whether there are other mobile devices in the vicinity. For instance, a handheld computer should be able to automatically determine the closest printer to print a document in an office environment. Such nontraditional applications also demand geolocation services.

There are several outdoor geolocation applications, the most common of which is simply the application of locating ones own self using GPS while traveling on the road. Information technology has increased the number of applications far beyond this simple self-location application. The term *telematics* is used to imply the convergence of telecommunications and information processing, and it has evolved to refer to automobile systems that combine the GPS location mechanism with wireless communications for services such as automatic roadside assistance, remote diagnostics, and content delivery (information and entertainment) to the automobile. A good example of such a system is General Motors' OnStar system [OnStar]. *Intelligent transportation systems* refer to the ability to autonomously navigate vehicles while making use of the latest traffic information, road conditions, travel duration, etc. This includes fleet management as well as automatic steering of vehicles. In order to obtain relevant information from service providers or servers across a network or the Internet, the vehicles should be able to provide their location and destination information. Alternatively, the service provider should be able to determine the vehicle's location.

12.2.1 Wireless Emergency Services

In this section we will provide an introduction to the requirements of location-based services and E-911 in wireless environments. The requirements are primarily in terms of performance metrics of interest, such as accuracy and precision. We will also describe some of the regulatory aspects related to the provisioning of location information and performance metrics. We provide a brief summary of different cellular standards that we refer to later in this chapter. Cellular systems are considered in more detail in Chapter 7.

Wireless E-911 services, by far, have proved to be the biggest catalyst for investment and development of geolocation technology suitable for cellular communications. A caller on a wired telephone to an E-911 service is immediately located because the location of the fixed telephone is known with an accuracy of within a couple of rooms in a building. If the same caller is on a mobile telephone, then the technology that can obtain the location is more complex. In the simplest case, the only knowledge that is available is that the caller was connected to a particular BS.

Cellular service providers may use the position information for network-related issues such as handoff or location management [3GPP 25.305, Dra98]. However, much of the accuracy and precision levels for mobile phones are being driven by the values mandated by the FCC for E-911 public safety applications in the USA. The FCC is responsible for regulatory issues related to telecommunications services in the USA. The E-911 service provides emergency assistance to callers through a PSAP. The FCC has specified accuracy and precision values for E-911 calls. The service provider must be able to locate the caller within an accuracy of 100 m at least 67% of the time and within 300 m at least 95% of the time and provide this information to the PSAPs. The reader is referred to [Mey96, FCC E-911, Ree98] for more details. In Europe, emergency services are referred to as E-112 services and regulatory authorities are in the process of specifying accuracy and precision levels for E-112.

Cellular systems use many different physical layer and networking standards, making position location different in different systems. The 1G cellular systems use analog modulation schemes and do not support any position LCS. The 2G systems use digital modulation and two primary multiple access technologies: TDMA in the GSM-based systems and CDMA in IS-95 or cdmaOne systems. The 3G systems are based entirely on CDMA, but again have two primary standards: cdma2000 and UMTS. UMTS is the successor of GSM and cdma2000 is the successor of IS-95. All of the 2G and 3G standards are expected to satisfy the FCC mandate. While the positioning schemes in all of them have some common features, the standards and terms are somewhat different, as discussed in Section 12.3.3.

There are no mandated positioning requirements for WLANs. Most deployed WLAN equipment follows the IEEE 802.11 standard or its enhancements like 802.11a, b, or g. There are proprietary solutions for positioning with 802.11 WLANs, which primarily use RF signatures or variations thereof described below. The accuracy with such systems ranges from a few meters to tens of meters with varying levels of precision depending on the environment (building material, architecture, campus area, and so on).

12.2.2 Performance Measures for Geolocation Systems

In this section we consider performance benchmarking of geolocation systems, primarily employing the work and discussion in Tekinay *et al.* [Tek98]. Wireless systems have traditionally focused on telecommunications performance issues, such as QoS, grade of service, BERs, capacity, reliability, and coverage. For geolocation systems, some of these performance issues are still valid, although new performance benchmarks are necessary, as described below. Table 12.1 compares performance measures for telecommunications and geolocation systems based on Tekinay *et al.* [Tek98].

One of the most important performance measures of a geolocation system is the accuracy with which the location is determined. This is similar to the BER or packet error rate requirements in telecommunications systems. As in the case of BER, the actual benchmark values may be different depending on the application in question. For example, voice packets can tolerate a BER of 1%, but data packets need a BER of at least 10^{-6}. In the same way, outdoor position location applications demand a lower accuracy than indoor applications.

Location system accuracy is often defined as the area of uncertainty around the exact location, where a percentage of repeated location measurements are reported. For example, 67% of the measurements of the location of an MS lie within 50 m of the actual location or 95% of the measurements lie within 100 m of the actual location. This accuracy depends heavily on the radio propagation environment, receiver design, noise and interference characteristics, number of redundant measurements available for the same location, and the complexity of signal processing performed.

TABLE 12.1 Comparison of Performance Measures for Telecommunications and Geolocation Systems

Telecommunications systems	Geolocation systems
QoS	*Accuracy of service*
• Signal-to-interference ratio	• Percentage of location requests located to within an accuracy of δ meters
• BER	
• Packet error rate	• Distribution of distance error at a geolocation receiver
Grade of service	*Location availability*
• Call blocking probability	• Percentage of location requests not fulfilled
• Availability of resources	• Unacceptable uncertainty in location
• Unacceptable quality	
Coverage	*Coverage*
• Where communications are possible	• Where location information is available
Capacity	Capacity
• Subscriber density that can be handled	• Location requests/frequency that can be handled
Miscellaneous	*Miscellaneous*
• Delay in call setup	• Delay in location computation
• Reliability	• Reliability
• Database look-up time	• Database look-up time
• Management and complexity	• Management and complexity

In [Pah98, Kri98], the distribution of the distance error in indoor areas, which ultimately affects the area of uncertainty, is discussed in detail based on measurements and simulations.

The grade of service for telecommunication systems is usually the call-blocking rate during the peak hour. In a similar manner, the probability that a location request will not be fulfilled is a measure of the grade of service for a geolocation system. The location request will not be fulfilled if location-sensing measurements are not available in sufficient numbers or the measurements lead to unacceptable location accuracy.

Coverage in telecommunications systems is related to the service area where, at a bare minimum, access to the wireless network is possible. For geolocation systems, coverage corresponds to the availability of a sufficient numbers of measurements of a sensed characteristic to perform a location computation.

Finally, several other issues [Tek98] are also important in geolocation systems in manners similar to telecommunications systems: delay in triggering a location measurement, location algorithm calculation time, network transmission delay, database look-up time, end-to-end delay between the time a location request is made and the location information is received, and so on. Reliability, i.e. the mean time between failures and the mean time to repair, network management, and complexity are also important issues.

12.3 RADIO-FREQUENCY LOCATION SENSING AND POSITIONING METHODOLOGIES

In this section of the article we will describe *basic* positioning methodologies. In the following section we consider a generic geolocation system architecture and define different positioning approaches. Section 12.3.2 describes RF location-sensing approaches and algorithms. Section 12.3.3 includes a discussion of standards for LCS in cellular systems.

12.3.1 Generic Architecture

A functional architecture of a geolocation system is shown in Fig. 12.1*a* [Cha99]. The two essential functional ingredients for position location are the location estimation of the MS and sharing this information with appropriate attributes with some entity in the wireless network.

Geolocation systems *measure* or *sense* RF parameters of radio signals that travel from a mobile to a fixed set of receivers *or* from a fixed set of transmitters to a mobile receiver. There are thus two ways in which the actual estimate of the location of the MS can be obtained. In a *self-positioning system*, the MS locates its own position using measurements of its distance or direction from known locations of transmitters (for example, GPS receivers). In some cases, *dead reckoning*, a predictive method of estimating the position of the mobile by applying the course and distance traveled by an MS since a location was previously determined, could be employed. Self-positioning systems are often referred to as mobile-based or terminal-centric [Mey96] positioning systems. In *remote positioning systems*, receivers at known locations on a network together

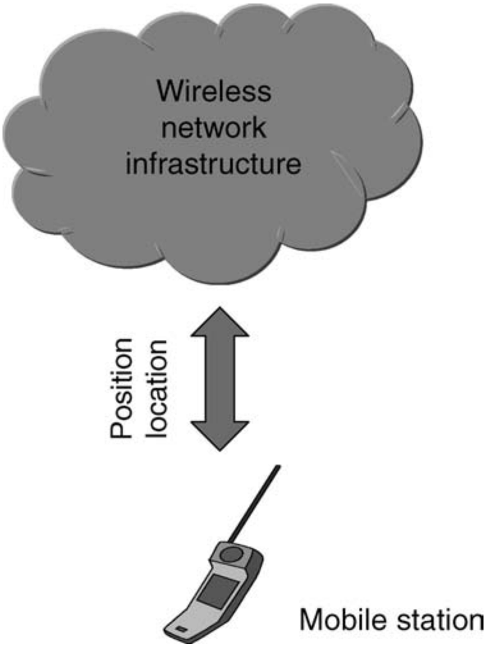


FIGURE 12.1 Functional architecture of geolocation system.

compute the location of a mobile transmitter using the measurements of the distance or direction of this mobile from each of the receivers [Dra98]. Remote positioning systems are also called network-based or network-centric [Mey96] positioning systems. Network-based positioning systems have the advantage that the MS can be implemented as a simple transceiver with small size and low power consumption for easy carrying or attachment to assets that need tracking as a simple and inexpensive tag. In addition, it is possible to have *indirect* remote- or self-positioning systems where the mobile may transmit information about its location to a location control center or the location control center transmits the location of the mobile to itself through an appropriate communications channel.

Example 12.1: Indirect Remote Positioning In indirect remote positioning, an E-911 PSAP requires the location information of a caller. If a mobile-based positioning system is used, the MS determines its own position either using GPS or signals from multiple BSs. This information has to be transmitted to the location control center by the mobile terminal through one of the BSs.

An example of a geolocation system architecture [Kos00] is shown in Fig. 12.2. A geolocation service provider provides location information and location-aware services to subscribers. Upon a request from a subscriber for location information about an MS, the service provider will contact a location control center querying it for the coordinates of the MS. This subscriber could be a commercial subscriber desiring to track a mobile device or a PSAP trying to answer an E-911 call. The location control

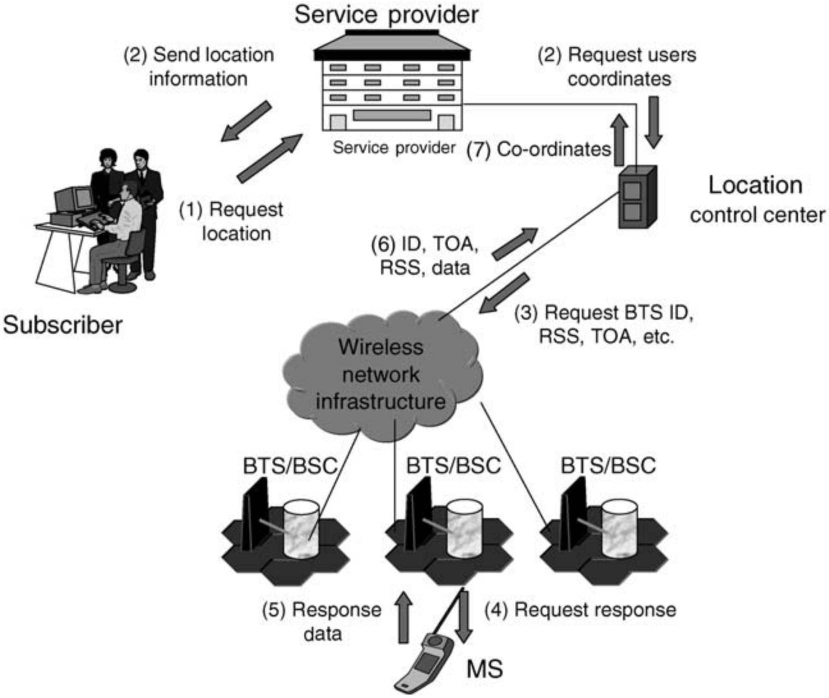


FIGURE 12.2 General architecture of a geolocation system.

center will gather information required to compute the MS's location. This information could be parameters such as RSS, BS ID, TOA of signals, etc. that we discuss later. Depending on past information about the MS, a set of BSs could be used to page the MS and directly or indirectly obtain the location parameters. These are sometimes called *geolocation* BSs (GBSs). Once this information is collected, the location control center can determine the location of the mobile with a certain accuracy and convey this information to the service provider. The service provider will then use this information to visually display the MS's location to the subscriber. Sometimes the subscriber could be the MS itself, in which case the messaging and architecture will be simplified, especially if the application involves self-positioning.

12.3.2 Positioning Algorithms

Positioning processes use *closeness* to the point of association (POA) of an MS with the network, *measures of distance* such as the TOA, TDOA, phase difference or RSS of signals, *measures of direction* such as the AOA or DOA of signals, a fingerprint or signature of signal characteristics at a location, or a combination of these to estimate the location of an MS [3GPP 25.305]. Figure 12.3 shows how some of these positioning methodologies operate. In the following discussion, we explain positioning approaches based on self-positioning or remote positioning. However, the alternative approach (or even indirect positioning) is also possible in similar ways.

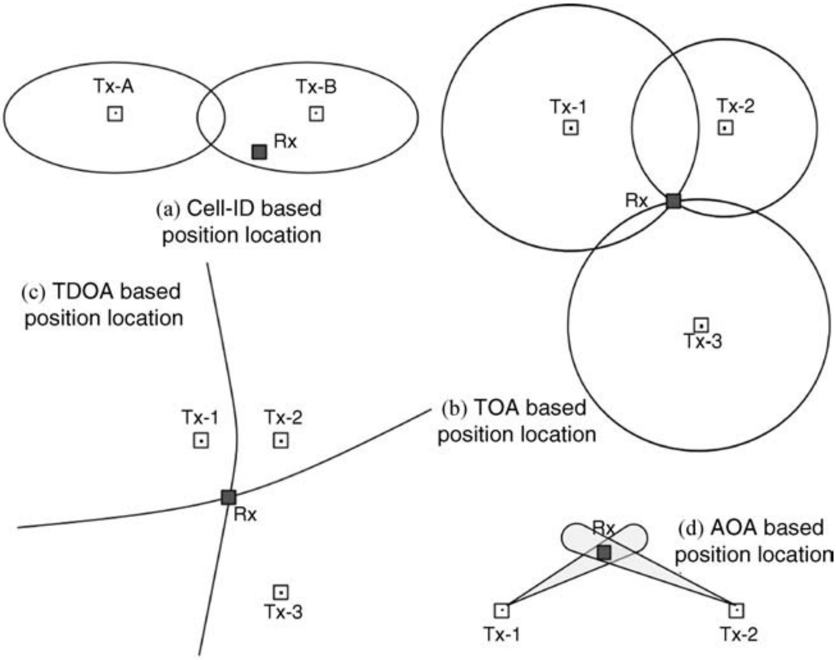

Tx-A
Tx-B
Rx
(a) Cell-ID based position location
Tx-1
Tx-2
Rx
Tx-3
(b) TOA based position location
(c) TDOA based position location
Tx-1
Tx-2
Rx
Tx-3
Rx
(d) AOA based position location
Tx-1
Tx-2

FIGURE 12.3 Location-sensing methodologies.

The location of a mobile terminal can be determined as follows. Consider, for example, a remote-positioning system where the GBSs are together determining the MS's position (a similar approach is applicable for self-positioning systems). It is possible to exploit the characteristics of radio signals transmitted by an MS to fixed receivers of known location to determine the location of the MS. The GBSs measure certain signal characteristics and make an estimate of the location of the MS based on the knowledge of their own location. The general problem can be stated as follows:

> The locations of N receivers (GBSs) are known via their coordinates (x_i, y_i) for $i = 1, 2, 3, \ldots, N$. We need to determine the location of the MS (x_m, y_m) using characteristics of the signals received by these transmitters.

In order to determine (x_m, y_m), traditionally, the distance or direction (or both) of the MS must be estimated by several of the GBSs from their received signals. Distances can be determined using the properties of the received signal, such as the signal strength, the signal phase, or the TOA. The direction of the MS can be determined from the angle of arrival of the received signal. Alternative approaches such as closeness of match of RF signatures or closeness to a POA have been used in recent times.

Closeness to the Point of Association. Most wireless networks use a fixed point of access to the network. These points of access could be BSs in cellular networks and APs in WLANs. BSs or APs have radio transceivers that provide a service over a specific geographical area called a "cell" (see Chapters 7 and 9). In the closeness to the point of association (POA) approach, the position of the MS is known to be somewhere within the cell. For example, in Fig. 12.3*a*, the receiver (Rx) is known to be in the coverage area (cell) of transmitter (Tx) A. Unfortunately, the size of a cell may be as small as 100 m in microcellular areas and as large as 15 km in macrocellular areas. Thus, the accuracy and precision can vary significantly depending on the density of deployment of BSs or APs and their coverage areas. In this case, the MS location (x_m, y_m) is associated with the location of the BS with coordinates (x_i, y_i).

Distance-based Techniques. If the distance to an MS is known from at least three distinct transmitters (whose locations are known), then it is possible for the mobile receiver to construct three overlapping circles, as shown in Fig. 12.3*b*. The point where the three circles intersect is the position of the receiver. The distances between transmitters and receivers are estimated using RSS, the TOA, or the TDOA of a transmitted signal.

Three measurements are required to estimate the position of the mobile in two dimensions and four measurements are required for estimating the position in three dimensions. In Fig. 12.3*b*, the need for three measurements for estimating the position in two dimensions is illustrated. If the distance between the receiver and the mobile is estimated to be *d*, then it is obvious that the mobile could be located on a circle of radius *d* centered on the receiver. A second measurement reduces the position ambiguity to the end points of the chord that is common to the two circles. The third measurement provides a fix on the location of the mobile.

TOA/TDOA: A transmitted signal travels 0.3 m/ns in air or free space and this property can be exploited to determine the distance between the transmitter and receiver. This is the TOA technique that is employed with some modifications in current GPS receivers [Kap96], as well as in certain E-911 location systems. When a GBS detects a signal, its absolute TOA is determined. If the time at which the MS transmitted the signal is known, then the difference in the two times will give an estimate of the time taken by the signal to arrive at the GBS from the MS. Errors in determining the TOA often make the intersection of the circles used to determine the location a region rather than a point. In such a case, the position-location estimation algorithm will pick some point within this region as the estimated position. Wireless systems that employ the TOA (or TDOA) technique employ pulse transmission, phase information or spread-spectrum techniques to form time estimates. For instance, the time difference between two signals received for either self-positioning or remote positioning can be estimated from their cross-correlation.

Example 12.2: Commercial Indoor Location Systems based on TOA Recently, a few commercial products for indoor geolocation have appeared in the market [Wer98]. The overall system architecture of these systems is shown in Fig. 12.4. These systems use simple-structured *tags* that can be attached to valuable assets or personnel badges. Indoor areas are divided into cells with each cell being served by a *cell controller*. The cell controller is connected to a number of antennas (Ref. 16 in [Wer98]) located at known positions. To locate the tag position, a cell controller transmits a 2.4 GHz spread-spectrum signal in the unlicensed ISM bands through different antennas in time-division multiplexed mode. Upon receiving signals from the cell controller antenna network, tags simply change the frequency of the received signal to another portion of the available

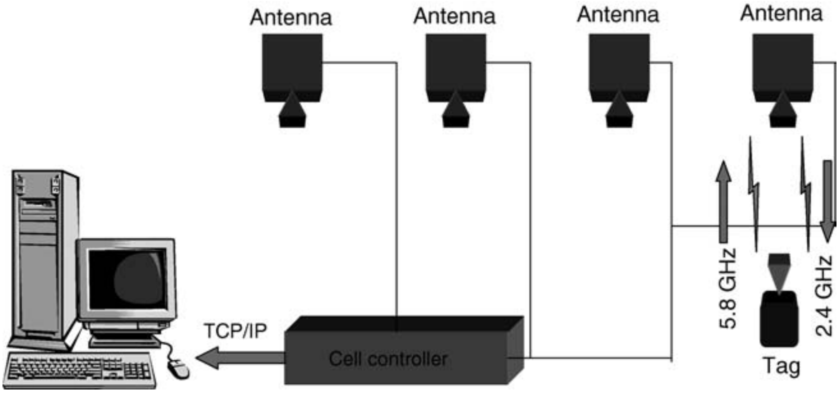


FIGURE 12.4 Local indoor positioning system.

unlicensed bands, either in 2.4 GHz or 5.8 GHz, and transmit the signal back to the cell controller with tag ID information phase-modulated onto the signal. The distance between the tag and antenna is determined by measuring the round-trip time of flight. With the measured distances from the tag to antennas, the tag position can be obtained using the TOA method. A host computer is connected to each cell controller, through a TCP/IP network or other means, to manage the location information of the tags. Since the cell controller generates the signal and measures round-trip time of flight, there is no need to synchronize the clocks of tags and antennas.

The multipath effect is one of the limiting factors for indoor geolocation (see Chapter 2 for a discussion). Without multipath signal components, the TOA can be easily determined from the autocorrelation function of the spread-spectrum signal. The autocorrelation is 2 chips wide and the time to rise from the noise floor to the peak is 1 chip. If the chipping rate were 1 MHz, then it would take 1000 ns to rise from the noise floor to the peak, providing a "ruler" with a thousand 30 cm increments. In this manner, a 40 MHz chipping rate, chosen for the PinPoint system, provides a ruler of 25 ns that provides real-world increments of about 3.8 m. Because of regulatory restrictions in the 2.4 and 5.8 GHz unlicensed bands, faster chipping rates are not easy to achieve, and signal-processing techniques must be used to improve the accuracy further. If different frequency bands are used for uplink and downlink communications, then the interference between the channels can be further isolated.

If the absolute times of arrival are unknown, then the TDOA technique, which uses time differences between pairs of transmitters, is preferred. In this case, the measured times from two transmitters are subtracted and the result is the intersection of two hyperbolas, as shown in Fig. 12.3*c*. In GPS, the TDOA technique is used where the differences in the times of arrival of signals from satellites are used to locate the mobile. The TDOA technique defines hyperbolas (rather than circles) on which the transmitter must be located with foci at the receivers. Three or more TDOA measurements provide a position fix at the intersection of hyperbolas. Exact solutions and Taylor series approximations are available [Com98] for solving these equations. Compared with the TOA method, the main advantage of the TDOA method is that it does not require knowledge of the transmit time from the transmitter. As a result, strict time synchronization between the MS and the GBSs is not required. However, the TDOA method requires time synchronization among all the receivers used for geolocation.

While geometric interpretation can be used to calculate the intersection of circles or hyperbolas, estimates have to be used when there are errors. Multipath propagation of signals (see Chapter 2) impacts errors in the time (difference) of arrival of signals, as signals may take more time to reach the receiver because of multiple reflections. A recursive least-squares estimate is used when there are errors in the distance measurements. Let the distance d_i from the ith GBS be determined from the absolute TOA of the signal it receives as $d_i = c\tau$, where c is the velocity of light and τ is the time taken by the signal to reach the GBS. If the location of the ith GBS is (x_i, y_i) and the location of the mobile is (x, y), then we have N equations (to be solved together) of the form

$$f_i(x,y) = (x-x_i)^2 + (y-y_i)^2 - d_i^2 = 0 \tag{12.1}$$

for $i = 1, 2, \ldots, N$. As a geolocation problem, extensive research has been done to improve upon the accuracy of algorithms that are used to estimate the position of a mobile. In particular, when N is more than three or four (thus providing redundancy in the measurements), information in the redundant measurements can be used to reduce errors that are introduced by noise, environment, multipath, etc. [Kap96]. Figure 12.5 shows an example of using the recursive least-squares technique to arrive at the location of the MS.

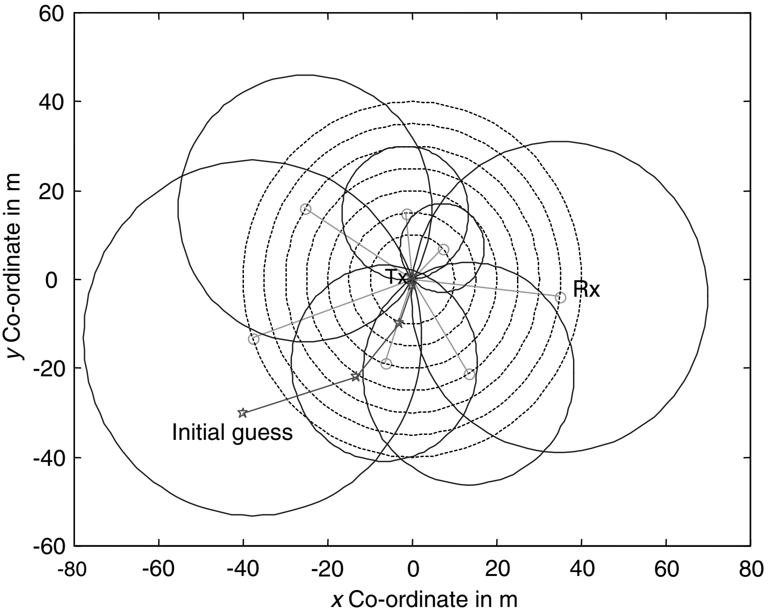


FIGURE 12.5 Recursive least squares to determine the MS location using measurements at seven GBSs.

RSS: If the transmitted power at the MS is known, then measuring the RSS at the GBS can provide an estimate of the distance between the transmitter and the receiver using known mathematical models for radio signal path loss, which depends on distance (see Chapter 2). As with the TOA method, the measured distance will determine a circle, centered on the receiver, on which the mobile transmitter must lie. This technique results in a low-complexity receiver for a self-positioning system. This method, however, is very unreliable because of the wide variety in path-loss models and the large standard deviations in the errors associated with these models due to shadow fading effects. Receivers do not distinguish between signal strength in the LOS path and in reflected paths [Mey96]. Especially indoors, the power–distance gradients can vary anywhere between 15 and 20 dB/decade to as high as 70 dB/decade. Also, these gradients and other parameters employed in path-loss models are site specific. As a result, this technique cannot be employed in situations where the required accuracy is a few meters. The accuracy of this method can be improved by utilizing premeasured RSS contours centered at the receiver [Fig69] and multiple measurements at several BSs [Kos00]. A fuzzy logic algorithm was shown in [Son94] to be able to significantly improve the location accuracy.

Example 12.3: A Commercial Geolocation System based on RSS The infrastructure of Paltrack indoor geolocation system [PalTr], developed by Sovereign Technologies Corp., consists of tags, antennas, cell controllers, and an administrative software server system. The PalTrack system utilizes a network structure that resides on an RS-485 node platform. A network of transceivers is located at known positions within the serving area and the transmitter tags are attached to assets. The tag transmitters transmit a unique identification code at 418 MHz to a network of transceivers when in motion or at predefined time intervals. Transceivers estimate the tag location by measuring the RSS and utilizing a robust RSS-based algorithm patented by Sovereign Technologies Corp. The master transceiver collects measured information from the transceivers and relays it to a PC-based server system. The accuracy for PalTrack is 0.6–2.4 m. The key component of the PalTrack system is the RSS-based geolocation algorithm.

Received Signal Phase: The received signal phase is another possible geolocation metric. It is well known that, with the aid of reference receivers to measure the carrier phase, differential GPS (DGPS) can improve the location accuracy from about 20 m to within 1 m compared with standard GPS, which only uses pseudorange measurements [Kap96]. One problem associated with the phase measurements lies in the ambiguity resulting from the periodic property (with period 2π) of the signal phase, while the standard pseudorange measurements are unambiguous. Consequently, in DGPS, the ambiguous carrier phase measurement is used for fine-tuning the pseudorange measurement. A complementary Kalman filter is used to combine the low noise ambiguous carrier phase measurements and the unambiguous but noisier pseudorange measurements [Kap96]. For an indoor geolocation system, it is possible to use the signal phase method together with the TOA/TDOA or the RSS method to fine-tune the location estimate. However, unlike DGPS, where the LOS signal path is always observed, the serious multipath condition of the indoor geolocation environment causes more errors in the phase measurements.

Direction-based Techniques. If the positions of two fixed transmitters are known, then a receiver can compute its own position by determining the angles at which these transmitters are located with respect to itself, as shown in Fig. 12.3*d*. This is called the DOA or AOA technique for position location. If the accuracy of the direction measurement (roughly the beam width of the antenna array) is $\pm\theta_s$, then the AOA measurement at the receiver will restrict the transmitter position around the LOS signal path with an angular spread of $2\theta_s$. Two such AOA measurements will provide a position fix, as illustrated in Fig. 12.3*d*.

Consequently, one of the problems with this technique is the wide angular ranges for transmissions of most signals in wireless systems. Antennas in most 2G cellular systems are either omnidirectional or transmit over angles as wide as 120°. DOA/AOA techniques are thus not specified in any of the positioning standards for cellular systems. Moreover, the accuracy of the position estimation depends on where the receiver is located with respect to the transmitters. If the receiver lies in between the two transmitters along a straight line, then AOA measurements will not be able to provide a position fix. As a result, more than two transmitters are usually needed to improve the location accuracy. Also, radio signals propagate through reflections and diffraction, rendering the direction from which a signal arrives potentially random. For macrocellular environments, where the primary scatterers are located around the transmitter and far away from the receivers (the GBSs), the AOA method can provide acceptable location accuracy [Caf98]. But dramatically large location errors occur if the LOS signal path is blocked and the AOA of a reflected or a scattered signal component is used for estimating the direction. In indoor environments, surrounding objects or walls mostly block the LOS signal path. Thus, the AOA technique is not suitable for indoor geolocation systems. In addition, this requires placing expensive array antennas at the receivers to track the direction of arrival of the signal. While this is a feasible option in next-generation cellular systems, where smart and narrow-beam antennas may be deployed to increase capacity, it is in general not a good solution for low-cost indoor applications.

Signature-based Techniques. Problems in TOA or AOA schemes arising due to multipath propagation can be overcome by exploiting multipath propagation for position location. The idea behind this technique is that specific positions in a given environment have specific or unique *radio signatures* or *fingerprints* [Kos00]. The received signal can be extremely site specific because of its dependence on the terrain and intervening obstacles. So the multipath structure of the channel is unique to every location and can be considered as a fingerprint or signature of the location if the same RF signal is transmitted from (or received at) that location. This property has been exploited in proprietary systems to develop a "signature or fingerprint database" of a location grid in specific service areas. The received signal is measured as a vehicle moves along this grid and recorded in the signature database. When another vehicle moves in the same area, the signal received from it is compared with the entry in the database and thus its location is determined. Such a scheme may also be useful for indoor applications, where the multipath structure in an area can be exploited. By creating a database of signatures and the associated positions, it will be possible to estimate the position of an MS if it can measure the radio signature at its own position and compare it with entries in the database.

Example 12.4: Commercial Geolocation Systems Based on Fingerprinting A company called US Wireless Corporation has designed and implemented a system based on this scheme called RadioCamera in downtown Oakland in California [RadCa]. The mobile transmits RF signals, which scatter because of multipath conditions. RadioCamera™ takes measurements of the RF signals and collects all the multipath rays. A location pattern signature is developed using the multipath rays. The location signature is compared with a learned database and a location is determined. Continuous measurements of the location pattern signature provide tracking. Ekahau [Eka] is a vendor of a geolocation system for WLANs that uses RSS fingerprints for estimating the location of an MS.

In WLANs, the received signal strength from multiple different APs at a given position can be used as a location-specific signature [Bah00]. Since WLANs are widely deployed and MSs that can connect to WLANs can scan for multiple APs and measure the RSS, it is possible to deploy a fingerprint-based geolocation system without additional infrastructure or need for special hardware. In this case, typically, the RSS from N APs forms an N-dimensional vector. The mean vector $\mathbf{R}_{(i)} = [r_{1(i)}\, r_{2(i)}\, r_{3(i)}\, \ldots\, r_{N(i)}]$ is stored in the database for locations $i = 1, 2, 3, \ldots, L$, typically on a square or rectangular grid in the physical service area. Thus, the database has L entries, each corresponding to the mean RSS vector at a location. When an MS wants to determine its location, it measures a *sample* RSS vector $\mathbf{S} = [s_1\, s_2\, s_3\, \ldots\, s_N]$. The Euclidean distance between $\mathbf{S}$ and $\mathbf{R}_{(i)}$ is given by

$$D_{(i)} = ||\mathbf{S} - \mathbf{R}_{(i)}|| = \sum_{n=1}^{N} (s_n - r_{n(i)})^2 \tag{12.2}$$

The Euclidean distances between the sample vector and all entries in the database are computed (i.e. $D_{(\mathrm{i})}$ is computed for $i = 1, 2, 3, \ldots, L$). The location associated with the entry in the database that has the smallest Euclidean distance to the measured sample is chosen as the estimate of the location of the mobile. This process, illustrated in Fig. 12.6,

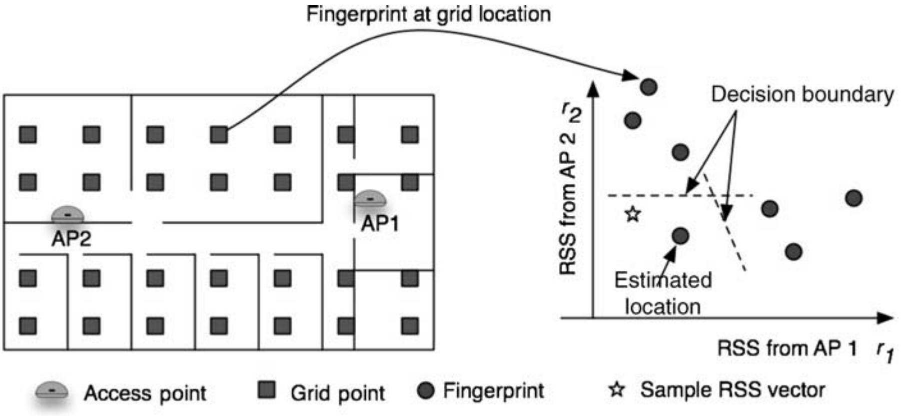


FIGURE 12.6 Illustration of estimating locations using RSS fingerprints.

is similar to matching a received signal to one in the signal constellation, as described in Chapter 3. The primary difference is that the “noise” in this case need not be Gaussian (although this assumption is commonly made). Further, the entries in the database do not necessarily form a structured constellation. More details of matching samples to database entries and the associated errors are described in [Kae04, Swa08].

Unfortunately, for several reasons, signature-based techniques are not free of disadvantages. Radio signatures change with time due to changes in the environment. Moreover, the size of the database needs to be reasonable and may not capture all possible signatures. It is very laborious to collect RSS data to populate the fingerprint database. Finally, it is unlikely that N APs are visible at a location all the time. It is possible to distinguish between locations based on the visibility of APs, but the accuracy is likely to be lower.

Hybrid Positioning Schemes. Recently, attempts have been made to use a combination of sensing methods and technologies to improve the location estimates. A good example is the ability of PDAs to use both WLANs and cellular signals to estimate the location of the MS. A popular commercial enterprise involved in using WLANs for location estimates is Skyhook. Recent research has also considered the use of both Bluetooth and WLANs to improve the accuracy of location estimates within buildings.

12.3.3 Positioning Standards for Cellular Telephone Systems

In this section we will describe techniques specified by cellular telephone standards for RF location sensing. These standards employ the techniques described in the previous section. The options for E-911 services when they were first mandated included traditional GPS or a network-centric approach based on TDOA techniques. GPS, especially after the elimination of *selective availability* of the signal by the US Government, provides sufficient accuracy for E-911 systems. The one disadvantage of GPS is that the time to first fix can be very long, depending on what satellite constellation an MS may be able to see. There is also the problem of using GPS in urban canyons. However, compared with a standalone GPS, network-centric approaches can provide a faster time to first fix but are unreliable and inaccurate. A variety of approaches are now part of cellular standards for geolocation.

Techniques Based on Closeness to the Point of Association: The standard specified for locating MSs using POA is called the cell-ID technique [Trev04, 3GPP 25.305]. In most cellular standards, the BS serving a cell broadcasts information about itself. In GSM, the broadcast control channel carries this information. In IS-95 or cdma2000 systems, the pilot channel and sync channels together provide information about the BS. In 3G UMTS, the cell-ID technique can use paging, location or routing area updates, or cell update messages to get information about the serving BS [3GPP 25.305]. When an MS associates itself with the BS, it is aware of the cell (or cell sector) in which it is located. One may assume that this will be the BS closest to the MS. This is true only if the RSS of the broadcast control channel or pilot channels from the nearest cell are the strongest. In most cases, MSs latch on to the BS with the strongest signal. Owing to radio propagation effects, an MS may sometimes associate itself with a BS that is farther away. As reported by Trevisani and Vitaletti [Trev04], this could be as high as 43% of the cases.

The coverage of a cell is irregular and needs to be known a priori to know in what region an MS is located. The accuracy reported by Trevisani and Vitaletti [Trev04] was 800 m in the New York area and 500 m in Italy using the cell-ID technique. The accuracy of the cell-ID technique can be improved by using the round-trip time in CDMA systems or timing advancement information used in the framing structure in TDMA systems.

Techniques Based on TOA/TDOA: Most standard cellular positioning schemes use TOA or TDOA for estimating the position of the MS. The standards include enhanced observed time difference (E-OTD), observed time difference of arrival – idle period on downlink (OTDOA-IPDL), uplink TDOA (UTDOA), assisted GPS (AGPS), and advanced forward link trilateration (A-FLT). We discuss these standards briefly below.

The E-OTD standard was the earliest standard for positioning of MSs and it is suggested for GSM and EDGE systems. The idea here is that the MS will determine the TDOA from multiple BSs and determine its position using standard TDOA techniques or using improved algorithms. It is called "enhanced" because additional *location measurement units* are required to compute the timing difference between the clocks in different GSM BSs (that are not synchronized). The accuracy and precision with this method does not usually meet the FCC mandate because of clock accuracy and timing accuracy issues with E-OTD. The accuracy of E-OTD has been reported to be between 50 and 400 m, with availability varying between 70 and 95% depending on whether the area covered is urban or rural [Mar02]. Moreover, the time to obtain the position estimate can be as large as 5 s and software changes are necessary in the handset to enable it to work with E-OTD.

The A-FLT standard is similar to E-OTD except that it uses the pilot signals in IS-95-based cellular systems. Pilot signals from the serving BS and a neighboring BS are used to compute the TDOA hyperbolas. The advantage here is that BSs in IS-95 are already time synchronized using GPS. The chip duration in the spread-spectrum signals used in IS-95 is 0.813 μs, which results in better accuracy. In A-FLT, the resolution used for reporting is actually $^1/_8$ of the chip duration. Measurement reports in Nissani and Shperling [Nis00] indicate that the accuracy is 48 m 67% of the time and 130 m 90% of the time. Enhanced forward link trilateration (E-FLT) is similar to A-FLT, but it uses a different signaling protocol with the network and it also could use additional information from location measurement units or radio signatures to improve accuracy. Because of cdma2000's similarity to IS-95, A-FLT and E-FLT are also used with that standard.

The OTDOA standard also uses TDOA measurements, but in the UMTS standard. The MS measures the time difference between frames transmitted by multiple BSs. Location measurement units are used in a manner similar to E-OTD to account for asynchronous transmissions from BSs. Since UMTS uses CDMA as the access scheme, an MS close to a BS may not be able to hear signals from other BSs because they are swamped by the high power of signals from the closer BS. To allow the MS to make measurements from other BSs, each BS ceases its transmission for short periods of time (called idle periods). This technique is called OTDOA-IPDL for this reason. Either the MS can compute its own position or it can report the measurements to the network where a stand-alone mobile location center computes the MS's position. Simulation results in [Por01] indicate that the OTDOA method can provide good accuracy and precision. In rural areas, the accuracy is 17 m 67% of the time and 27 m 95% of the time. In urban areas, the accuracy is between 68 and 86 m 67% of the time and between 156 and 193 m 95% of the time, depending on the type of urban environment. In the UTDOA standard, different location measurement units compute the position of an MS by using the TDOA of a signal transmitted by the MS that they all receive. This method does not need the MS to do anything, but location measurement units have to be deployed by the service provider.

Network AGPS is a method that significantly improves the accuracy of position estimates in a cellular system [Dju01]. It is possible to install a GPS receiver in each MS so that an MS can determine its own position. GPS by itself in an MS is not a viable solution for many reasons. The time to first fix of a cold receiver can be several minutes. The MS needs a clear view of the skies to observe at least four satellites, so this approach does not operate well in urban canyons. If the MS has to scan for satellites, acquire the signals from the satellites, and then determine its own position, this could consume the MS battery power significantly. In the case of AGPS, a partial (or low-complexity) GPS receiver is built into the MS. Additionally, AGPS servers are placed in the network as appropriate. By predicting what signal an MS may see and sending that information to the MS, the network entity can enable a faster time to first fix, shortening it from minutes to a second or less [Dju01]. The wireless network signals information about the reference time, the visible satellite list, the satellite signal spread-spectrum code phase, and search windows appropriate for these signals to the MS, reducing the burden on the MS. This improves the time to first fix and also the accuracy of position estimates. AGPS also enables the network entity to detect signals with weaker signal strength than an MS and send a *sensitivity assistance message* to the MS. AGPS can be used in GSM, IS-95, cdma2000 and UMTS systems and in conjunction with other techniques, such as cell-ID, E-OTD, A-FLT or OTDOA when GPS signals are not available. Accuracy estimates for AGPS range between 5 and 30 m.

Techniques Based on Location Signatures: While there is no standard specified for using RF location signatures for estimating the position of an MS, some companies [PolW] and research studies have demonstrated the feasibility of this approach. In Ahonen and Laitinen [Aho03], the multipath intensity profile at a given location is used as the RF signature. A database correlation mechanism is used where the measured multipath profile is correlated with profiles stored in a database. The profile in the database with the highest correlation is chosen as the estimated position of the MS. Simulation results by Ahonen and Laitinen [Aho03] show that 67% of the position estimates are within 25 m of the exact position and 95% are within 140 m.

Table 12.2 provides a summary of these different geolocation approaches based on some of the performance measures discussed previously. The table is self-explanatory.

12.4 LCS ARCHITECTURE FOR CELLULAR SYSTEMS

The focus of this section is to provide an overview of the network architecture and protocols that are required to be in place in cellular wireless systems to provide location-based services once the location sensing is completed. The complexity and large number of protocols, the different types of cellular networks, and the numerous

TABLE 12.2 Comparison of Geolocation Techniques in Cellular Systems

Geolocation technique	Coverage	Accuracy	Delay	Complexity/cost/others
Cell-ID	Almost everywhere	Poor – as high as 800 m	Low; information is included in signaling messages	Low
E-OTD	70–95% depending on type of area	Medium – because of clock accuracy (50–400 m)	Can be as high as 5s	Needs additional measurement units to account for lack of synchronization
A-FLT or E-FLT	Good	Reasonable – 48–130 m	Medium	Makes use of the small chip duration in CDMA systems No changes to handsets Privacy is network controlled
OTDOA	Good	Good; dependent on type of area	Medium	Needs BSs to stop transmissions periodically to avoid near–far effect
AGPS	Good; indoor coverage is suspect	Superior	Short time to first fix but needs a lot of signaling between network and mobile	Computation is offloaded to the network so that there is minimum impact on handset battery life
Radio fingerprinting	Good	Good	Computational effort to match fingerprints	Collecting a fingerprint database may be labor intensive

standards that apply under different circumstances make it difficult to document such information completely in a short section. Consequently, we present a simplified overview with a classification that closely follows how someone outside the field would perceive the network and its services. Readers should refer to the detailed standards produced by the 3rd Generation Partnership Project (3GPP) [e.g. 3GPP 25.305] and its equivalents in the USA for a complete understanding of the subject. We would also like to point out that the messaging considered in this article is primarily related to control messages (i.e. signaling required to enable the actual communication or transaction).

As described in Section 12.3, there are several techniques, such as cell-ID, OTDOA, A-FLT, etc., that are used to determine the position of the MS in cellular networks. In cellular wireless systems, the normal infrastructure for transporting voice and data traffic is leveraged in an *LCS architecture* for transporting the location-sensing information and query management for location-based services. For simplicity, we break up the signaling and communication of location information into three parts:

1. *Over the air* transport (or access network communications), which requires communication between the MS and the rest of the network. For example, the IS-801 standard handles this part of the communication for CDMA-based cellular systems such as IS-95 and cdma2000.
2. *Signaling within the fixed part of the cellular system* (core network communications), which is necessary to enable position determination and account for mobility issues. For example, the JSTD-036 standards specified by the EIA/TIA for wireless emergency services and LCS extend the ability of the signaling network to transport location- and emergency-related information in standardized formats.
3. *Application protocols that make use of the LCS architecture*. We consider the *mobile location protocol* (MLP) developed by the Open Mobile Alliance (OMA) that defines a set of constructs and services to enable location-based services for applications.

This section summarizes the three types of communications required for location-based services in cellular networks. The goal is not to cover all standards and architectures, but to provide an overview and summary of some of the standards and messaging that happens in cellular networks for location-based services.

12.4.1 Cellular Network Architecture

Figure 12.7 shows a schematic of the architecture of a cellular system. While this schematic is not particular to a specific standard, it provides an idea of the different components in the network. The reader is referred to Chapter 7 for more details of architectures of cellular systems. In the radio access subsystem, the MS, sometimes called user equipment, is the device whose position is to be determined. BSs, as before, are fixed transmitters that are points of access to the rest of the network. An MS communicates with a BS during idle periods (signaling), cellular phone calls (voice), or other data transmission. BSs are controlled by radio network controllers (RNCs), which also manage the radio resources

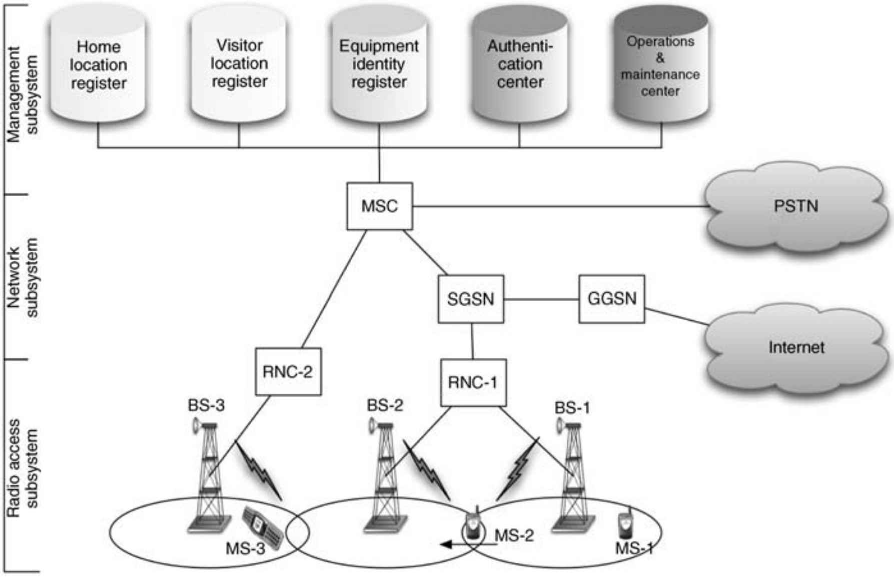


FIGURE 12.7 Generic cellular network architecture.

of each BS and MS (frequency channels, time slots, spread-spectrum codes, transmit powers, and so on).

The network subsystem carries voice and data traffic and also handles routing of calls and data packets. The MSC and the serving and gateway GPRS support nodes (SGSNs and GGSNs) are responsible for handling voice and data respectively. These network entities perform the task of mobility management, where they keep track of the cell or group of cells where an MS is located and handle routing of calls or packets when an MS performs a handoff, i.e. when it moves from one cell to another (for example, see MS-2 in Fig. 12.7). They connect to the PSTN or the Internet. Several databases in the management subsystem are used for keeping track of the entities in the network that are currently serving the MS, security issues, accounting, and other operations, as shown in Fig. 12.7.

The above architecture was designed specifically to handle voice and data communications and needs enhancements to enable support for LCS. In particular, new entities are required to determine position location information, communicate this information appropriately to the concerned parties (PSAP – for emergency services, LCS clients, and so on). These changes are described next. This is similar to the changes required in cellular systems to support data traffic, as described in Section 7.7.1, where new entities (primarily the support nodes) were introduced to interface the cellular system to the Internet. In the case of the LCS architecture, new entities to support location estimation are introduced.

12.4.2 Location Services Architecture

As shown in Fig. 12.8, additional network entities are required to support LCS. The architecture shown in Fig. 12.8 does not correspond to any particular standard,

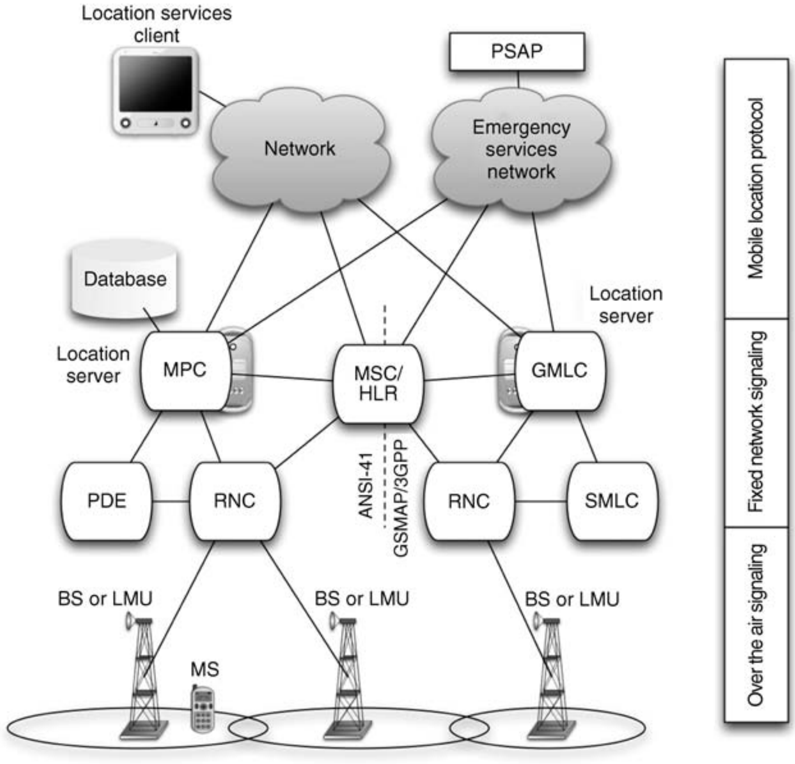


FIGURE 12.8 Architecture for LCS in cellular networks.

but tries to present some of the important network entities that are part of different standards, such as the J-STD-036 and 3GPP TS 25.305. The goal here is to provide a general discussion of these entities rather than to describe each standard individually. Also, some of these entities may be collocated, although they are shown separately in Fig. 12.8.

The location measurement unit (LMU) is a device that assists the MS in determining its position or uses signals from the MS to determine the position of the MS. It is used with AGPS to help the MS determine its position. With other positioning techniques, such as UTDOA, it makes measurements of radio signals and communicates this information to network entities such as the RNC. An LMU may be associated with a BS, in which case it communicates with the RNC over a wired link. Alternatively, it may be a stand-alone LMU which uses the air interface to communicate with the RNC.

The mobile positioning center (MPC) is the entity that handles position information in cellular networks that use ANSI-41 for signaling – typically North American cellular systems [TR-45-02]. It uses a position-determining entity (PDE) to determine the MS's position using a variety of technologies, such as AGPS or some OTDOA. The PDE can determine an MS's position while the MS is in call or when it starts a call (using information from the MS or LMUs). There may be multiple PDEs that are used by one MPC. The MSC is associated with an MPC. The same MPC may be associated with multiple MSCs. The MPC and MSC communicate with the emergency services network (ESN), as described later. The MPC also handles access restrictions to the position information.

In 3GPP-based networks [3GPP 25.305] such as the UMTS or GSM, the gateway mobile location center (GMLC) and the serving mobile location center (SMLC) take up the responsibilities for positioning similar to the MPC and PDE. The GMLC is the first point of contact when the position of an MS is required. The SMLC coordinates the resources necessary to determine the MS's position, sometimes calculating the position and accuracy itself (using information from the MS or LMU).

An emergency call is ultimately answered by a PSAP. The PSAP connects to the fixed infrastructure in the cellular system through the ESN, which interfaces with the MPC or GMLC and the MSC. Two types of call are considered by the ESN – the first where the position information is pushed to the ESN along with the signaling that occurs with the emergency call, and the second where the ESN has to pull the position information. In the former case (called call-associated signaling or CAS), an ESN entity (ESNE) communicates with the MSC serving the emergency call and obtains the position information. In the latter case (called non-call-associated signaling or NCAS), an emergency services message entity (ESME) interfaces with the MPC or GMLC to pull the position information. The database shown attached to the MPC in Fig. 12.8 translates the position of the MS into a number specifying the emergency service zone where the MS is located. It is the emergency service zone that is assigned to a PSAP and emergency services such as police, fire, or ambulance.

Finally, at a higher layer we can consider LCS clients and location servers (often collocated with the GMLC or MPC). These are often independent of the underlying network technology. The LCS client may be requesting position information either for emergency services or for some other purposes (e.g. concierge services). In any case, it has to communicate over some network (usually using the IP) with the location server to obtain the MS's position.

12.4.3 Over the Air (Access Network) Communications for Location Services

Certain communications have to take place over the air interface in the process for locating an MS, which are specified by standards that are different for IS-95/cdma2000 and 3GPP systems. Although such signaling happens over the air, it is also important from the point of view of functionality: it enables querying MSs for measurements related to positioning, such as signal strength, timing, round-trip times, GPS information, and so on.

The IS-801 standard [IS-801-99] defines the signaling messages between an MS and a BS to support position determination in IS-95 or cdma2000 networks. The standard specifies the formats of messages and procedures to be adopted by the MS and BS when messages are received. Most of the messages are in the form of requests and responses. Some of the request messages sent by the MS include those that ask for BS capabilities, GPS assistance, GPS almanac, and GPS ephemeris. Some of the response messages sent by the MS include MS information, pilot phase, time offset measurements, and pseudorange measurements.

The BS makes requests for the response messages from the MS and provides responses to the MS requests.

In 3GPP, the signaling messages to support position determination are carried by the radio resource control (RRC) messages [3GPP 25.331]. The form of messaging is similar to the request–response scheme in IS-801. The serving RNC generates RRC measurement control messages that are sent to the MS through the BS. They include data about the SMLC, assistance data related to GPS, or instructions to the MS to perform measurements. In response, the MS sends an RRC measurement report that contains the position of the MS or other measurements that will help the SMLC to determine the position. Several control and report messages may be necessary before success or failure of the position determination.

12.4.4 Signaling in the Fixed Infrastructure (Core Network) for Location Services

When a request for the position of the MS arrives, entities within the fixed infrastructure in Fig. 12.8 have to communicate information between them to support the determination of the position and to deliver this information to the appropriate destination. The signaling for this is specified in the JSTD-036 and 3GPP TS 25.305 V. 7.2.0 standards. These standards consider a variety of scenarios, such as CAS and NCAS calls, automatic detection of emergency calls, handling position determination and delivery in the case of calls that are in handoff, and so on. We provide a brief summary of some of the signaling in the fixed infrastructure.

When an MS initiates an emergency services call, the delivery of position information to the PSAP is called emergency location information delivery (ELID). The MSC serving the MS that makes the emergency call contacts the MPC or GMLC for position information. A PDE or SMLC may have automatically detected the invocation of an emergency call and started procedures for computing the MS's position. Alternatively, the MPC or GMLC may contact the PDE or SMLC to start such procedures and obtain the information. Once the MPC or GMLC has the position information, it will send such information to the MSC. In the case of a CAS call, the MSC sends the position information to the ESNE. The information sent includes information such as the calling party number and the position of the MS. In the case of an NCAS call, the MSC serving the MS making the emergency call will send the emergency service routing digits (essentially information about the BS or cell sector serving the MS) to the ESNE. Then, the ESME associated with the ESNE autonomously requests the position information from the appropriate MPC or GMLC. Two or more MSCs may be involved in delivering position information if the MS is in handoff during the call.

Several standard formats are used for protocols between different networking entities involved in ELID. In ANSI-41-based systems (such as cdma2000 and IS-95), ANSI-41 is used for communication between MSCs, and also between MSCs and other entities. Two special protocols are also used in such systems. The *LCS protocol* is used between the PDE and MPC or between the routing database and MPC. The *emergency services protocol* is employed between the MSC and the ESME. Protocols developed for ISDN are used for communication between the MSC and ESNE. The GSM Mobile Applications Part has been extended in 3GPP systems to enable similar communications.

12.4.5 The Mobile Location Protocol

The MLP is an example of an application-level mechanism used by an LCS client to obtain position information about an MS from a location server (such as the MPC or GMLC). This protocol can also be used for emergency services. The MLP was initially developed by the Location Interoperability Forum (LIF) [LIF TS 101-02]. The goal of the MLP specification was to develop standard methods (using extensible markup language (XML)) for Internet applications to obtain position information from cellular network entities. The work of the LIF was later rolled over into the activities of the OMA [Bren05], an industry forum consisting of hundreds of telecommunications and related companies for generating market-driven specifications for mobile services to ensure interoperability between these services.

Some protocol/service layers associated with the MLP are shown in Fig. 12.9. At the top, three types of MLP service are defined. The basic MLP service corresponds to emergency services and ELID as it is defined by 3GPP. Advanced and other services can be developed to follow the MLP specifications as required. Some services that have already been defined include standard and emergency LCS (further classified into *immediate* for delay-sensitive single-location response and *reporting* for an LCS client requesting position information). A triggered location reporting service is specified for the instance when an MS's location needs to be reported when an event occurs or a certain time elapses. The core location elements are in the form of document type definitions that form the building blocks for the XML. The transport of these XML messages is specified in the lowest layer. Mapping to standard web/web services protocols like HTTP and SOAP are specified here, as well as mappings to the wireless session protocol, which is part of the wireless application protocol developed by the OMA.

The way the MLP operates is fairly simple. An LCS client requests position information from a location service using an MLP request, which may, for example, be transported using XML in HTTP and SSL (see Chapter 11 for a brief discussion of SSL). The location request is an XML document that could include the MS's identification in either the North American mobile identification number format or the GSM mobile subscriber identifier format, the age of the position information, the response time, and accuracy. The response from the location server is also delivered to the LCS client using the MLP. The location response is also an XML document, which provides information such as the accuracy, response time, and so on.

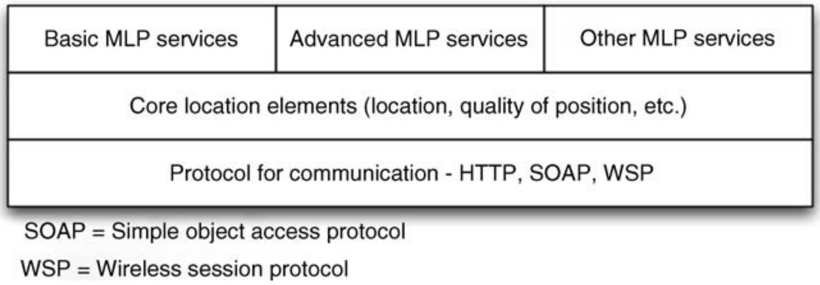

Basic MLP services
Advanced MLP services
Other MLP services
Core location elements (location, quality of position, etc.)
Protocol for communication - HTTP, SOAP, WSP
SOAP = Simple object access protocol
WSP = Wireless session protocol

FIGURE 12.9 Part of the MLP stack.

12.5 POSITIONING IN AD HOC AND SENSOR NETWORKS

In infrastructure-free ad hoc networks or sensor networks, geolocation is referred to as "localization," where each node (mobile or stationary) in the network needs to determine its location. In sensor networks, the location of a sensor is important because the physical sensed quantity (e.g. temperature in a reactor) is often correlated with the location (e.g. where in the reactor is the temperature too high?). The resource constraints of nodes make it infeasible to embed GPS chips in them. The lack of fixed infrastructure in many cases makes geolocation a challenging problem.

The problem of localization in ad hoc and sensor networks can be explained with reference to Fig. 12.10. The connectivity of nodes in this network is indicated by dashed lines (only nodes with lines between them can communicate directly; to reach other nodes, multi-hop communications are necessary). The assumption in these networks is that a certain fraction of nodes in the network are location aware (either because they have a GPS chip embedded or because the network administrator has manually included the location information). In general, the problem of localization in an ad hoc network then becomes one of determining the location of the remaining nodes in the network. For example, in Fig. 12.10, node B can directly see three location-aware nodes. By measuring the distance to these three nodes (using signal strength or TOA), node B can estimate its own location. Once node B knows its own location, it can help node F determine its location, since node F can see node B and two location-aware nodes. This way, most nodes can estimate their locations. However, there are errors that occur when node B estimates its location. These errors carry over to the location estimation of node F. Some nodes may never be able to see three nodes that are aware of their location (even nodes that have iteratively determined their locations, like node F). However, they may collaboratively be able to compute their locations, as in the case of nodes C and G, which connect to different location-aware nodes. In extreme circumstances the location error may be significant; for example, in

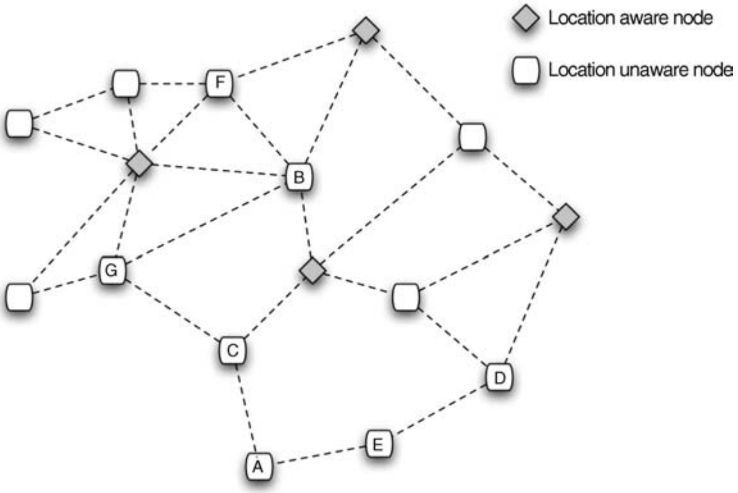


FIGURE 12.10 Localization in ad hoc networks.

the case of node A, which has no direct connection to a location-aware node and has very few neighbors.

Most work in this area [e.g. Sav01] employs either the signal strength or TOA measurements to determine the distance. Connectivity has also been suggested as a metric to estimate the range to a set of location-aware nodes [e.g. Bul00]. The use of multiple ultrasound devices to estimate the direction of arrival is proposed by Niculescu and Nath [Nic03], with ranging to estimate the positions of nodes in infrastructure-free networks.

QUESTIONS

1. Why are cellular service providers interested in location based services? Give some examples of location based services.
2. Explain the differences between GPS, wireless cellular assisted GPS, and indoor geolocation systems.
3. What are E-911 services and who has mandated these services?
4. Explain three differences between performance metrics for telecommunications and geolocation services.
5. Compare mobile centric and network centric geolocation techniques in terms of complexity and accuracy.
6. What are the basic elements of a wireless geolocation system?
7. Differentiate between remote and self-positioning systems.
8. What is the advantage and what is the disadvantage of using closeness to the point of association as the estimate of a mobile station's location?
9. Name the three major distance based techniques used for location finding and explain how they are implemented in a system.
10. How is TOA different from TDOA for geolocation?
11. Why is RSS not a very good measure of the distance between a transmitter and a receiver? How can distance estimates with RSS be improved?
12. Why are AOA techniques not popular in indoor geolocation applications?
13. What are the advantages and disadvantages of RSS fingerprint based location techniques in indoor areas?
14. Why is it not certain that a MS is associated to the base station or access point that is physically closest to it?
15. In using TOA/TDOA in cellular systems, what two reasons make the accuracy better in CDMA systems compared to GSM?
16. How is the OTDOA scheme different from the UTDOA scheme?
17. Why do cellular networks employ assisted GPS instead of a full GPS receiver in each MS?
18. Compare the accuracy performance of the different TOA/TDOA based location estimation schemes in cellular systems.
19. Describe the functionality of a mobile positioning center in cellular systems.
20. What new entities are required in the cellular network architecture to support location services?
21. How can nodes in an ad hoc or sensor network determine their locations even if they are not directly connected to nodes that are location aware?

PROBLEMS

Problem 1:

Two base stations located at (x, y) coordinates (500, 150) and (200, 200) are measuring the angle of arrival of the signal from a mobile terminal with respect to the x-axis. The first base station measures this angle as 45^{o} and the second as 75^{o}. What are the co-ordinates of the mobile terminal?

Problem 2:

In Problem 1, what happens if the first base station incorrectly measures the AOA from the mobile terminal as 50^{o}? 30^{o}?

Problem 3:

Base stations A, B and C located at (50, 50), (300, 0), and (0, 134) are found to be at distances 90, 200 and 100 m from a mobile terminal. Draw circles corresponding to these values and try to determine the location of the mobile terminal.

Problem 4:

In Problem 3, what happens if the mobile incorrectly measures the distance from base station B as 100 m? 300 m?

Problem 5:

Consider a fingerprint database with five fingerprints given by [−40 −65 −70], [−75 −33 −57], [−67 −55 −71], [−38 −59 −59], and [−55 −55 −55] (all values are in dBm) associated respectively with five rooms labeled A, B, C, D and E respectively. An MS samples the received signal strength from the three access points in the area. The sample RSS vector is [−41 −60 −66] dBm. Compute the Euclidean distance between the sample RSS vector and the five fingerprints. Sort the locations according to the ascending order of Euclidean distances. Where is the MS most likely located? Is there is chance that this estimate is erroneous? Why?

PROJECTS

1. In this project, you will develop a crude positioning system that makes use of the observed SSIDs of WLANs in the area. Use any of the tools mentioned in the projects section in Chapter 9 like Netstumbler, iStumbler, or Inssider to discover the *visible* SSIDs of networks at different locations along your street. Create a table that associates a set of visible SSIDs with a particular location. The sets of visible SSIDs must be different, so that you can distinguish between locations. Then try to predict where you are by looking at the list of visible SSIDs at random locations along your street. Describe how accurate this method is according to your experiments. Suggest alternatives to improve the accuracy.

2. Let us suppose that there are N estimates d_i of the distance of an MS from N known locations with coordinates (x_i,y_i) for $i = 1,2,3,\ldots,N$. We then have N equations of the form:

$$f_i(x,y) = (x_i-x)^2 + (y_i-y)^2 - d_i^2 = 0$$

where (x, y) is the unknown location of the MS. The *least squares* technique provides a method of estimating x and y when there are errors in the estimates d_i. The technique works as follows. Let $\mathbf{F} = [f_1(x, y)\ f_2(x, y)\ \ldots\ f_N(x, y)]^{\mathrm{T}}$. Construct the *Jacobian* matrix given by:

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial f_1(x,y)}{\partial x} & & \dfrac{\partial f_1(x,y)}{\partial y} \\ & \cdots & \\ \dfrac{\partial f_N(x,y)}{\partial x} & & \dfrac{\partial f_N(x,y)}{\partial y} \end{bmatrix}$$

Pick an estimate of the solution $\mathbf{U} = [x^*\ y^*]$. Determine the error in the solution as $\mathbf{E} = -(\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{F}$ evaluated at $\mathbf{U}$. The new solution is $\mathbf{U} + \mathbf{E}$. Iteratively, the error in the solution is reduced by computing a new error that is added to the previous solution to obtain a new solution till a point is reached where the solution does not change. Write a Matlab program that takes as input the N known locations, the N range estimates, a guess of the solution, and provides as output the final solution for the location of the MS. Also, add additional code to plot the locus of the intermediate solutions starting from the initial guess.

13

WIRELESS SENSOR NETWORKS

13.1 INTRODUCTION

Over the last decade, sensor networks and applications relying on sensor networks have become prevalent, as also has research in academia and a burgeoning industry in sensor networking devices. Sensor devices have evolved over the last decade to support various

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

applications, such as asset monitoring, surveillance, structural health monitoring, habitat monitoring, and even underwater sensing. In this chapter, we provide an overview of wireless sensor networks.

We start in Section 13.2 by describing several applications that have considered implementations of sensor networking, such as habitat monitoring and structural health monitoring. In Section 13.3 we provide an overview of sensor network architectures. We also discuss wireless sensor devices, the actual hardware that makes sensor networks work, and some of the platforms that sensor devices employ. Sections 13.4 and 13.5 discuss the physical and MAC layers of sensor networks and Section 13.6 provides a discussion of higher layer issues for sensor networks. In these sections, some of the discussion is devoted to IEEE 802.15.4 low-rate wireless personal area networking devices and the standard. This has some commonality with the treatment of Zigbee and 802.15.4 in Chapter 10. We include this redundancy in treatment for completeness and to have a self-contained chapter on sensor networks.

13.2 SENSOR NETWORK APPLICATIONS

Sensor network applications are diverse, ranging from habitat monitoring (e.g. studying the climate on redwood trees and their impact on the ecosystems therein) to surveillance and physical intrusion detection (e.g. detecting breaking into museums). Applications can be of academic or scientific interest, or commercial where the sensor network can have significant impacts; for example, in improving crop yields. There still appears to be a significant disconnect between the actual deployment of sensor networks for real applications and academic research in sensor networking as discussed by Raman and Chebrolu [Ram08]. We ignore this issue in this chapter. In the following subsections, we look at examples where sensor networks have been or are being actively deployed. Where possible, we link the applications to the networking issues in subsequent sections.

13.2.1 Habitat Monitoring

Habitat monitoring is a scientific application that has benefited significantly from the deployment of sensor networks. In habitat monitoring applications, it is necessary to monitor a variety of environmental characteristics, such as temperature, humidity, barometric pressure, and other physical parameters, over significantly long periods of time and/or significantly large geographical areas. When such characteristics are monitored underneath the soil surface, inside a lake, on the surface of a tree, etc., the corresponding conditions are referred to as "microclimates."

An example of habitat monitoring using sensor networks, widely cited in the literature [e.g. Sze02a, Sze02b, Hu07], is that of understanding the impact of microclimates on habitat selection by sea birds. In particular, the objective of this project was to improve the understanding of sensor networks by using them to monitor the occupancy of nesting burrows on Great Duck Island in Maine. The hope here was that passive IR sensors used to detect heat from birds that were nesting and measures of temperature and humidity to indicate extended inhabitation of burrows could eventually replace laborious manual sampling and direct inspection of the island. A two-tier sensor network (see Example 13.3) was deployed by researchers over a 4-month period with 150 sensor devices. The researchers evaluated the performance of the sensor network, its lifetime, how nodes failed, and how nodes could be recovered after the completion of the project (about 50% of the devices were recovered).

Sensor networks are extremely useful for monitoring physical phenomena over large geographical areas requiring dense deployment of sensors. One such example, of what has been called a "macroscope" because of its similarity to a microscope in terms of its ability to reveal complex details, is the monitoring of air temperature, relative humidity, and amount of active solar radiation along the length of a 70 foot ($\sim$21.3 m) coastal redwood tree [Tol05]. The microclimate along a redwood tree can have significant differences, both spatially and temporally. The treetop gets sunlight, while the bottom is typically moist and cool because of the shade from leaves. The tree also moves large volumes of water from the soil upwards and eventually into the air. A wireless sensor network that collected data over 44 days, every 5 min, with nodes placed every 2 m along the redwood tree is described by Tolle *et al.* [Tol05]. The monitoring and subsequent analysis of data did reveal dynamic gradients of phenomena, as expected by the biologists.

An example of a mobile sensor network for habitat monitoring is the ZebraNet project [Ju02], where a sensor network was deployed to study the behavior of zebras in the Mpala Research Centre in Kenya. The idea in this project was that sensors equipped with GPS and peer-to-peer ad hoc networking can yield more information about animal behavior than simple radio collars that are sampled during the day by researchers driving around the natural area. No fixed infrastructure was available, which makes this network different from the previous examples.

13.2.2 Structural Health Monitoring

Structural health monitoring refers to the continual or periodic monitoring of the *health* of large structures such as bridges, buildings, or ships. The vibration data from bridges can be used to detect the health of bridges (e.g. whether it is ambient vibrations or some other serious condition). Examples of projects that have looked at monitoring bridges include those of Xu *et al.* [Xu04] and Kim *et al.* [Kim06]. Kim *et al.* [Kim06] monitored a 260 foot ($\sim$80 m) long suspension footbridge using 13 sensors that measured the vibration of the bridge using accelerometers and proposed extending the work to the Golden Gate Bridge in San Francisco.

13.2.3 Miscellaneous Applications

Other applications of sensor networks include asset monitoring, surveillance, etc. As described in Section 13.3, many vendors now carry sensor devices and the corresponding networking infrastructure to perform asset monitoring.

13.3 SENSOR NETWORK ARCHITECTURE AND SENSOR DEVICES

We start with a brief overview of the typical architecture of a sensor network to provide a view of how sensor devices fit into the network. Our objective in this section is to discuss the capabilities of sensor devices and the different commercial product platforms that are available on the market. We do not explore the details of the integrated circuit chips, radios, or other components on the sensor network platforms in great detail, except as necessary to illustrate their capabilities. Section 13.3.2 includes a brief history of sensor devices and their development. Section 13.3.3 describes a variety of commercially available sensor devices. Section 13.3.4 considers the evolution of sensor devices in the future and emerging issues.

13.3.1 Sensor Network Architecture

A typical sensor network architecture consists of a *sensor field*, which is the physical environment where the sensor nodes or devices are deployed (see Fig. 13.1). Sensor nodes can possibly be deployed in extremely large numbers, on the order of thousands of sensor nodes in the field. Consequently, the cost of these nodes should be low. A low-cost device can thus be expected to have fairly limited computational and communication capabilities, considering the fact that sensing capabilities are also to be included in the device. Moreover, in many applications, sensor nodes are deployed in hostile areas or physically inaccessible regions where it is not easy to have human intervention to maintain sensor nodes. Such sensor nodes have to operate on limited battery power and the batteries cannot be replaced easily. In such cases, sensor nodes have to be designed so that power-consuming operations such as the central processing unit or the radio used for communications are shut down when they are not being used. Of course, for specific applications (e.g. physical intrusion detection using cameras), sensor nodes may have more advanced capabilities. Thus, sensor devices may range from millimetre-sized devices fabricated on custom silicon to more general purpose cell-phone-sized devices with advanced capabilities.

Figure 13.1 shows a schematic of a simple sensor network architecture. Sensor nodes with limited capabilities deployed in the sensor field communicate to a powerful BS that links them to the Internet and a central manager for processing the sensed data. Communications to the BS have to go through several sensor nodes first, because all sensor nodes will not be typically able to communicate directly with the BS. This may be due to limited communication range, distance from the BS, intermittent sensor activity, and so on.

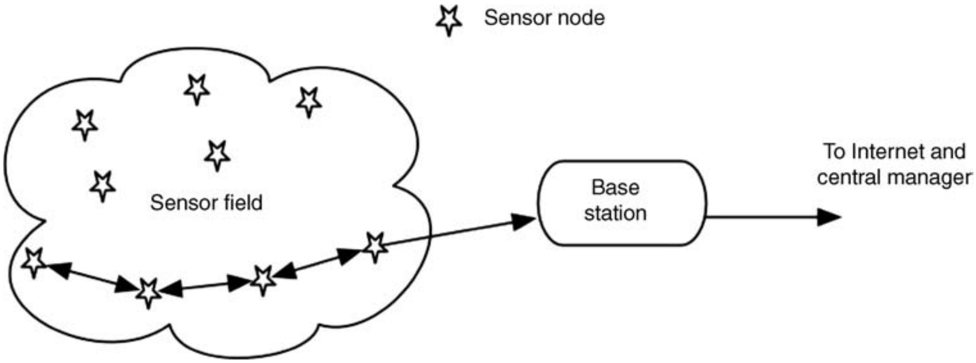


FIGURE 13.1 Typical sensor network architecture.

The simple architecture shown in Fig. 13.1 is expanded upon by Hill *et al.* [Hil04], where four classes of sensor networking devices are described. At the lowest level of the hierarchy, the actual sensing device could be very specialized, with a tiny form factor and very limited capabilities. Such devices may not even be capable of receiving information and may simply transmit information when they sense an event or perform other application-related activity. We will call these devices *submotes* in this chapter for ease of reference.

It is the responsibility of the second tier in the hierarchy (called the *mote* class) to receive this information and convey it using multiple hops towards the gateway or BS that forms the highest level in the hierarchy. The word "mote" itself means a tiny piece of substance, but the mote-class sensor device is larger in form factor and capability than the very basic sensing device. Figure 13.1 explicitly shows only the mote-class device and the BS.

The third device described by Hill *et al.* [Hil04] is a sensor device that is more complex than a mote in terms of its ability to communicate using higher bandwidths (e.g. using Bluetooth radios) and may have more random access memory (RAM) on its chip, and a processing unit that is more powerful than the one on the mote-class device. We will refer to this as a *supermote* in this chapter for easy reference.

The gateway or BS is the most powerful sensor device in the network. It typically has two interfaces: one that can connect to a mote-class device or a supermote device and the other that can connect to a larger network (cell-phone network, WLAN, or a wired LAN). The gateway may have other processing and storage capabilities that will be useful for the sensor application.

Example 13.1: Redwood Microclimate Monitoring In the Redwood "macroscope," mote-class sensor devices were deployed along the length of a redwood tree. The mote-class devices were equipped with sensors that measured total solar radiation by inspecting the spectrum and barometric pressure, but pressure measurements were not used. The solar radiation impacts photosynthesis and, thus, was a quantity of interest. Motes were deployed from about 15 m above the ground to 70 m above the ground on the west side of the tree very close to the tree trunk. The sensors delivered data to a BS-class device that connected to the Internet using a GPRS cellular network (see Chapter 7 for a discussion of data services over cellular networks).

Example 13.2: Hybrid Sensor Networks for Cane Toad Monitoring Monitoring of cane toad populations in Australia using sensor networks is reported by Hu *et al.* [Hu07]. This application is resource intensive, in that several FFTs have to be computed by the sensor devices to identify vocal characteristics of toads against background and other noise. One approach suggested by Hu *et al.* [Hu07] is to employ supermote- or BS-class devices that have more computational and communication resources. This approach is quite expensive due to the cost of these devices. Instead, a hybrid network of mote-class devices and BS-class devices can make the system more cost effective. In the work described by Hu *et al.* [Hu07], mote-class devices deployed on a larger scale take acoustic samples, perform some preliminary processing (compression) to reduce communication costs (see Section 13.6 on data aggregation and in-network processing), and send them to BS-class devices. The BS-class sensor devices use the received information from mote-class devices to determine the existence of cane toads.

Example 13.3: Network Architecture in Great Duck Island A tiered architecture was implemented on Great Duck Island [Sze02a] to monitor seabird occupancy. One network used single-hop transmissions from mote-class devices to a BS-class device. A second network used a multi-hop topology where mote-class sensors would route information to a BS-class device. Two types of mote were deployed: burrow motes that had IR and temperature/humidity sensors for use inside burrows and weather motes that monitored temperature, humidity, and barometric pressure on the surface. Node lifetimes were longer in the single-hop network, with most weather motes in the single-hop network being functional even after 4 months. Burrow motes performed additional sensing and had shorter lifetimes.

13.3.2 Overview of Sensor Devices

In this section we present a very brief history of sensor devices and also present a generic architecture of sensor devices. The actual architecture of a sensor device will be different, but the functional components can be expected to be similar.

Brief History. The concept of a network of extremely small devices that can sense physical phenomena (light, temperature, vibrations, motion, etc.) or even induce some activity (actuator) is believed to have its origins in the early 1990s with military applications based on microelectromechanical systems being the primary driver for this concept. The idea of *smartdust* that can be scattered in an environment, elements of which would self-configure themselves into a network, originated in the mid 1990s through several workshops organized by DARPA in the USA. While dust-sized sensor devices are not yet available (see discussion in Section 13.3.4), research in the area of very small devices that could sense phenomena and network with one another has been going on at full steam for more than a decade now.

Much of the initial effort was dedicated to developing sensor devices using off-the-shelf components, thereby reducing the cost of development. The development of WLANs during the 1990s was also useful in terms of enhancing radio communications in unlicensed bands that could also be exploited by sensor devices. The earliest mote-class devices were built using printed circuit boards and were an inch or two in size. The *Rene* mote was developed in 1999 [Hil04]. It had 512 bytes of RAM, 8 KB of flash memory, several expansion connectors, the ability to communicate at a data rate of 10 kb/s and consumed 60 mW of power when active. It was around 1999 that Bluetooth was also reaching stability as a cable replacement technology. In parallel, work on developing an operating system and programming tools for sensor devices was ongoing. The TinyOS operating system [Tiny] and the nesC programming language [Gay03] are examples of this development. Recently, Sun Microsystems has advocated the use of Java 2 micro edition as the software platform for sensor nodes. The early commercial off-the-shelf sensor devices used their own nonstandard radios. The IEEE 802.15.4 working group started standardization of the 802.15.4 physical and MAC layer standards for low-rate WPANs in the early 2000s and was completed in 2003. Most sensor nodes of today use the 802.15.4 standard for medium access and radio communications at the physical layer. The Zigbee standard [Whe07] (ratified in 2004) considers standardization of higher layer issues (such as routing, addressing, and application messaging) for embedded sensors, and many commercial vendors have now adopted the standard.

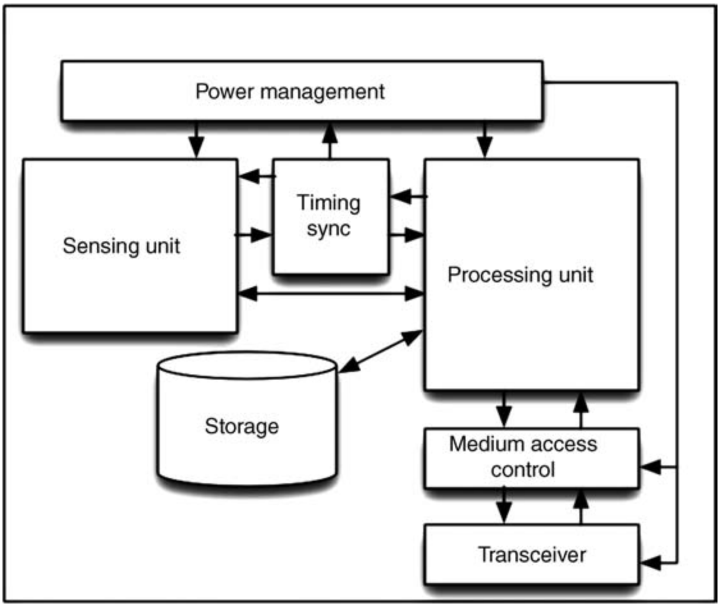


FIGURE 13.2 Simple schematic of a sensor device.

Architecture of a Sensor Device. Figure 13.2 shows a simple schematic of the internal architecture of a sensor device. The interested reader is referred to [Coo06, Whe07, Aky02] for other examples of architectures. In the simple schematic shown here, a sensor device primarily has a sensing unit, a processing unit, and a power management unit. The sensing unit performs the actual sensing tasks (e.g. detecting changes in temperature). The processing unit (with some internal RAM) is responsible for performing computations on the sensed data in conjunction with a storage unit. It is also responsible for handling communications (by running the operating system code) and working collaboratively with other sensor nodes towards accomplishing the application objectives. The power management unit is important because of the necessity to reduce the power consumption in the sensor node to the maximum extent possible. Also shown in Fig. 13.2 are the storage and time synchronization components. Usually, the storage on sensor devices is comprised of flash memory. A MAC unit works with the transceiver to access the shared air interface. Not shown in the figure are potential external sensing units, the battery that actually powers the sensor device, the antenna, or other components, like those used for localizing a sensor (determining its location).

Components and Operation. In the same spirit of general overview, we do not discuss the details of the hardware used to make up sensor devices, except at a high level. We also avoid providing an exhaustive list of chips used in sensors. Some details are available in Section 13.3.4.

The processing units are typically simple. Some sensor devices use custom chips (e.g. the submote-class *spec* sensor [Hil04]), while others employ low-power off-the-shelf microcontrollers from ATMEL [Atmel], Jennic [Jennic], Texas Instruments, or Intel. A common standard 802.15.4-compliant radio from Chipcon (now part of Texas Instruments) is used for the transceiver MAC. Electrically erasable programmable read-only memory or flash memory storage is commonly employed in most sensor devices.

TinyOS is the most common operating system that is run on sensor devices. However, other operating systems are also used in some sensors, such as J2ME, Contiki [Cont], and MantisOS [Mant]. TinyOS was developed at the University of California at Berkeley and has been explicitly designed to support concurrency operations on embedded devices and it is open source. TinyOS has been used in the sensors deployed in several projects, including the monitoring of seabird occupancy on Great Duck Island described in Section 13.2. TinyOS 2.0 is a clean-slate design that foregoes backwards compatibility with the earlier versions to overcome their limitations. MoteWorks is a commercial sensor network software development environment from Crossbow (see Section 13.3.3) that is built using TinyOS 1.1 as the basis. Contiki is an open source operating system with small memory requirements, a microIP stack for TCP/IP communications, and a lightweight communication stack for low-power radios that supports multi-hop communications. MantisOS (also open source) was developed at the University of Colorado at Boulder and is a multithreaded operating system for sensors written in C. Many mote-class sensor devices can run any of the three operating systems.

13.3.3 Commercial Sensors

In this section we consider some example commercial sensor devices, their capabilities, and application environments (recommended by the vendors). We look at a couple of popular vendors and their products and cite appropriate references to the devices. This is not intended to be an exhaustive list, but more of a representative list to illustrate the current state of the art. As we discuss below, most sensor devices available on the market use the IEEE 802.15.4 standard for the lowest two layers of communications (the MAC and physical layers). Heterogeneity exists at higher layers, with different vendors selecting different options, such as Zigbee, WirelessHART, and 6LoWPAN. Some vendor products have the ability to organize themselves into a mesh and perform multi-hop routing to a manager or BS. Finally, there are several customized sensor nodes that have been implemented [e.g. Leo03, Sam06, Ayl07] in academic projects. An exhaustive discussion of such sensor devices is beyond the scope of this chapter.

Crossbow. One of the earliest vendors of mote-class sensor devices, Crossbow [Xbow] now makes a variety of sensor devices and supports applications in industrial automation, environmental or climate monitoring, and asset tracking. Crossbow's catalog consists of evaluation and development kits, end-user mote systems, and gateway or BS-capable sensor nodes. There are numerous types of sensor devices, evaluation kits, boards, and processor/radio units in Crossbow's catalog that are suitable for specific applications or for research and development. Different units have different capabilities (e.g. combinations of ability to sense quantities such as light, pressure, or temperature, GPS, accelerometers, external analog inputs, etc.). Some of the units come with software that enables them to create a self-healing mesh network (the software is called Xmesh by Crossbow). They have support for the commercial software development tool (MoteWorks). Most of the units are IEEE 802.15.4 compliant and operate in the 2.4 GHz ISM unlicensed bands (see Chapter 3 for more details). However, some units also operate in the

888 or 916 MHz unlicensed bands. We briefly consider some products from Crossbow in more detail.

The TelosB mote from Crossbow is especially suitable for research and development with support for TinyOS. It has a Texas Instruments microcontroller with 10 KB of RAM, additional flash memory, two AA batteries, and the ability to collect data using a USB port. This port can also be used for programming the sensor. TelosB operates in the 2.4 GHz bands.

The Cricket Mote includes both an RF transceiver and an ultrasound transceiver. It is an example of a sensor that has two different transceivers. The latter can be used for obtaining ranging information using the time of flight of the ultrasound signal. Ranging information can be used to localize sensor nodes after deployment. It is an enhancement of Crossbow's low-power Mica2 mote, which operates in the 868/916 MHz bands.

The Stargate BS from Crossbow includes a 400 MHz Intel processor, an embedded Linux platform, the ability to communicate using Ethernet or other kinds of interfaces using PCMCIA connectors, and support for communications with mote-class sensor devices.

Dust Networks. Like Crossbow, Dust Networks is a vendor targeting applications such as process control, asset management, environmental monitoring, and health safety monitoring [Dust]. The SmartMesh products from Dust Networks include those specially designed for industrial automation, low power (using system-on-a-chip), and for harsh wireless environments. Most of the products operate using IEEE 802.15.4 radios in the 2.4 GHz ISM bands, although some of the products operate around 900 MHz. Evaluation kits are also available.

The SmartMesh-IA products from Dust Networks, built for industrial automation, also employ the WirelessHART protocol [HART]. Several million intelligent instruments in the process control industry already employ the HART communication protocol for configuration, status checks, and exchanging information using wired networking. The Wireless HART protocol also allows HART-enabled devices to communicate with sensors. It is reliable and secure and employs a combination of mesh networking for redundancy, channel hopping to avoid interference, time-synchronized messaging, and 128-bit encryption and authentication.

Sun. Sun Microsystems, which has been a major player in the computer industry in developing chips, operating systems, and in enabling services, has invested in the so-called "Small Programmable Object Technology" (SPOT) as part of its research and development efforts [Sunspot]. SunSPOT sensors are not being sold with particular commercial applications in mind, but rather as more powerful wireless platforms running a Java virtual machine that can perform complex tasks for use in exotic applications such as swarm intelligence or rocket launch monitoring. Sun makes its sensor nodes available for education, government, military, and industrial applications, as well as for hobbyists. SunSPOT sensors are IEEE 802.15.4 compliant.

Others. Other vendors of sensor networks include Ember Corporation, Sensinode [Sensinode] in Finland, and Sentilla, which is focused on software solutions for sensors using Java. Companies such as Jennic manufacture integrated solutions for Zigbee where an

TABLE 13.1 List of Some Vendors of Sensor Devices and Features Of Some Products

Company	Types of sensor device	Frequency bands	Standards supported	Applications supported
Crossbow	Evaluation kits, BSs, motes, data-acquisition boards	2.4 GHz, 868 MHz, 916 MHz	802.15.4, ZigBee	Asset monitoring, climate control, surveillance
Dust Networks	Evaluation kits, motes, managers	2.4 GHz, 902–928 MHz	802.15.4, WirelessHART	Process control, asset monitoring, health safety monitoring
Ember Corporation	ZigBee chips, software, development tools	2.4 GHz	802.15.4, ZigBee	Home and building automation, asset management, defense applications
Sensinode	Development kits, Nanoseries sensors and routers	2.4 GHz	802.15.4, 6LoWPAN	Hospital asset management

802.15.4 radio is integrated with a microcontroller and software that supports Zigbee. Such products are not marketed as complete sensor networking solutions, but can be used by other companies to develop their own sensor network solutions for specific applications.

Table 13.1 lists some of the vendors, their products, and targeted applications.

13.3.4 Future Directions

In this section we briefly discuss some future possibilities for sensor devices that are being explored in the research literature (and in practice). We first consider the reduction in form factor of sensor devices, which can result in efficient use of computing and energy resources. Then we consider mobility and other issues.

Reducing the Size of Sensor Devices. Sensor devices are becoming smaller and yet more capable with time in a manner similar to computers, which started out as being huge devices in the second half of the twentieth century and whose capabilities are now matched by today's laptops and PDAs. One of the technologies driving the miniaturization of sensors is the ability to fabricate an entire "system" on a single silicon chip. Today's sensor devices include several commercial off-the-shelf components that have been "put together" to create the device. While this reduces the cost of manufacturing sensor devices, such devices do not have the extremely small form factors required for some applications and they are quite inefficient in the use of power and other resources. If all the digital (processing and communication), analog (typically sensing), and RF (communication) components can be integrated onto a single chip, then the form factor of sensors can be reduced significantly. For example, some of the SmartMesh products claim to be more efficient than other comparable 802.15.4 devices because of tight integration of components on a chip. The feasibility of a single-chip sensor mote is discussed by Cook *et al.* [Coo06]. Issues of integrating elements such as the antenna for wireless communications and quartz crystal for timing and synthesis of high-frequency signals are challenges that are difficult to overcome when considering integration of RF components on a chip. The sensing capabilities may also be hard to integrate with the other components on a single chip, as also is the energy source.

At the extreme, sensor devices are expected to be of the size of specks of dust. Nanosensors employing nanotechnology are considered by Sailor and Link [Sai05] in their discussion of sensor devices that are really of the size of dust motes. We recall that a sensor has to perform sensing, processing, and communications. Consequently, identifying or addressing a sensor device is important. Sailor and Link [Sai05] discuss the technique of etching spectral "bar codes" (that can be scanned using lasers at distances of up to 20 m) on porous silicon dust for identifying sensors the size of dust. Similarly, sensing with such small form factors is difficult, but it is possible for some applications, such as sensing the presence of biological molecules. Processing and communications are also potentially possible at the nanoscale.

Mobile Sensors. Deploying sensors that can move to better locations for sensing or for delivering information can benefit applications that deploy a large number of sensors. Mobile sensors can improve the application performance as well as improve the efficiency of deployment. Mobility, however, is costly, especially since it is difficult to guide, monitor, and modify the movement of sensor devices. Typically, only large sensors are mobile and employ robots for moving the sensor devices, although efforts have been made to reduce the size of such robots [Ber00, Sib02]. Alternatively, BS-class sensors can be made mobile so that they can move around the deployment area to gather data from mote-class or submote-class sensors. Such sensors have been called "data mules" in the research literature [Sha03].

The real challenges occur when submote-class sensors or sensors that are the size of dust specks need to be mobile. As discussed by Sailor and Link [Sai05], there are great challenges in enabling autonomous mobility of sensors that are the size of specks of dust, even if the mobility is random. It is possible, however, to enable *directed* mobility, where sensors can be made to move in specific directions using the environment in which they are deployed or external mechanisms or forces such as electrical, magnetic, or photonic stimuli. Work in all areas related to making extremely small sensor devices mobile is in the nascent stages. If and when mobile sensors that are of the size of dust specks become reality, one can expect several new applications to emerge.

Other Issues. An important issue with sensor devices owing to their small size, limited resources, and reduced functionality is the ability to maintain and verify consistency of applications and reprogrammability of the applications or retasking the sensors as needed. Such needs may include changing some applications in response to the environment and maintaining scalability as more sensors are deployed. A short survey of software approaches for sensor devices is presented by Hadim and Mohamed [Had06]. Recent work is focusing on improving and enabling such reprogrammability of sensors in an efficient manner. For example, a dynamic operating system SOS is described by Han *et al.* [Han05] that has a common kernel for all sensor nodes. Applications can be dynamically loaded at run time in SOS. This is different, for instance, from TinyOS, which is statically compiled. Consequently, the entire system image must be redistributed if any changes need to be made in the applications, making it expensive in a large-scale sensor network.

13.4 THE PHYSICAL LAYER IN SENSOR NETWORKS

In sensor networks, the PHY layer has to be simple for several reasons. The cost of fabricating a sensor should be low if several hundreds or thousands of devices have to be deployed to satisfy application objectives. At the same time, a sensor has to have a small form factor, implying that the transceiver cannot be made very complex. Chapters 2–4 provide a detailed treatment of radio propagation, modulation, and error control coding. In this section, we discuss aspects of these in relationship to sensor networks.

13.4.1 Spectrum

The characteristics of radio propagation depend upon the frequency of operation, as discussed in Chapter 2. At frequencies beyond 500 MHz or so, propagation of radio waves can be approximated as if they are optical rays propagating from the transmitter. While this somewhat simplifies theoretical analyses and simulations, the fact remains that radio propagation is extremely site specific and frequency specific. In general, the higher the frequency is, the larger is the attenuation of the signal strength with distance because the larger is the absorption in the surrounding medium and the larger is the loss in the antennas.

Because sensor networks are expected to be deployed by a variety of organizations for a variety of applications, it is beneficial to employ *unlicensed* frequency spectra for transmissions. The use of such spectrum does not require permission from the FCC, nor widespread testing to ensure that transmissions do not interfere with other licensed applications. Instead, a reasonable spectrum etiquette that allows low-power transmissions and multiple operators to coexist is implemented. The unlicensed spectra of choice for sensor networks are the 868 MHz band in Europe, the 916 MHz band in the USA, and the 2.4 GHz band that is available almost everywhere in the world. The 2.4 GHz band has the widest available bandwidth and is becoming the most popular choice for sensor networks. The only downside to this choice is that IEEE 802.11 WLANs, cordless phones, Bluetooth, and many other wireless devices also operate in these bands, resulting in significant interference in areas of dense deployment. Recently, the IEEE 802.15 standards committee's task group 4c has been considering standards for the 779–787 MHz bands in China and task group 4d is looking at the 950–956 MHz bands in Japan.

The base IEEE 802.15.4 standard that was approved in 2003 and extended in 2006 specifies three channels in the 868 MHz band, 30 channels in the 916 MHz bands and 16 channels in the 2.4 GHz bands (see later discussion for clarification on channels). Originally, the data rates were dependent on the frequency bands of operation. The lower frequencies supported 20 kb/s and 40 kb/s, while the 2.4 GHz bands supported up to 250 kb/s. Recent revisions to the standard enable higher data rates in all bands and an option for 100 kb/s in the 868 MHz band.

13.4.2 Path Loss

One of the most important parameters for design and operation of communication networks is the *transmission range* – essentially the distance up to which reliable communications is possible between a transmitter and a receiver. The *path loss* (see Chapter 2) or the reduction in signal strength as a function of distance is an important parameter that determines how far apart two sensor devices can be and still have reliable communications.

As discussed in detail in Chapter 2, in free space, where there are no obstacles, and with ideal isotropic antennas (antennas that transmit with equal power in all directions), it is possible to demonstrate theoretically that the signal strength drops as the square of the distance. In other words, if the transmit power is P_t and the received power is P_r, then the relationship between them is $P_r = L_0P_t/d^2$, where d is the distance between the transmitter and receiver and L_0 is a constant that depends upon the frequency of transmission. The path loss in this case is $P_t/P_r = d^2/L_0$. However, in reality, there are obstacles in the environment and, depending on the site, the drop in the signal strength with distance can be very different. A general model for path loss that is widely used in the research literature uses a general exponent n to define the drop in signal strength. That is, the relationship between the transmit power and received power is represented by

$$P_r = \frac{L_0P_t}{d^n} \tag{13.1}$$

Writing this in decibels, the relationship becomes

$$P_r(\text{dBm}) = P_t(\text{dBm}) + L_0(\text{dB}) - 10n\ \log_{10}(d) = P_t(\text{dBm}) - L_p(\text{dB}) \tag{13.2}$$

The path loss is given by $L_p = 10n\ \log_{10}(d) - K$ (dB). Note that the factor of 10 appears in Eq. (13.2) because the units are in decibels, not bels. Transmit and received powers are usually written in dBm, which expresses the power in comparison to 1 mW. The value of n has been empirically determined for a variety of types of sites and transmission frequencies in cellular networks (e.g. big cities, rural areas, hilly areas) and WLANs. In sensor networks, the placement of the sensor determines how n may behave. Sensors are typically places close to the ground and sometimes are in very cluttered environments. It is not uncommon to find values of n that may range from 2 to 6 depending on the environment.

Example 13.4: Transmission Range for Communication between Sensors Let the transmit power of a sensor be 1 mW (0 dBm) and the receiver sensitivity be −94 dBm. This implies that the received signal strength or received power must be at least −94 dBm, and in many cases several decibels higher for reliable reception. In the IEEE 802.15.4 standard, the receiver sensitivity is defined as the signal strength that results in a packet error rate that is smaller than 1% for a given size of a PHY layer packet. If the constant L_0 (sometimes called the loss at the first meter) is −45 dB at 2.4 GHz, then the distance-dependent loss can at most be $94 - 45 = 49$ dB. If the signal strength drops as the third power of distance (i.e. $n = 3$), then $d = 10^{49/30} = 42$ m. In other words, the transmission range is

42 m. This compares with an indoor range of 30 m specified for a comparable vendor product. Other experiments have shown that, in outdoor areas with little clutter, the range can be as large as 75 m.

The range could be reduced due to a variety of factors. Experiments have shown that humidity, obstacles in the sensor field, the environment (e.g. wooded areas, bamboo plantations, sandy beaches can have different values of *n*), and in some cases the directionality of sensors (because their antennas as not really omnidirectional) have an impact on the transmission range. One experiment showed that the transmission range of Mica2 motes dropped from 55 m to 10 m in the presence of fog (or rain) in an outdoor area. This could have a serious impact for applications like climate monitoring.

13.4.3 Gray Zone

In sensor networks, like other communication networks, it is the higher layer performance that is more crucial than simpler lower layer metrics like transmission range. The question in most cases is what the packet transmission reliability (or packet success ratio) is. In most of the simulations or analysis, it is common to use the transmission range as a hard threshold. All communications within this distance are assumed to be 100% reliable and all packets that are received beyond the transmission range are assumed to be received incorrectly. If the sensor network is modeled as a graph, then this is also called the *unit disk model*.

Experiments in real sensor networks over the last few years have shown the presence of a significant *gray zone* where a model with such a hard threshold does not work well. Consider Fig. 13.3, which shows an illustrative plot (not from real data) of the packet success ratio as a function of distance between a sensor transmitter and receiver. Clearly, there is a close to 100% chance that the packet is received successfully when the

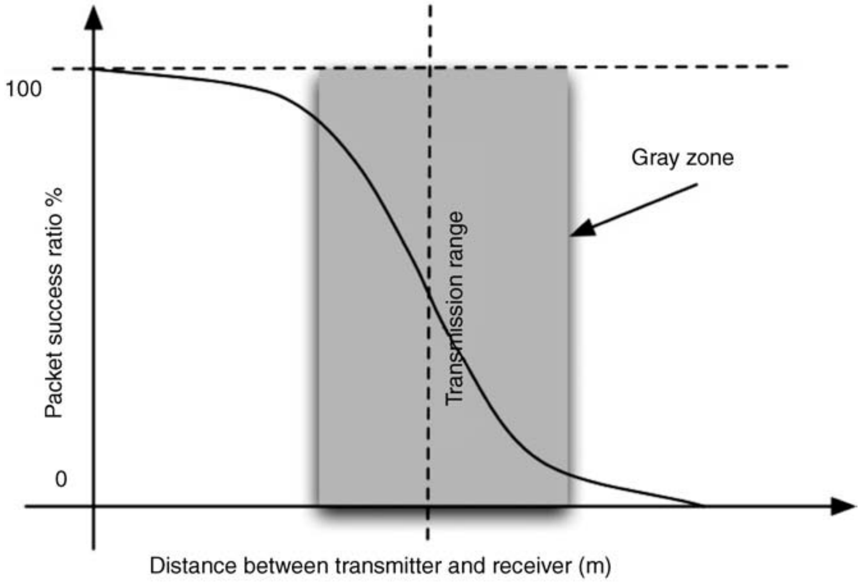


FIGURE 13.3 Illustration of the gray zone.

transmitter and receiver are very close. As expected, a packet has no chance of successfully being delivered when the receiver is very far away from the transmitter. However, in between, the packet success ratio varies from something less than 100% to something greater than 10% or so. In most experiments, this region, where the packet success ratio is not really predictable, has been found to be of significant size. In many cases, it is found to be dominating. Unfortunately, unlike Fig. 13.3, it is not really easy to predict what the success ratio will be for a given transmitter–receiver separation, and this is very site specific.

Example 13.5: Link Stability in the Sensor Network on Great Duck Island In the Great Duck Island sensor network [Sze02a], the stability of links was examined by looking at how long links existed and how often a parent node changes (the node to which data is delivered). Most links were found to have small life spans, but the stable links were used by the network to deliver most of the data. Up to 80% of the data was delivered using 20% of the links. While 55% of the time the parent node did not change, there were some occasions when the entire network topology would be modified.

Experiments have shown that the gray zone exists not only for sensor networks, but also for IEEE 802.11 WLANs. Some experiments have shown that the gray zone is less prevalent at higher data rates in IEEE 802.11 WLANs. The packet size impacts the size of the gray zone (usually, larger packets are more prone to errors). The transmission scheme and the receiver complexity and circuitry also impact the size of the gray zone, as it has an effect on how reliably the receiver can detect the transmitted data. The experimental performance of motes and IEEE 802.15.4 and IEEE 802.11 devices in several settings and scenarios can be found in the literature [e.g. Fan03, Ana05, Pet06].

Finally, the packet success ratio depends on interference. When sensor networks coexist with other technologies employing the same spectrum, the performance can drop significantly. Also, transmissions between a given pair of sensors can impact transmissions between another pair of sensors at the same time, depending on where the sensors are located.

13.4.4 Modulation Schemes

Some of the earliest sensors used on–off keying, as this modulation scheme is simple and easy to implement, both in the transmitter and receiver. The IEEE 802.15.4 PHY layer uses more complicated quasi-orthogonal sequences in a modification of orthogonal modulation. We discuss below the IEEE 802.15.4 PHY layer primarily from a conceptual standpoint. We ignore intricate details of framing and exact fields and formats in our treatment (see Chapter 10 for some of these details). Details of the IEEE 802.15.4 standard are available in [IEEE06]. The PHY layer in most standard technologies is responsible for functions other than simply transmitting and receiving data. Some of these functions are in support of the protocols. The 802.15.4 PHY layer, for example, also handles activation and sleep of the transceiver, detecting energy on the air for collision avoidance, providing some estimate of the quality of the link, and selecting appropriate frequency channels.

The standard defines *channel pages* and *channel numbers*. Channel pages are used to distinguish between the different possible modulation schemes. The 2.4 GHz band is associated with a single channel page (0) and has 16 channel numbers, while the 868 MHz and 916 MHz bands are associated with three channel pages, 0, 1, and 2. The 868 MHz band has one channel number and the 916 MHz band has 10 channel numbers for a total of 3 and 30 channels respectively in these bands. However, note that, at a given time, only one possible (channel page, channel number) combination is possible. Thus, in reality, there is only a single channel in the 868 MHz band and 10 channels in the 916 MHz band. However, these channels can support different modulation schemes and, thus, different data rates. In the 868 MHz band, the channel is 0.6 MHz wide, in the 916 MHz bands it is 2 MHz wide, and in the 2.4 GHz bands it is 5 MHz wide.

In each channel, irrespective of the band, the bits that are received from the higher layers are first converted to symbols (if necessary), which are then mapped to the sequence of chips and finally the chips are modulated on a carrier. As an example, consider the 2.4 GHz bands. Here, an octet is broken into groups of 4 bits. Each group of 4 bits (there are 16 possible combinations) – now a symbol – is mapped to one of 16 quasi-orthogonal sequences of length 32 chips. The chips are then modulated using a 4-ary modulation scheme called MSK. MSK is a modified form of QPSK with a half-sine pulse shape. The modulation scheme is more bandwidth efficient than QPSK, with a rectangular pulse shape. The symbol rate is 62.5 kS/s (with 4 bits/S, the bit rate is 250 kb/s). Each symbol is converted into a 32-chip sequence. Consequently, the chip rate is 2 Mc/s.

In the 868/916 MHz bands, two base modulation schemes are allowed. If BPSK is used as the base modulation scheme, then the data rates supported are 20 kb/s and 40 kb/s respectively in the two bands. Each bit is first mapped onto a sequence of 15 chips (derived from a maximal length sequence) and each chip is transmitted using BPSK with raised cosine pulse shaping. The chip rates for the two bands are 300 kc/s and 600 kc/s respectively. With 15 chips/bit, this results in the corresponding data bit rates. An optional base amplitude-shift keying (ASK) modulation scheme allows the data rates to be increased by using a form of CDM (where bits are sent in parallel by spreading them using almost-orthogonal sequences).

The receiver sensitivity in the 2.4 GHz bands is specified to be at least −85 dBm. For BPSK modulation in the 868/916 MHz bands, the receiver sensitivity should be at least −92 dBm or −85 dBm with ASK modulation.

13.5 THE MAC LAYER IN SENSOR NETWORKS

Like the physical layer, the MAC layer cannot make use of very complicated algorithms since they have to reside in each sensor and sensors are expected to be low-cost devices. In wireless sensor networks the shared medium is air, which makes the design of MAC protocols more challenging due to interference and lack of reliability. This section has to be considered in conjunction with Chapter 5, which discusses the MAC layer in detail. In this section we discuss MAC-related issues for sensor networks, followed by a description of the IEEE 802.15.4 MAC. A good tutorial on IEEE 802.15.4 is the article by Callaway *et al.* [Cal02] in the *IEEE Communications Magazine*. Then we present an overview of MAC protocols developed in the research literature that focus either on energy efficiency or latency. A good survey of MAC protocols for sensor networks is that by Demirkol *et al.* [Dem06], which also points to references to some of the MAC protocols discussed in this section, like sensor MAC (SMAC) and its variants, Sift, and the traffic-adaptive medium access (TRAMA) protocol. The design of MAC protocols typically considers metrics such as throughput, bandwidth utilization, fairness in providing access to the medium, latency, scalability, and efficiency or control overhead for evaluating the MAC protocol. For sensor networks, the metrics for evaluating a MAC protocol are often different, and we explore these briefly next.

13.5.1 Issues in Medium Access for Sensor Networks

Unlike MAC protocols that are designed for fairness or efficiently utilizing the bandwidth, in sensor networks the two issues of importance are energy efficiency and latency.

Energy. Energy efficiency is crucial because of the scale and application environments in which sensors are deployed. Sensors are expected to be of extremely small form factor and several thousands of them may be deployed in a sensor field. The sensor field may not be accessible to human beings; and even if accessible, it may not be practical to replenish batteries in several thousand sensors on a regular basis. For most applications, it is expected that sensor devices will be in sleep mode (as far as communication goes) for a majority of the time and will be awake only periodically or as events occur to transmit information. The low duty cycle of sensors must be exploited by MAC protocols.

Some of the common sources of energy waste in MAC protocols are collisions, unnecessary reception of data, idle listening, and control overhead. Collisions occur when two nodes transmit at the same time and both transmissions (whether intended to a receiver or not) arrive simultaneously at a receiver. The receiver cannot recover the transmission. Energy is wasted for the transmission and for the attempted reception. This is a bigger problem with carrier-sensing-based MACs (see similar discussion in Chapter 9 with respect to WLANs). Suppose all sensor nodes are identical and have identical transmission and reception ranges, as shown in Fig. 13.4. The transmission from sensor A can be heard by

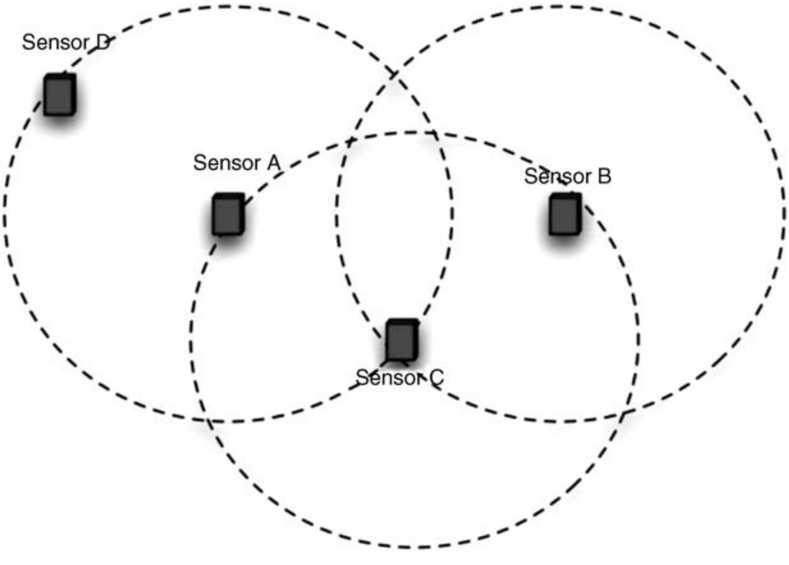


FIGURE 13.4 Illustrating hidden nodes, collisions, and exposed nodes.

sensor C but not sensor B. So, when sensor A is transmitting a frame to sensor C, sensor B will not sense the channel as busy and sensor A is *hidden* from sensor B. If both sensor A and sensor B transmit frames to sensor C at the same time, then the frames will collide. This problem is called the hidden terminal problem. There is a dual problem called the exposed terminal problem. In this case, sensor A is transmitting a frame to sensor D. This transmission is heard by sensor C, which then backs off. However, sensor C could have transmitted a frame to sensor B and the two transmissions would not interfere or collide. In this case, sensor A is called an *exposed node*. Both hidden and exposed nodes cause a loss of throughput. Since sensor C is listening to sensor A's transmission to sensor D (which is of no consequence to sensor C), it is engaging in wasteful overhearing, which leads to energy waste. In most carrier-sensing-based MAC protocols, there is some amount of idle listening to ensure that the medium is free for access. This wastes energy as well. Finally, in trying to reduce the hidden node problem or in scheduling-based MACs, nodes exchange short control packets to announce their presence, impending transmissions, or schedules. These control packets do not contribute to the application objectives directly and waste energy.

Latency. Latency is an important issue in several applications where the response from a group of sensors needs to be received at the sink (BS or gateway) quickly after the occurrence of an event so that appropriate action can be taken to mitigate the impact of an event. For example, intrusion detection requires alarms to be delivered without significant latency to the BS. If the vibration data in a section of a bridge changes such that it makes it necessary to clear vehicular traffic rapidly and close the bridge, such alarms must be received in a timely manner. In such cases, it is important that the MAC protocol is not a bottleneck for the delivery of data to the sink.

Data Gathering Tree and Energy/Latency Issues. In many sensor applications, sensors deliver data towards a sink in a tree-like manner, as shown in Fig. 13.5. Sensors at the highest level (level 3 in Fig. 13.5) send their collected data to sensors at the next lowest level and so on till the data is delivered at the BS or sink. This causes additional problems in terms of

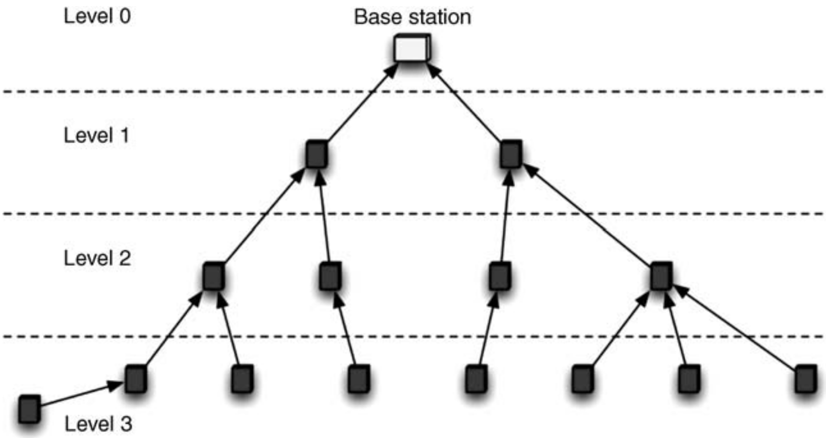


FIGURE 13.5 Tree-like data delivery.

disproportionate energy consumption in nodes that are closer to the BS (e.g. those at level 1) that are required to transmit not only the data sensed by themselves, but also the data that is sent to them from nodes at higher levels. This also causes congestion closer to the BS, since more data needs to be transmitted there compared with the transmissions at the higher levels.

13.5.2 IEEE 802.15.4 Medium Access Control

As mentioned in Chapter 10, the IEEE 802.15.4 standard specifies FFDs and RFDs. An RFD can only communicate with an FFD, whereas an FFD can communicate with RFDs or other FFDs. A network with two or more devices is called a WPAN. A WPAN can be of two topologies: a star topology, where one FFD acts as the WPAN coordinator and controls other RFDs and FFDs in the network, or a peer-to-peer topology, where FFDs and RFDs exist and can communicate with one another as long as they are in range and have the correct functionality (e.g. two RFDs cannot communicate with each other directly). The formation of a network is part of higher layer issues discussed briefly in Section 13.6. However, each WPAN is assumed to have a WPAN coordinator which is an FFD.

The MAC protocol in 802.15.4 operates under a superframe structure (see Fig. 13.6). A superframe is defined as the period between two *beacons*, which are special management packets transmitted by the coordinator. Beacons synchronize the WPANs and provide information about the network. Within the time between two beacons, i.e. the superframe, sensor nodes can have an active period and an inactive period. The active period can be divided into a contention period and a contention-free period. The contention period is slotted. In each slot, nodes use CSMA/CA to access the channel. The process is quite simple. Each device waits for a random period to see whether the channel is idle. If it is idle, then it simply transmits. Otherwise, it backs off for another random period and tries again (see bottom of Fig. 13.6). This access suffers from the disadvantages mentioned previously

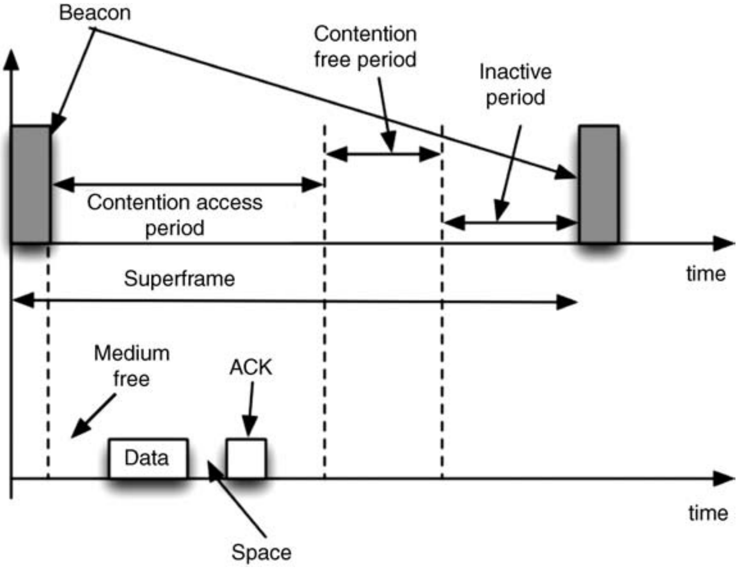


FIGURE 13.6 Illustration of medium access in IEEE 802.15.4.

with carrier sensing. Scheduled access is possible through the use of the contention-free period, where the WPAN coordinator creates guaranteed time slots, which nodes can use without contention from other nodes. It is possible to have a WPAN without beacons, in which case nonslotted CSMA/CA is adopted by all nodes.

The standard considers "transactions" that are initiated by the low-power devices, which will otherwise have the choice to be in a low-power mode. In the star topology, nodes can send data to a coordinator or receive data from a coordinator. In the former case, the node simply sends data to the coordinator using CSMA/CA and gets an acknowledgment if requested. In the latter case, the node should first request data from the coordinator, get an acknowledgment for its request, followed by the data. Upon receipt of the data, it acknowledges it to the coordinator. Acknowledgements do not wait for the medium to be idle as in the case of data frames. Alternatively, the beacon can have information about whether or not there is pending data for a sensor node (see related discussion of SMAC and its variants next). Peer-to-peer transmissions occur in a similar manner, except that special steps may be necessary for synchronizing transmissions of two peer nodes.

In order to allow the MAC layer to process frames, there must be some time that elapses between successive frame transmissions. The time between receipt of a frame and the transmission of an acknowledgment is the smallest, while long and short IFSs are used to separate long or short frames.

The MAC layer in IEEE 802.15.4 is also responsible for starting and maintaining WPANs (scanning through the various PHY channels, recognizing IDs in the beacons), synchronization with the WPAN coordinator, association and disassociation with a WPAN, allocation of guaranteed time slots, and frame security (encryption and authentication).

We next briefly consider some MAC protocols that have been proposed in the literature, again from a conceptual standpoint. The reader is referred to references previously mentioned in the chapter for additional details.

13.5.3 Low-Duty-Cycle Medium Access Controls

As mentioned previously, energy waste in sensor networks occurs due to the fact that there are collisions, unnecessary overhearing, and wasted awake times when sensor nodes are listening to the channel but not receiving anything. This is especially a problem when the sensors actually transmit or receive data very infrequently (low duty cycles). Random access schemes that try to address this problem are described below.

Researchers at the University of Southern California proposed SMAC, which looks at this issue specifically rather than issues such as fair access to the medium (when all nodes get equal opportunities to transmit data). Several variations of SMAC have been proposed since, including *timeout* MAC (TMAC) by researchers in Delft University and dynamic SMAC by the University of Buffalo. The underlying mechanism behind SMAC and its variants is still CSMA/CA. The idea behind SMAC is that sensor nodes can form virtual clusters where they are loosely synchronized. This synchronization can help them coordinate their sleep cycles so that they do not have to be awake when they are unlikely to receive any transmissions. Note that this is a distributed approach compared with the approach taken in IEEE 802.15.4, which requires nodes to ask for data from the coordinator or receive information about pending data in the beacons. Nodes that are in the vicinity of two virtual clusters may have to follow two different sleep schedules in SMAC. SMAC also supports message passing, where a sensor node occupies the medium until it has completed transmitting a message (which may be fragmented into frames). This may make the medium unfair, but avoids the unnecessary waiting times in typical CSMA protocols. Moreover, SMAC supports the use of RTS and CTS control frames that ameliorate the hidden node problem by announcing the impending data transmissions in the vicinity of *both* the sender and the receiver nodes, like the IEEE 802.11 protocol. The RTS/CTS frames also indicate the size of transmission, allowing nodes to sleep longer if necessary. TMAC introduces a variable sleep cycle for nodes, compared with SMAC, to improve upon the energy consumption in the presence of variable loads. Latency, which could be high in SMAC, especially for multi-hop transmissions (e.g. in Fig. 13.5, the time taken for a packet to reach the sink may be very high if the sleep schedules of the nodes in different levels are arranged in a worst-case manner), can be reduced by halving the sleep schedule of some nodes which experience high latency and yet maintain synchronization with other nodes.

Asynchronous MAC protocols for low-duty-cycle sensor nodes rely on a mechanism called “low-power listening,” where sensor nodes sample a (periodically) transmitted preamble to see whether there is data that is intended for them. If so, the nodes wake up; else they return to sleep. The low-power listening reduces the power to detect intended transmissions; but there are other disadvantages, where nodes should stay awake to listen to the preamble whether or not the preamble indicates a transmission towards them, and latency is a problem, as data transfer can occur only after the preamble is completed even if a receiving node is awake at the beginning of the preamble. As in the case of SMAC, latency accumulates over multiple hops. The Berkeley MAC from the University of California at Berkeley and X-MAC from the University of Colorado are examples of asynchronous MAC protocols that rely on low-power listening. The X-MAC protocol employs a shorter preamble to reduce both energy consumption and latency.

13.5.4 Low-Latency Medium Access Controls

A second group of MAC protocols for sensor networks considers latency as the issue of interest.

Of these, so-called *event-driven* MACs are still based on carrier sensing and random access. When an event occurs, it is likely that sensors that detect the event will start a flurry of transmissions, many of them spatially correlated, that attempt to deliver data towards a sink, all at times very close to one another. In many applications, it is not necessary that all of this data be delivered reliably and correctly to the sink. It is sufficient if a subset of sensors from a spatially correlated set delivers data reliably and correctly to the sink. This observation is exploited in *Sift*, a MAC protocol developed at the Massachusetts Institute of Technology. Again, fairness in channel access is a sacrifice for improvement in latency performance. In typical CSMA/CA MAC protocols, it is common for nodes to pick a slot within a contention window to transmit. The slot is picked randomly, from a uniform distribution, so that there is roughly fair access for all nodes. The node with the earliest slot gets to transmit, while others back off and wait for their chance to transmit in succession. The contention window expands with time in the case of collisions to alleviate the problem. In Sift, the probability of picking a slot is picked from a nonuniform distribution within a fixed contention window, which changes depending on whether or not a node observes a transmission in the earlier time slots. If no transmissions are seen, then a node increases the probability with which it will transmit in the coming time slots. Sift has a much better latency performance than standard IEEE 802.11-based CSMA/CA.

Finally, we mention TDMA-like MACs that attempt to schedule transmissions of sensors to prevent collisions and achieve gains in latency. The TRAMA protocol, which is itself a modification of node activation multiple access (NAMA) to better suit sensor networks, considers a two-hop neighborhood of sensor nodes to schedule transmissions. The schedules ensure that there are no collisions between transmissions (which waste energy). The schedules also indicate the intended receivers, allowing other nodes to go to sleep when no packets are expected for them. In the *neighbor protocol* phase, sensors exchange information about their two-hop neighborhood. In the *schedule exchange protocol* phase, nodes exchange traffic information and schedules. Then, an *adaptive election algorithm* picks the transmitters and receivers to ensure collision-free transmissions. The first two phases make use of a contention access period (like CSMA) to enable exchange of information.

13.6 HIGHER LAYER ISSUES IN SENSOR NETWORKS

In previous sections we provided an overview of sensor devices and the lowest two layers (namely the PHY and MAC layers) of sensor networks. However, sensor networks are *application-oriented networks* and the operation of the higher layers plays an important role in realizing application objectives. For a sensor network to be operational, it is necessary for nodes to self-organize into a network to establish a service and have protocols in place that enable routing of sensed and processed data from source sensor nodes to destinations (sinks). Instead of transmitting all of the sensed data towards the sink, sensor nodes can save communication and energy costs by performing in-network processing and aggregation of data. How sensor nodes should be deployed to *cover* a region that needs to be sensed such that areas do not suffer gaps nor the resulting deployment cause partitioned networks is a challenging issue. Many applications need sensors to have some ability to determine their location (also called localization – see Chapter 12 for more details). Several MAC and routing protocols, as well as applications, need synchronous operation of sensor networks, requiring some protocols for enabling synchronization between nodes and handling clock drifts. Finally, it is necessary to keep sensors and their information secure.

A large number of research papers have been published over the last decade addressing these issues and, in some cases, the impact that these issues have on one another. For example, some MAC protocols, not discussed in this chapter, consider joint MAC and routing. As another example, there are protocols that look at routing correlated data towards a sink. It is extremely difficult to survey all of these papers in a short chapter devoted towards background information. Thus, the objective of this section is to provide the reader with an idea of some of the issues related to higher layers in sensor networks and to provide information about some common protocols and research efforts briefly. The references described next provide pointers to many surveys and tutorials that address these topics in greater detail.

The chapter on sensor networks in Siva Rama Murthy and Manoj [Srm04] discusses many of the topics considered in this chapter, such as routing, establishment of the network, and coverage. Recent developments in node clustering for sensor networks are described in Younis *et al.* [You06]. A treatment of Zigbee is presented by Wheeler [Whe07], including random addressing and routing. The paper by Woo *et al.* [Woo03] describes MintRoute, the routing protocol used in many real implementations of sensor networks (such as the structural health monitoring network described by Kim *et al.* [Kim06]). Al-Karaki and

Kamal [Alk04] provide a comprehensive survey of routing protocols for sensor networks. A survey of coverage and connectivity issues in wireless sensor networks is available in Ghosh and Das [Gho08], and a thorough treatment of topology control is presented by Santi [San05]. Our treatment of the issues of coverage is heavily influenced by the latter two papers. A survey of synchronization protocols for sensor networks is available in Sivrikaya and Yener [Siv04]. The issue of localization in ad hoc/sensor networks has been discussed in many papers [e.g. Sav01, Bul00, Nic03]. Stinson [Sti03] is a good reference on encryption/ authentication schemes and the AES. Details of the security process in IEEE 802.15.4 are available in [IEEE06].

13.6.1 Establishing the Sensor Network

When sensor nodes are first deployed in a sensor field, some type of organization is essential to ensure that the network operates smoothly. In this section we provide a quick overview of some of the approaches implemented or proposed for sensor networks.

In the case of the IEEE 802.15.4 standard (along with the ZigBee higher layer protocols), this is accomplished through the use of WPAN coordinators. As mentioned in Chapter 10, two different topologies are possible. In the star topology, nodes communicate directly to a WPAN coordinator and all communications go through the coordinator. In many ways, this is like nodes communicating to an AP in WLANs. In the peer-to-peer topology, devices can communicate directly with one another as long as they are in radio range.

In the star topology, as shown in Fig. 13.7, an FFD (see Chapter 10 for information on FFDs and RFDs), when deployed, becomes the WPAN coordinator. This is accomplished simply by the device, by transmitting a beacon and announcing itself as the coordinator if it does not hear any other device when it powers up. In the peer-to-peer topology, a similar mechanism works and there is a WPAN coordinator (usually the first FFD to power up). However, devices may communicate directly with one another where allowed. If two or

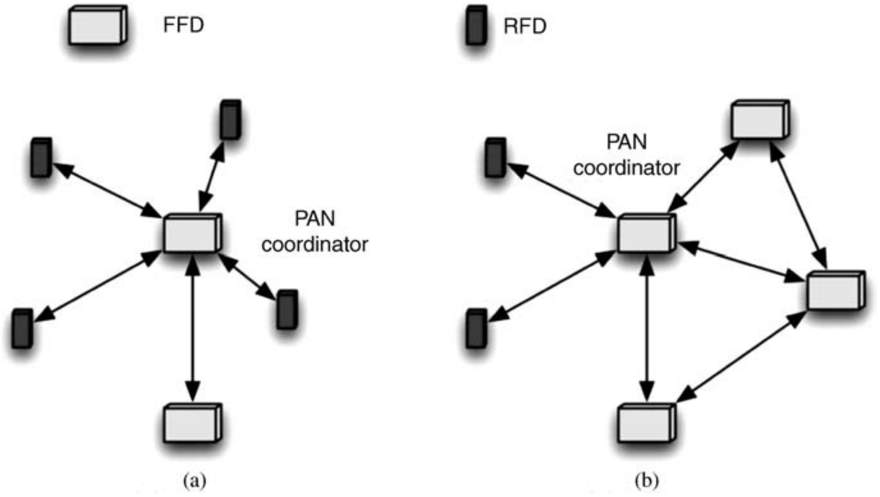


FIGURE 13.7 Star and peer-to-peer topologies in IEEE 802.15.4.

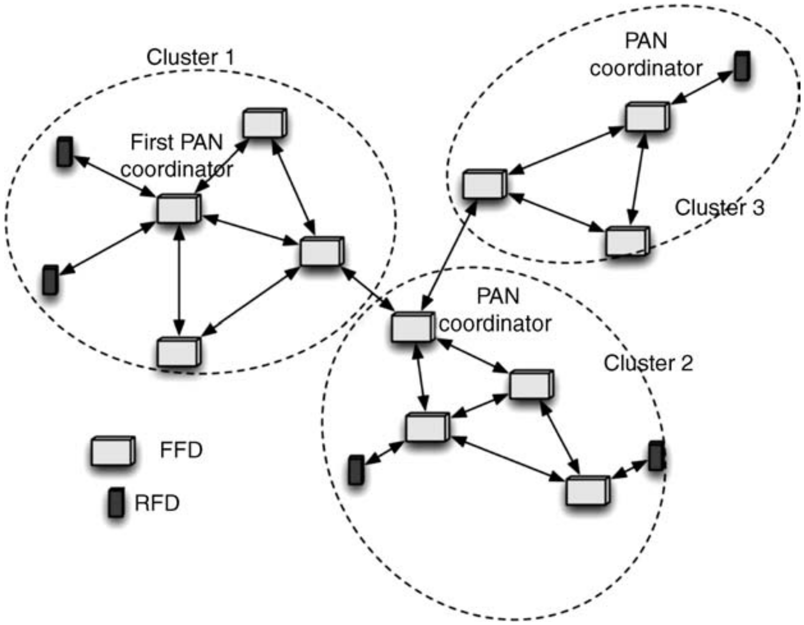


FIGURE 13.8 Cluster tree formation in IEEE 802.15.4 WPANs.

more FFDs attempt to become coordinators, then some contention resolution beyond the scope of the IEEE 802.15.4 standard becomes necessary.

A *cluster tree* is a generalized form of the peer-to-peer topology in IEEE 802.15.4 networks. Figure 13.8 shows an example of a cluster tree with three clusters. Here, the assumption is that most of the devices are FFDs (although RFDs can connect to clusters as leaf nodes). The first FFD that announces itself (or a device with more power or capabilities) becomes the overall WPAN coordinator. It transmits beacon frames and provides other nodes (and other coordinators) synchronization. As the network grows, the overall WPAN coordinator instructs another device to become a WPAN coordinator of its own cluster. Clusters can develop in this way into a large network. The standard specifies resolution of conflicts between WPAN IDs, transmitting beacons, etc.

Clustering is an approach that is also suggested in the research literature for enabling communications between sensors in a field and the BS. Groups of sensors create clusters with cluster heads. These cluster heads are responsible for communicating information from sensor nodes in their clusters to the BS, perhaps through other cluster heads. While the proposed approaches in the research literature are not very specific as to the functionality of the cluster heads (e.g. they are not like WPAN coordinators), in order to distribute the load across sensors in the clusters, cluster heads are periodically changed by an election process. This improves the number of nodes in the network that are still alive after a given time. One popular protocol that changes cluster heads periodically is low-energy adaptive clustering hierarchy (LEACH; see Section 13.6.2). Several other clustering approaches have been suggested (e.g. picking nodes with highest degree as cluster heads to improve connectivity, or picking them among nodes that provide redundancy of coverage).

A second approach to initialization and establishment of the network is a flat architecture consisting of levels (see Fig. 13.5, for example). A BS could broadcast a signal containing its ID which is received by all sensors in the field. This is because the BS is powerful and can transmit at a high power. Next, all sensors respond to this broadcast with their IDs. The BS only receives the responses from nodes that are in level 1 (those that can directly communicate with the BS using a single hop). The BS broadcasts a list of these nodes, enabling nodes that are in level 1 to recognize their level. Next, nodes that are not in level 1 transmit their IDs. Level 1 nodes that receive these IDs forward them to the BS, which broadcasts a list of the new nodes as belonging to level 2, and so on. This way, nodes can determine where they belong and how many hops from the BS they are away.

One of the issues with sensor networks is the challenge in providing addresses to sensor nodes, as they may be several hundred or thousands in number, and careful deployment of addresses in nodes is quite challenging. The ZigBee Pro higher layer standard that operates on the IEEE 802.15.4 lower layers handles this issue by allowing new nodes that join a network to randomly pick 16-bit addresses. With more than 60 000 addresses, the expectation is that collisions will be negligible.

13.6.2 Routing

Routing in sensor networks has received special attention in the research literature. The number of routing protocols that have been proposed makes it impossible for a survey in a short overview chapter like this one. Moreover, many of these routing protocols are not actually implemented in real sensor networks as of now making them mostly of academic interest. The objective of this section is to provide a short summary with pointers for further reading.

Routing in ad hoc networks is typically classified into proactive and reactive types. In proactive routing protocols, nodes exchange information periodically and maintain routes to all possible destinations in the network (i.e., they have information about the network topology). In reactive or on-demand routing, nodes only try to find routes to those destinations to which they are interested in transmitting data. This may be accomplished just before data has to be sent by sending route requests to neighboring nodes that forward these on till the destination is reached. In sensor networks, often times, the data is directed towards a sink from many source sensor nodes. On occasion, traffic may be directed between sensor nodes in the network. The situation is exacerbated by constraints on energy, limited memory of nodes, and potential for links to break due to nodes dying or sleeping.

In the ZigBee Pro protocol, multiple routing algorithms are employed depending on the type of traffic that is involved. For traffic between any two nodes in the network, a version of the AODV protocol is employed. This is a reactive protocol that is suitable for occasional traffic between two nodes, especially if they are close to one another in the network topology. A proactive approach is selected for data from many sensor nodes to the BS or gateway. The BS periodically broadcasts its presence to nodes that are one hop away and these nodes can inform their neighbors, and so on. When sensors send data to the BS, packets carry the source routes (i.e. the nodes through which they have reached the BS). The BS thus does not need to store routes to thousands of devices. Instead, it can simply send a response back to a sensor node using the source route embedded in the received packet. Using source routes increases the overhead in packets, but this has been used as a compromise between complexity and overhead.

One of the early routing protocols developed for sensor networks is MintRoute, which is proactive and maintains information about the next hop that can take a packet towards the BS. It recognizes the fact that, in a dense network, there will be a few good links to one-hop neighbors and many weak links to other nodes and develops appropriate metrics to account for the best route towards the BS. We note that neither of these two routing schemes considers energy efficiency or optimizations of any kind for sensor networks.

Several routing schemes have been defined with the recognition that sensor networks are data oriented – in that sensors have to sense and collect data for delivery to a sink. It is likely that such data is spatially correlated and it may be possible to reduce the energy consumption by eliminating redundancies through processing of the data en route to the BS. In the *sensor protocol for information via negotiation* (SPIN), sensor nodes perform some metadata negotiation before transmitting the data towards the sink. The nodes can also take into account the remaining energy they have to decide which node transmits the data. *Directed diffusion* aggregates data along a route towards the BS. The BS queries sensor nodes with *interests* (e.g. data from nodes that see a particular range of values). The interest is transmitted through the network by sensor nodes. Depending on the interest, a gradient can be set up for different flows from leaf nodes to the BS. The flows with the strongest gradients are reinforced to prevent flooding. The protocol has been evaluated using random sources in the sensor field and a circular *event region*, which generates the event of interest. Both SPIN and directed diffusion save energy because of elimination of redundancies. In both cases, aggregation occurs only when routes intersect (or when nodes are close together). Among other routing protocols of interest is the COUGAR data-centric protocol that assumes that the entire network is a huge distributed database and abstracts query processing from networking functions to identify nodes that need to deliver data. In-network data aggregation can provide further savings.

The routing protocols considered so far are flat routing protocols that assume that all sensors are involved in routing in the same manner. As described previously, one of the methods for establishing a sensor network is to create clusters with cluster heads. This can make routing more scalable by introducing a hierarchy into the network architecture. Several routing schemes have been specified with clustering in view. The earliest, called LEACH, also lets cluster heads aggregate data. The assumption that cluster heads can directly send data to the BS is a weakness of LEACH. In the *threshold-sensitive energy-efficient sensor network* (TEEN) protocol, cluster heads inform nodes in their cluster the hard and soft thresholds for the sensed quantity, which indicate the interest in the data. This reduces unnecessary transmissions to the cluster heads. TEEN supports changing cluster heads and thresholds as necessary.

Location-based routing protocols have received attention in the literature. These protocols assume that sensor nodes are aware of their locations (see Chapter 12) and can be addressed accordingly using their location information. Location awareness can be combined with a hierarchy. For example, in the *geographic adaptive fidelity* protocol, nodes form virtual grids based on their location. They elect one node in the grid to be awake while the others go to sleep. The node that is awake will be responsible for sensing and communicating.

Other routing protocols consider maximizing battery life, using multiple paths for routing, QoS-based routing that balances data quality and energy consumption, mobility, and multiple BSs.

13.6.3 Coverage, Connectivity, and Topology Control

As mentioned in the introduction, one of the key issues in sensor network deployment is the interaction between coverage, connectivity, and topology control. Loosely speaking, coverage implies that sensors have been deployed so that they do not miss sensing events of interest in the geographical area deployed. Assuming that the area is "covered," sensors must now be able to communicate such that there are always paths available to report the sensed data to the BS or sink (connectivity). Topology control seeks to optimize the "connectivity" of the network by changing the transmission range of nodes while reducing energy consumption. This may have the by-product of reducing contention and collisions in the network. In this section, we provide a brief overview of the concepts related to coverage, connectivity, and topology control without going into the details of algorithms or approaches investigated in the research literature. Much of this work is still in the theoretical domain, with few actual applications employing the results from this research in real sensor networks.

We start with the observation that the coverage required by different applications may be different. The amount of coverage expected to monitor habitats or the temperature in an area may be quite different from that required to detect intrusions in a given area. Even the definitions of coverage can be quite different. *Blanket coverage* assumes that it is possible to deploy nodes in a static arrangement such that there is a maximum detection rate of events in the sensor field, while *barrier coverage* minimizes the chance that an event goes undetected. *Sweep coverage* envisions moving sensor devices in the sensor field so as to balance the detection rate and missed detections in a given unit area.

It is generally believed that the quality of sensing by a sensor device degrades with distance from the sensor, in a manner similar to how the signal strength drops with distance. In fact, this degradation is modeled in a similar manner: the sensitivity of the sensor drops as the mth power of the Euclidean distance between the sensor and the target point to be sensed. Like the unit disk model for communications (see Section 13.4), a threshold determines when the sensing is essentially nonexistent. A circle with a sensing radius R_s thus defines the coverage of a given sensor. Like the gray zone, this model can be enhanced to have a zone where the sensing coverage is probabilistic, at closer distances the sensing is perfect, and at farther distances there is no sensing (see Fig. 13.9). We note that specific models for specific types of sensor (e.g. temperature or seismic) do not appear to be readily available.

While it is not trivial to quantify the coverage provided by a particular sensor deployment, in order to obtain an idea of how good or how poor the coverage is, *paths* within the sensor field with certain coverage are employed. Let us suppose that sensors are deployed (say randomly) in a field. Each sensor has the ability to sense an event, which drops with distance from the sensor. Along a given path through the field, the sensitivity to detect an event changes.

A path that results in the least probability of detection represents the worst-case scenario. Two optimization problems that determine the *minimal exposure path* and the *maximal breach path* have been defined to quantify the worst-case path. In the former case, exposure is defined as the path integral of a sensing function (which depends on the distance from the

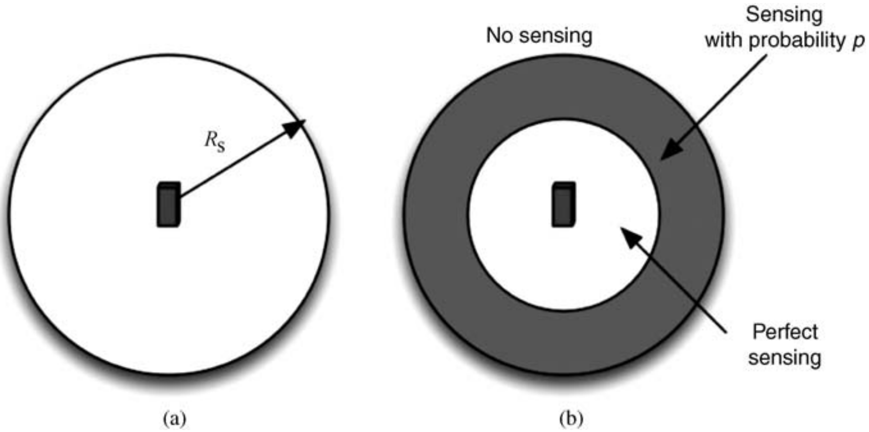


FIGURE 13.9 (*a*) Unit disk sensing and (*b*) sensing with a gray zone.

closest sensor or distances from a group of sensors that contribute a certain sensitivity to a given point) during a specific time interval. The path that results in minimal exposure represents one metric of the worst-case coverage in the sensor network. The maximal breach path uses a simpler idea. It corresponds to the path in the sensor field where each point on the path is at the maximum possible distance from the closest sensor. The maximal breach path can be constructed as follows. First draw line segments joining pairs of sensors in the deployed sensor field. Then draw lines that bisect these line segments and use the intersections of the bisectors as vertices and the bisectors as edges of polygons that tessellate the sensor field. This creates a Voronoi diagram where all points inside the polygons are closer to some sensor compared with points along the edges. The maximal breach path must lie along the edges of the polygons, but the amount of breach depends on how far a given edge is from the sensors within the polygon.

In a similar manner, it is possible to define paths that have the best coverage, and they are referred to in the literature as the *maximal exposure paths* and *maximal support paths.* The former corresponds to a path that has the largest exposure path integral in a manner similar to the minimal exposure path. The maximal support path corresponds to a path that has points closest to sensors on it.

Research literature evaluating the possibility of *moving* sensors to obtain the best coverage after initial deployment also exists. If mobile robots are employed, a gradient to repel and attract sensors has been proposed to distribute sensors in a field. The Voronoi diagram can also be used to determine coverage holes and fill them by repelling sensors that are close to one another or to the edge of a sensor field.

One question that arises is whether coverage automatically implies connectivity in the network. Connectivity is often measured by whether or not the graph of communication links that exist between pairs of nodes is connected or partitioned. A result from the research literature indicates that, given a communication radius that is two times the sensing radius (assuming unit disk models for both sensing coverage and communications), the network is connected if it also provides complete coverage over the given area. Coverage here assumes that at least one sensor can sense any point in the given region. Similar results exist for situations where n sensors are required to cover any point in the given region.

Topology control ensures connectivity of the network by changing the transmission range of nodes (by increasing the transmit power). While this is not practical in the real sensor networks of today, one can anticipate benefits of this approach in the future. Two types of topology control technique are considered: homogeneous and nonhomogeneous. In the homogeneous case, all sensors must have the same transmit and receive ranges, and the question is what range is the smallest to provide connectivity in the network. In the nonhomogeneous case, sensors are allowed to have different transmission ranges (which may result in some asymmetric links).

Related to the idea of topology control is the idea of developing optimal sleep schedules. In many applications, many more sensor nodes are deployed than necessary to cover the region and keep the network connected. In such cases, it is not necessary to keep all the sensors active all the time. Developing optimal sleep schedules, where a subset of nodes sleeps while others are active, such that the network is still covered and connected, is another problem that has been considered in the research literature.

13.6.4 Synchronization

Synchronization is an important function in sensor networks. As already described in Section 13.5, many MAC protocols rely on synchronization between nodes. Synchronization is also necessary for localization schemes that rely on time of flight of a signal. When nodes perform data aggregation, it is necessary for them to be synchronized to recognize the age of data. This is also necessary for some applications at the BS, which needs temporal information related to sensed data to make decisions. Synchronization can be used to save energy in the network by having correct sleep schedules and transmissions.

It is not easy to synchronize all sensor nodes to a common clock for several reasons. The devices are supposed to be inexpensive and may not have accurate clocks. The clock drifts between different sensors can be high. There are several sources of error in computing a common time, such as the time at which the synchronization message is composed may be different from the time at which it was sent due to medium access delays, the propagation delay is unpredictable, and the time required for a receiving node to process the synchronization message may be hard to determine. The challenges only increase in multi-hop networks.

In sensor networks, one can think of local synchronization and global time synchronization. In the case of local synchronization, the argument is that nodes should run without any synchronization for most of the time. When sensed data need to be transmitted, nodes can temporarily synchronize in a relative manner, perhaps to the first node that initiates synchronization. This makes sense in low-duty-cycle networks where communications is infrequent and the importance of aggregation and ordering is mostly in local groups of sensors. Implementations of local synchronization through the reference broadcast synchronization (RBS) protocol have shown synchronization on the order of a few microseconds on mote-class sensor devices. Global synchronization schemes have also been proposed for sensor networks. For example, the timing-sync protocol for sensor networks (TPSN) uses a root node and levels (similar to Fig. 13.5) for achieving network-wide synchronization. The way it operates is as follows. The root node sends a level discovery message. Its immediate neighbors form level 1. Their immediate neighbors that cannot reach the root form level 2, and so on. Then a two-way message exchange occurs between level 1 nodes and the root node with time stamps of local times of transmission and reception. These allow the clock drift and propagation delay to be computed by level 1 nodes, which can then synchronize themselves to the root node. Level 2 nodes then synchronize themselves with level 1 nodes, and so on.

In IEEE 802.15.4, the beacon from the WPAN coordinator can be used for synchronization purposes.

13.6.5 Security

Security in sensor networks is an important topic, although there are opinions that suggest that work in this area may be of academic interest only. In this section, we briefly discuss the security features provided in the IEEE 802.15.4 standard and briefly mention some of the research work describing security threats in sensor networks.

Security Threats. Sensor networks are especially vulnerable to security attacks for the following reasons: (a) the transmission medium is air, and so it is easy to obtain remote access to the medium using powerful antennas for eavesdropping, jamming, or injecting malicious traffic; (b) sensors may be deployed in huge numbers and are supposed to be of low cost, with the implication that the possibility of some of the sensors being captured, tampered with, and compromised being very real. It is possible for adversaries to deploy their own sensors into a sensor field, but this requires physical protection of the sensed region. There are numerous security threats that have been considered in the research literature of wireless sensor networks, making it difficult to consider them all together. Instead, it is easier to group the threats into categories. While there are overlaps between them, we can classify these threats into the following categories.

Physical Layer Threats: At the physical layer of the communications protocol stack, common threats against sensor networks are disruptions to communications through jamming and node disabling. By jamming, an attacker may disrupt reliable communications by transmitting signals that interfere with the radio signals of sensor nodes. This may result in partitioning of the network, lower reliability of the sensed data because of the lack of availability of data from certain sensed regions, and ultimately result in the battery exhaustion of nodes repeatedly transmitting data till they are acknowledged or receiving bogus data. Jammers can be classified as those that may be outsiders employing a constant radio signal, deceptive jammers that inject regular packets into the network, random jammers that alternate between sleep and awake states, and reactive jammers that cleverly disrupt communications upon sensing channel activity. Experimental studies indicate that packet delivery ratios are adversely impacted by all of these types of jammer. Jamming may adversely impact sensor nodes at the edges of a network or those that are towards vulnerable physical areas.

Eavesdropping Threats: One of the most common threats in wireless sensor networks is information leakage, where an adversary may obtain the sensed information by simply passively eavesdropping on the radio signals being transmitted by sensor nodes. Eavesdropping is especially problematic even with encryption, because of the potential for sensor nodes possessing keys to be compromised or captured by adversaries. One model for computing the eavesdropping vulnerability is based on the adversary interested in predicting the behavior or aggregate output of the sensor network. In addition to information leakage, radio transmissions may reveal the location of sensors and the sink node and allow other kinds of analyses on the traffic patterns. An adversary may also be able to poll sensor nodes actively for information if there is no authentication of queries in the network.

Threats Impacting Routing: In networks, it is important for nodes to know where to send data packets so that they reach the destination in an efficient way. Such routing protocols in sensor networks are still evolving, since attempts to directly use routing protocols designed for mobile ad hoc networks in sensor networks have faced challenges due to the scalability and energy requirements of sensor networks. Geographical and geometric routing that makes use of the knowledge of the Euclidean coordinates of sensors is proposed as an efficient means of routing data to the destination. However, it is likely that, in general, routing in sensor networks faces the same threats as those in mobile ad hoc networks. Such threats include location disclosure, replay of old routing information, disruption by fabricating routing information, and route table poisoning. In addition, wormhole, blackhole, and Sybil attacks are possible. In blackhole attacks, malicious nodes advertise themselves as closer to the destination, thereby making themselves part of most routes. They can then disrupt network operation by dropping packets or get information by eavesdropping. In Sybil attacks, a single malicious node claims to be more than one node. This way, it could claim a disproportionate amount of resources and also perform blackhole attacks. If there are collaborating nodes, then they may create a wormhole (a tunnel) between them and create the impression of a false network topology.

Threats Impacting Position Information: The position of a sensor node has importance in several applications. For example, temperature variations over a given area may have to be accurately characterized, in which case the position location of the sensor reporting the temperature reading needs to be known to a certain accuracy. Such position information may be used for routing or even in security measures. Further, the location of a sensor monitoring a critical quantity may itself need to be kept secure (location privacy). Malicious nodes can interfere with the reporting of position location information in many ways. They can fabricate the position information or interfere with the support infrastructure used by sensors to determine their own positions. In the latter case, there are many different approaches for determining the position of sensor nodes, such as using beacons from nodes at known positions, determining the number of hops a node is away from a reference node, and so on. Malicious nodes can interfere with such position-determining activity.

Threats Impacting Data Aggregation and in-Network Processing: Data aggregation and in-network processing is an important feature of data-intensive sensor networks. Because sensor nodes collect a huge amount of data and sometimes only aggregate information is necessary at the sink (e.g. average value or the sum of the sensed quantity), intermediate nodes can process the received data (in-network processing) or fuse data and forward those values. This reduces the communication costs and delays in the network. However, such functionality makes it extremely easy for malicious nodes to introduce false values that corrupt the processed or fused values. If a malicious node is responsible for fusing or aggregating data, then the problem could be worse. If a Sybil attack is launched, then a node can claim multiple identities and further skew the aggregated data by creating multiple false reports.

Threats Against Time Synchronization: Sensor networks often require nodes in the network to be time synchronized for many reasons, such as data fusion, scheduled transmissions for saving power, tracking duplicate sensed data, and so on. Time synchronization can be achieved using reference broadcasts or sender–receiver synchronization. It is possible to disrupt the sensor network operation by misleading different nodes about the time at which they have to perform operations like sensing or transmissions of packets.

Miscellaneous Threats: If sensor nodes are compromised, then they can disrupt a sensor network in many ways. For instance, a compromised node may not follow the medium access protocol and hog the medium. If a node assumes several identities, as in the case of a Sybil attack, it could access the medium more often than it should normally have fair access. These may both deny access to radio resources by legitimate sensor nodes. In many types of sensor network, a sink node is used to collect data after a query to many sensor nodes in the networks and for other types of network maintenance. In some cases, mobile sink nodes are employed to poll sensors or collect data from a set of static sinks. Compromise of sink nodes can lead to damages or disruption of a sensor network.

Security in IEEE 802.15.4. In IEEE 802.15.4, no attempt is made to address the many different vulnerabilities or threats that exist for the sensor network applications described above. However, cryptographic protection of communications over links is part of the standard. The cryptographic operations that are part of the standard provide for confidentiality, integrity, and authentication of the communicated data using encryption and message authentication codes. The standard assumes that secrets that are to be shared between sensor nodes is an issue that is beyond its scope. So it is necessary to employ additional key establishment and key management schemes with IEEE 802.15.4 sensor devices. Keys may be pairwise or shared by a group of nodes. The rest of this section assumes that, somehow, pairs of communicating nodes share keys with each other (group or pairwise).

Protection in IEEE 802.15.4 can be adopted on a per frame basis with message authentication (includes integrity and replay protection) and optional encryption of contents for confidentiality. This enables deploying security as needed without expending energy for cryptographic operations that are not necessary. The size of the message authentication code can be varied (32, 64, and 128 bits), offering various levels of protection. Similarly, encryption may or may not be enabled. It is also possible to send frames without any protection.

The most general form of protection in IEEE 802.15.4 involves the *c*ounter mode with *c*ipher-block-chaining *m*essage authentication code (CCM) operation of a block cipher. This operation is used in the IEEE 802.11i standard for wireless local area networks as well. The block cipher specified in the IEEE 802.15.4 standard is the AES, an encryption scheme that was standardized by NIST in 2001. This scheme works as follows. A counter is incremented and encrypted with an encryption key. The resulting output stream is XOR-ed with the data to provide confidentiality. The data is broken into bocks of 128 bits. Each block is XOR-ed with the previous block's ciphertext and then encrypted using an authentication key. The first block is XOR-ed with an initial vector. The final encrypted block is truncated to the appropriate number of bits to form the message authentication code. A receiving node can locally perform the same operations to decrypt the data or to compare the received message authentication code with the locally computed message authentication code to see if the message has been modified or fabricated.

QUESTIONS

1. What is habitat monitoring? Why are sensor networks suited for this application?
2. Name two applications other than habitat monitoring that sensor networks are used for.
3. What are the four classes of sensor devices?
4. What are the differences between a base station class sensor device and a mote class sensor device?
5. What challenges in size reduction and mobility are expected in the future for sensor devices?
6. Why is it beneficial to employ unlicensed spectrum for sensor networks? What are the disadvantages?
7. What is the unit disk model for communications in sensor networks? Why is this model not correct?
8. Explain the concept of the grey zone for links between sensor devices.
9. What is the difference between a channel page and a channel number in IEEE 802.15.4?
10. What issues are more important for sensor networks at the MAC layer than traditional wired networks?
11. Why are sensors closer to the base station more likely to have depleted batteries than those farther away from the base station?
12. What is the difference between a full function device and a reduced function device in IEEE 802.15.4?
13. What is the primary objective of low duty cycle MACs?
14. What techniques are suggested to reduce energy waste in accessing the medium sor sensor networks?
15. What are event-driven medium access control protocols?
16. Differentiate between the star topology and the cluster-tree topology in IEEE 802.15.4 based sensor networks.
17. Name three types of routing protocols for sensor networks. What objectives do they try to meet?
18. Differentiate between blanket coverage and barrier coverage in sensor networks.
19. What is a minimal exposure path?
20. Why is synchronization important in sensor networks?
21. What is a blackhole attack?
22. What security services are provided by IEEE 802.15.4? Which of them are optional?
23. Describe the most general form of data protection in IEEE 802.15.4.

REFERENCES

[3GPP 25.305] 3GPP TS 25.305 V. 7.2.0 (2006-03), Technical Specification: Group Radio Access Network; Stage 2 Functional Specification of User Equipment Positioning in UTRAN (Release 7), 3rd Generation Partnership Project, 2006.

[3GPP 25.331] 3GPP TS 25.331 V7.1.0 (2006-06), Technical Specification: Group Radio Access Network; Radio Resource Control (RRC); Protocol Specification (Release 7), 3rd Generation Partnership Project, 2006.

[Abr70] N. Abramson, The ALOHA system: another alternative for computer communications, *Proc. AFIPS*, Houston, 1970, Vol. 37, pp. 281–285.

[Ah003] S. Ahonen and H. Laitinen, Database correlation method for UMTS location, *Proc. IEEE Vehicular Technology Conference (VTC)*, Vol. 4, April 2003, pp. 2696–2700.

[Aka08] Akamai Report: State of the Internet, Vol. 1, No 1, 1st Quarter, 2008.

[Aky02] I. F. Akyildiz *et al.*, A survey on sensor networks, *IEEE Commun. Mag.*, **40**(8), 102–114 (2002).

[Ala98] S. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE J. Sel. Areas Commun.*, **16**, 1451–1458 (1998).

[Alk04] J. Al-Karaki and A. E. Kamal, Routing techniques in wireless sensor networks: a survey, *IEEE Wireless Commun.*, **11**(6), 6–28 (2004).

[Amo99] E. G. Amoroso, *Intrusion Detection: An Introduction to Internet Surveillance, Correlation, Trace Back, Traps, and Response*, Intrusion.net Books, 1999.

[Ana05] G. Anastasi *et al.*, Understanding the real behavior of mote and 802.11 ad hoc networks: an experimental approach, *Pervasive Mobile Comput.*, **1**(2), 237–256 (2005).

[APWG] Anti-phishing Working Group: http://www.antiphishing.org

[ARS08] http://arstechnica.com/journals/apple.ars/2008/06/07/apple-ranked-third-among-smartphone-vendors

[Atmel] Atmel Corporation: http://www.atmel.com

[Ayl07] R. Aylward and J. A. Paradiso, A compact, high-speed, wearable sensor network for biomotion capture and interactive media, *Proc. IPSN*, April 2007.

[Bah00] P. Bahl and V. N. Padmanabhan. RADAR: an in-building RF-based user location and tracking system, *Proc. IEEE INFOCOM'00*, March 2000, pp. 775–784.

[Bar03] S. Barnes, Location-based services: the state of the art, *e-Service J.*, **2.3**, 59–70 (2003).

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

[Bej04] R. Bejtlich, *The Tao of Network Security Monitoring*, Addison-Wesley, 2004.

[Ben00] P. Bender *et al.*, CDMA/HDR: A bandwidth efficient high-speed wireless data service for nomadic users, *IEEE Commun. Mag.*, **38** (7), 70–77 (2000).

[Ber00] C. Bererton, L. Navarro-Serment, R. Grabowski, C. Paredis and P. Khosla, Millibots: small distributed robots for surveillance and mapping, *Proc. Government Microcircuit Applications Conference*, 2000.

[Ber87] D. Bertsekas and R. Gallagher, *Data Networks*, Prentice Hall, 1987.

[Ber94] H. L. Bertoni *et al.*, UHF propagation prediction for wireless personal communications, *Proc. IEEE*, **82** (9), 1333–1359 (1994).

[Blu00] Bluetooth Special Interest Group, Specifications of the Bluetooth System, vol. 1 v. 1.1, 'Core' and vol. 2 v. 1.0 B 'Profiles', 2000.

[Bren05] M. R. Brenner, M. L. F. Grech, M. Torabi, and M. R. Unmehopa, The Open Mobile Alliance and trends in supporting the mobile services industry, *Bell Lab. Tech. J.*, **10** (1), 59–75 (2005).

[Bud97] K. C. Budka, H. J. Jiang, and S. E. Sommars, Cellular packet data networks, *Bell Lab. Tech. J.*, **2** (3), 164–181 (1997).

[Bul00] N. Bulusu, J. Heidemann, and D. Estrin, GPS-less low-cost outdoor localization for very small devices, *IEEE Personal Commun.*, **7** (5), 28–34 (2000).

[Caf98] J. Caffery, Jr. and G. L. Stuber, Subscriber location in CDMA cellular networks, *IEEE Trans. Veh. Technol.*, **47** (2), 406–416 (1998).

[Cai97] J. Cai and D. J. Goodman, General packet radio service in GSM, *IEEE Commun. Mag.*, **35** (10), 122–131 (1997).

[Cal02] E. Callaway *et al.*, Home networking with IEEE 802.15.4: a developing standard for low-rate wireless personal area networks, *IEEE Commun. Mag.*, **40** (8), 70–77 (2002).

[Cert] The United States Computer Emergency Readiness Team website at http://www.us-cert.gov.

[Cha99] S. Chakrabarti and A. Mishra, A network architecture for global wireless position location services, *Proc. ICC'99*, 1999, pp. 1779–1783.

[Cha00] M. V. S. Chandrashekhar, P. Choi, K. Maver, R. Sieber, and K. Pahlavan, Evaluation of interference between IEEE 802.11b and Bluetooth in a typical office environment, *Proc. PIMRC'01*, San Diego, 2001.

[Che03] Z. Chen, L. Gao, and K. Kwiat, Modeling the spread of active worms, *Proc. IEEE Infocom*, April 2003

[Che97] G. Cherubini *et al.*, 100BASE-T2: a new standard for 100 Mb/s Ethernet transmission over voice-grade cables, *IEEE Commun. Mag.*, **35** (11), 115–122 (1997).

[Ches03] W. R. Cheswick, S. M. Bellovin, and A. D. Rubin, *Firewalls and Internet Security*, Addison-Wesley, 2003.

[Com98] IEEE Communications Magazine Special Issue on Geolocation Applications, April 1998.

[Cont] The Contiki Operating System: http://www.sics.se/contiki/.

[Coo06] B. W. Cook, S. Lanzisera, and K. S. J. Pister, SoC issues for RF smart dust, *Proc. IEEE*, **94** (6), 1177–1196 (2006).

[Cro97a] B. P. Crow, I. Widjaja, L.G. Kim and P.T. Sakai, IEEE 802.11 Wireless Local Area Networks, *IEEE Commun. Mag.*, **35** (9), 116–126 (1997).

[Cro97b] B. P. Crow, I. Widjaja, L. G. Kim, and P. T. Sakai, Investigation of the IEEE 802.11 medium access control (MAC) sublayer functions, *Proc. IEEE Infocom*, Vol. 1, 1997, pp. 126–133.

[Cvi08] M. Cvijetic and P. Magill(eds), Delivering on 100 Gbe Promise, Special Issue on Applications and Practice, IEEE Commun. Mag., June 2008.

[Dem06] I. Demirkol, C. Ersoy, and F. Alagoz, MAC Protocols for wireless sensor networks: a survey, *IEEE Commun. Mag.*, **44** (4), 115–121 (2006).

[Den96] L. R. Dennison, BodyLAN: a wearable personal network, *Second IEEE Workshop on WLANs*, Worcester, MA, 1996.

[Dha02] N. Al-Dhahir, C. Fragouli, A. Stamoulis, W. Younis, and R. Calderbank, Space–time processing for broadband wireless access, *IEEE Commun. Mag.*, **40** (9), 136–142 (2002).

[Dju401] G. M. Djuknic and R. E. Richton, Geolocation and assisted GPS, *IEEE Comput.*, **34** (2), 123–125 (2001).

[Dra98] C. Drane, M. Macnaughtan, and C. Scott, Positioning GSM telephones, *IEEE Commun. Mag.*, **36** (4), 46–54, 59 (1998).

[Dru01] M.-A. Dru and S. Saada,Location-based mobile services: the essentials, Alcatel Telecommunications Review, 2001.

[Dust] Dust Networks website: http://www.dustnetworks.com

[Edn04] J. Edney and W. A. Arbaugh, *Real 802.11 Security: Wi-Fi Protected Access and 802.11i*, Pearson Education, 2004.

[Eka] http://www.ekahau.com.

[Ela04a] M. Elaoud, D. Famolari, and A. Ghosh, Experimental VoIP capacity measurements for 802.11b WLANs, *IEEE Consumer Communications and Networking Conference*, 2004.

[Ela04b] M. Elaoud and P. Agrawal, VoIP capacity in IEEE 802.11 networks, *Proc. IEEE PIMRC*, 2004.

[ENN98] G. Ennis,Doc: IEEE P802.11-98/319, Impact of Bluetooth on 802.11 direct sequence, September 15, 1998.

[Erc03] V. Erceg *et al.*, Indoor MIMO WLAN channel models, IEEE 802.11-03/161r0, June 11, 2003.

[Ert98] R. B. Ertel, P. Cardieri, K. W. Sowerby, T. S. Rappaport, and J. H. Reed, Overview of spatial channel models for antenna array communication systems, *IEEE Personal Commun.*, **5** (1), 10–22 (1998).

[Fan03] A. Fanimokun and J. Frolik, Effects of natural propagation environments on wireless sensor network coverage area, *Proc. 35th IEEE Southeastern Symposium on System Theory*, March 2003.

[FCC E-911] FCC's E-911 webpage: http://www.fcc.gov/e911.

[Fei00] J. Feigin, K. Pahlavan, and M. Ylianttila, Hardware-fitted modeling and simulation of VoIP over a wireless LAN, *52nd IEEE VTS Fall VTC, Vehicular Technology Conference*, Vol. 3, pp. 1431–1438, 2000.

[Fei99] J. Feigin and K. Pahlavan, Measurement of characteristics of voice over IP in a wireless LAN environment, *IEEE International Workshop on Mobile Multimedia Communications (MoMuC '99)*, 1999, pp. 236–240.

[Fer80] P. Ferert, Application of spread spectrum radio to wireless terminal applications, *Proc. NTC'80*, Houston, TX, December 1980, pp. 244–248.

[Fig69] W. Figel, N. Shepherd, and W. Trammell, Vehicle Location by a Signal Attenuation Method, *IEEE Trans. Veh. Technol.*, **VT-18**, 105–110 (1969).

[Fis80] M. J. Fischer, Delay analysis of TASI with random fluctuations in the number of voice calls, *IEEE Trans. Commun.*, **COM-28** (11), 1883–1889 (1980).

[For95] S. Fortune *et al.*, WISE Design of indoor wireless systems: practical computations and optimization, *IEEE Comput. Sci. Eng.*, **2** (1), 58–68 (1995).

[Fos98b] G. J. Foschini and M. Gans, On limits of wireless communications in a fading environment using multiple antennas, *IEEE Wireless Commun.*, **6**, 311–315 (1998).

[Frii46] H. T. Friis, A note on a simple transmission formula, *Proceedings of the IRE and Wave and Electrons*, May 1946, pp. 254–256.

[Gan91] R. Ganesh and K. Pahlavan, Modeling of the indoor radio channel, *IEE Proc. I: Commun. Speech Vision*, **138**, 153–161 (1991).

[Gar00] V. K. Garg, *IS-95 and CDMA2000*, Prentice Hall, Upper Saddle River, NJ, 2000.

[Gar03] S. Garg and M. Kappes, Can I add a VoIP call? *Proc. ICC'03*, 2003, pp. 779–783.

[Gar99] V. K. Garg and J. E. Wilkes, *Principles and Applications of GSM*, Prentice Hall, Upper Saddle River, NJ, 1999.

[Gas02] M. S. Gast, *802.11 Wireless Networks: The Definitive Guide*, O'Reilly & Associates, 2002.

[Gay03] D. Gay *et al.*, The nesC language: a holistic approach to networked embedded systems, *Proc. ACM SIGPLAN Conference on Programming Language Design and Implementation*, San Diego, 2003.

[Gfe80] F.R. Gfeller,Infranet: infrared microbroadcasting network for in house data communication, IBM research report, RZ 1068 (#38619), April 27, 1981.

[Gho08] A. Ghosh and S. K. Das, Coverage and Connectivity issues in wireless sensor networks: a survey, *Pervasive Mobile Comput.*, **4**, 303–334 (2008).

[Goo89] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, Packet reservation multiple access for local wireless communications, *IEEE Trans. Commun.*, **37** (8), 885–890 (1989).

[Goo91] D. J. Goodman and S. X. Wei, Efficiency of packet reservation multiple access, *IEEE Trans. Veh. Technol.*, **40** (1, Pt 2), 170–176 (1991).

[Goo97] D. J. Goodman, *Wireless Personal Communications Systems*, Addison-Wesley, 1997.

[Haa00] J. C. Haartsen and S. Mattisson, Bluetooth – a new low-power radio interface providing short-range connectivity, *Proc. IEEE*, **88** (10), 1651–1661 (2000).

[Had06] S. Hadim and N. Mohamed, Middleware challenges and approaches for wireless sensor networks, *IEEE Distributed Systems Online*, March 2006.

[Hal96] C. J. Hall, and W. A. Foose, Practical planning for CDMA networks: a design process overview, *Proc. Southcon'96*, 1996, pp. 66–71.

[Hal99] K. Halford, S. Halford, M. Webster, and C. Ander, Complementary code keying for RAKE-based indoor wireless communication, *IEEE International Symposium on Circuits and Systems*, Vol. 4, Orlando, FL, 1999, pp. 427–430.

[Ham02] M. Hamalainen *et al.*, On the UWB System Coexistence with GSM900, UMTS/WCDMA, and GPS, *IEEE J. Sel. Areas Commun.*, **20** (9), 1712–1721 (2002).

[Ham86] J. L. Hammond and P. J. P. O'Reilly, *Performance Analysis of Local Computer Networks*, Addison-Wesley, Reading, MA, 1986.

[Han05] C.-C. Han *et al.*, A dynamic operating system for sensor nodes, *MobiSys '05: The Third International Conference on Mobile Systems, Applications, and Services*, 2005.

[Har06] S. Harsha, A. Kumar, and V. Sharma, An analytical model for the capacity estimation of combined VoIP and TCP file transfers over EDCA in an IEEE 802.11e WLAN, *IWQoS 2006, 14th IEEE International Workshop on Quality of Service*, 2006.

[HART] HART Communication Foundation website: http://www.hartcomm2.org/.

[Hat80] M. Hata, Empirical formula for propagation loss in land mobile radio services, *IEEE Trans. Veh. Technol.*, **VT-29** (3), 317–324 (1980).

[Hau94] T. Haug, Overview of GSM: philosophy and results, *Int. J. Wireless Inform. Networks*, **1**(1), 7–16 (1994).

[Hay] V. Hayes, Standardization efforts for wireless LANS, *IEEE Network*, **5** (6), 19–20 (1991).

[Hei98] R. Heille,WPAN functional requirement, Doc. IEEE 802.11/98/58, January 22, 1998.

[Hil01] A. Hills, Large scale wireless LAN design, *IEEE Commun. Mag.*, **39** (11), 98–105 (2001).

[Hil04] J. Hill, M. Horton, R. Kling, and L. Krishnamurthy, The platforms enabling wireless sensor networks, *Commun. ACM*, **47** (6), 41–46 (2004).

[How90] S. J. Howard and K. Pahlavan, Measurement and analysis of the indoor radio channel in the frequency domain, *IEEE Trans. Instr. Meas.*, **39** (5), 751–755 (1990).

[Hu07] W. Hu *et al.*, The design and evaluation of a hybrid sensor network for cane-toad monitoring, ACM Transactions on Sensor Networks, December 2007.

[IEE00] IEEE 802.15 Working Group: http://grouper.ieee.org/groups/802/15/.

[IEE01] *Proc. IEEE Workshop on Wireless LANs*, Newton, MA, September 2001.

[IEEE06] IEEE Std 802.15.4-2006, *Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)*, IEEE LAN/MAN Standards Committee, 2006.

[IEEE07] Delivering on 100GbE Promise, Special Issue, IEEE Applications & Practice, November 2007.

[IS-801-99] IS-801, Position Determination Service Standard for Dual Mode Spread Spectrum Systems, Telecommunications Industry Association, 1999.

[Jennic] Jennic's website: http://www.jennic.com.

[Ju02] P. Juang *et al.*, Energy efficient computing for wildlife tracking: design tradeoffs and early experiences with ZebraNet,*Proc. ACM ASPLOS-X*, Oct. 2002

[Kae04] K. Kaemarungsi and P. Krishnamurthy, Modeling of indoor positioning systems based on location fingerprinting,*Proc. IEEE Infocom*, March 2004.

[Kap02] S. Kapp, 802.11a. More bandwidth without the wires, *IEEE Internet Comput.*, **6** (4), 75–79 (2002).

[Kap96] E. D. Kaplan, *Understanding GPS: Principles and Applications*, Artech House Publishers, 1996.

[Kau02] C. Kaufmann, R. Perlman, and M. Speciner, *Network Security: Private Communication in a Public World*, Prentice Hall PTR, 2002.

[Kav87] M. Kavehrad and P. J. McLane, Spread spectrum for indoor digital radio, *IEEE Commun. Mag.*, **25** (6), 32–40 (1987).

[Kei89] G. E. Keiser, *Local Area Networks*, McGraw-Hill, New York, 1989.

[Ker00] J.P. Kermoal, L. Schumacher, P.E. Mogensen and K.I. Pedersen, Experimental investigation of correlation properties of MIMO radio channels for indoor picocell scenarios,*52nd IEEE Vehicular Technology Conference*, 2000, pp. 14–21.

[Kim06] S. Kim *et al.*, Poster Abstract:*Wireless Sensor Networks for Structural Health Monitoring, 5th ACM Conference on Embedded Network Systems (SenSys)*, Nov. 2006.

[Kle75] L. Kleinrock, *Queing Systems: Volume 1: Theory*, John Wiley & Sons, Inc., New York, 1975.

[Kle75b] L. Kleinrock and S. S. Lam, Packet switching in multi-access broadcast channel: performance evaluation, *IEE Trans. Commun.*, **23**, 410–423 (1975).

[Koh04] R. Kohno, M. Welborn, and M. McLaughlin,DS-UWB Proposal, IEEE 802.15-04/140r2, March 2004.

[Kot00] [Kos00] H. Koshima and J. Hoshen, Personal locator services emerge, *IEEE Spectrum*, **37** (2), 41–47 (2000).

[Kot04] S. Kota., K. Pahlavan, and P. Leppanen, *Broadband Satellite Internet*, Kluwer Publishing Company, 2004.

[Kri98] P. Krishnamurthy, K. Pahlavan, and J. Beneat, Radio propagation modeling for indoor geolocation applications,*Proc. PIMRC'98*,September 1998.

[Kum74] K. Kummerle, Multiplexer performance for integrated line- and packet-switched traffic, *ICCC Proc.*, Stockholm, 1974, pp. 508–515.

[Kur01] J.F. Kurose and K.W. Ross, *Computer Networking – A Top-Down Approach Featuring the Internet*, Addison Wesley, 2001.

[Lee06] M. J. Lee and J. Zheng, Emerging standards for wireless mesh technology, *IEEE Wireless Commun.*, **13** (2), 56–63 (2006).

[Leo03] M. Leopold, M. B. Dydebnsborg, and P. Bonnet, Bluetooth and Sensor networks: a reality check,*Proc. ACM Sensys*, November 2003.

[Li08] P. Li, M. Salour, and X. Su, A survey of internet worm detection and containment, *IEEE Commun. Surveys Tutorials*, **10** (1), 20–35 (2008).

[LIF TS 1001-02] LIF TS 101 Specification, Location Inter-operability Forum (LIF) Mobile Location Protocol, Version 3.0.0, June 2002.

[Mac79] V. H. MacDonald, The cellular concept, *Bell Syst. Tech. J.*, **58** (1), 15–41 (1979).

[Mant] The MANTIS Group: http://mantis.cs.colorado.edu/index.php/tiki-index.php.

[Mar02] I. Martin-Escalona, F. Barcelo, and J. Paradells, Delivery of non-standardized assistance data in E-OTD/GNSS hybrid location systems,*Proc. IEEE PIMRC*, Vol. 5, September 2002, pp. 2347–2351.

[Mar85] M. J. Marcus, Recent US regulatory decisions on civil use of spread spectrum,*Proc. IEEE Globecom*, New Orleans, December 1985, pp. 16.6.1–16.6.3.

[MBO04] MBOA Physical Layer Specification, May 2004.

[McN04] C. McNab, *Network Security Assessment: Know Your Network*, O'Reilly Books, 2004.

[McN04] K. Medepalli *et al.*, Voice capacity of IEEE 802.11b, 802.11a, and 802.11g wireless LANs, *Proc. Globecom*, 2004.

[Met76] R. Metcalfe and D. Boggs, ETHERNET: distributed packet switching for local computer networks, *Commun. ACM*, **19**, 395–404 (1976).

[Mey96] M. J. Meyer, T. Jacobson, M. E. Palamara, E. A. Kidwell, R. E. Richton, and G. Vannucci, Wireless enhanced 9-1-1 service – making it a reality, *Bell Lab. Tech. J.*, **1**(2), 108–202 (1996).

[Mir04] J. Mirkovic and P. Reiher, A taxonomy of DDoS attack and DDoS defense mechanisms, *ACM SIGCOMM Comput. Commun. Rev.*, **34** (2), 39–53 (2004).

[MOB08] http://mobchina.blogspot.com/2008/01/4-million-iphone-sold.html.

[Moo02] D. Moore, C. Shannon, and K. Claffy, Code Red: a case study on the spread and victims of an Internet worm,*Proc. 2nd ACM SIGCOMM Workshop on Internet Measurement (IMW)*, 2002, pp. 273–284.

[Nag98] A. Naguib *et al.*, A space–time coding modem for high data rate wireless communications, *IEEE J. Sel. Areas Commun.*, **16** (8), 1459–1477 (1998).

[Nic03] D. Niculescu and B. Nath, Ad hoc positioning system (APS) using AOA,*Proc. IEEE Infocom*, April 2003, pp. 1734–1743.

[Nis00] D. N. Nissani and I. Shperling, Cellular CDMA (IS-95) location, A-FLT proof-of-concept interim results,*The 21st IEEE Convention of Electrical and Electronic Engineers in Israel, 2000*, 2000, pp. 179–182.

[Nor01] S. Northcutt and J. Novak, *Network Intrusion Detection: An Analyst's Handbook*, New Riders, Indianapolis, 2001.

[Nor05] S. Northcutt *et al.*, *Inside Network Perimeter Security*, New Riders, Indianapolis, 2005.

[Oku68] Y. Okumura *et al.*, Field strength and its variability in VHF and UHF land mobile service, *Rev. Electr. Commun. Lab.*, **16**, 825–873 (1968).

[OnStar] General Motors Onstar website: http://www.onstar.com.

[Pah02] K. Pahlavan and P. Krishnamurthy, *Principles of Wireless Networks: A Unified Approach*, Pearson Ed., 2002.

[Pah05] K. Pahlavan and A. Levesque, *Wireless Information Networks*, 2nd edn, John Wiley and Sons, Ltd, 2005.

[Pah85] K. Pahlavan, Wireless communications for office information networks, *IEEE Commun. Mag.*, **23** (6), 19–27 (1985).

[Pah88] K. Pahlavan and J. L. Holsinger, Voice-band data communication modems – a historical review: 1919–1988, *IEEE Commun. Mag.*, **26**, 16–27 (1988).

[Pah95] K. Pahlavan and A. Levesque, *Wireless Information Networks*, John Wiley and Sons, Inc., New York, 1995.

[Pah97] K. Pahlavan, A. Zahedi, and P. Krishnamurthy, Wideband local access: wireless LAN and wireless ATM, *IEEE Commun. Mag.*, **35** (11), 34–40 (1997).

[Pah98] K. Pahlavan, P. Krishnamurthy, and J. Beneat, Wideband radio propagation modeling for indoor geolocation applications, *IEEE Commun. Mag.*, **36** (4), 60–65 (1998).

[PalTr] PalTrack Tracking Systems: http://sovtechcorp.com.

[Pat03] W. Pattara-atikom, P. Krishnamurthy, and S. Banerjee, Distributed mechanisms for quality of service in wireless LANs, Special issue on QoS in Next-generation Wireless Multimedia Communications Systems, *IEEE Wireless Commun.*, **10** (3), 26–34 (2003).

[Pat96] S. Patel, H. W. Johnson, and J. R. Rivers,Method and apparatus for implementing a type 8B6T encoder and decoder, United States Patent 5525983, Issued on June 11, 1996.

[Ped00] K. I. Pedersen, J. B. Andersen, J. P. Kermoal, and P. Morgensen, A stochastic multiple-input–multiple-output radio channel model for evaluation of space–time coding algorithms,*52nd IEEE Vehicular Technology Conference*, 2000, pp. 893–897.

[Pet06] M. Petrova *et al.*, Performance study of IEEE 802.15.4 using measurements and simulations, *Proc. IEEE WCNC*, 2006.

[Pet07] L. L. Peterson and B. S. Davie, *Computer Networks – A Systems Approach*, Morgan Kaufman, 2007.

[Pet61] W. W. Peterson and D. T. Brown, Cyclic codes for error detection. *Proc. Inst. Radio Eng.*, **49**, 228–235 (1961).

[PolW] Polaris Wireless: http://www.polariswireless.com.

[Por01] D. Porcino, Performance of a OTDOA-IPDL positioning receiver for 3gpp-fdd mode,*Second International Conference on 3G Mobile Communication Technologies*, March 2001, pp. 221–225.

[Pro00] J. G. Proakis, *Digital Communications*, 4th edn, McGraw Hill, 2001.

[Qua01] Qualcomm Inc., 1x Evolution IS-856 TIA/EIA Standard, Airlink Overview, Rev. 7.2, November 7, 2001.

[Rad85] R. Perlman, An algorithm for distributed computation of a spanning tree in an extended LAN, *ACM SIGCOMM Comput. Commun. Rev.*, **15** (4), 44–53 (1985).

[RadCa] http://www.uswcorp.com/USWCMainPages/Applications/E-911.htm.

[Ram08] B. Raman and K. Chebrolu, Censor networks: a critique of 'sensor networks' from a systems perspective, *ACM SIGCOMM Comput. Commun. Rev.*, **38** (3), 75–78 (2008).

[Rap03] T. S. Rappaport, *Wireless Communications Principles and Practice*, Prentice Hall, 2003.

[Red95] S. Redl, M. K. Weber, M. Oliphant, and W. Mohr,*An Introduction to GSM*, The Artech House Mobile Communications Series, Artech House, 1995.

[Ree98] J. H. Reed, K. J. Krizman, B. D. Woerner, and T. S. Rappaport, An overview of the challenges and progress in meeting the e-911 requirement for location service, *IEEE Commun. Mag.*, **36** (4), 30–37 (1998).

[Sai05] J. M. Sailor and J. R. Link, 'Smart dust': nanostructured devices in a grain of sand, *Chem. Commun.*, 1375–1383, 2005.

[Sam06] M. Sama *et al.*, 3dID: a low-power, low-cost hand motion capture device,*Proc. IEEE Conference on Design, Automation and Test in Europe*, 2006.

[San05] P. Santi, Topology control in wireless ad hoc and sensor networks,*ACM Comput. Surveys*, **37** (2), 164–194 (2005).

[Sav01] A. Savvides, C.-C. Han, and M. B. Srivastava, Dynamic fine-grained localization in ad-hoc networks of sensors, *Proc. Mobicom*, 2001, pp. 166–179.

[Sch00] M. Z. Win and R. A. Scholtz, Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications, *IEEE Trans. Commun.*, **48** (4), 679–689 (2000).

[Sch87] M. Schwartz, *Telecommunication Networks: Protocols, Modeling and Analysis*, Addison-Wesley Series in Electrical & Computer Engineering, Addison-Wesley, 1987.

[Sensinode] Sensinode website: http://www.sensinode.com.

[Sha03] R. C. Shah *et al.*, Data MULEs: modeling a three-tier architecture for sparse sensor networks, *Proc. IEEE Workshop on Sensor Network Protocols and Applications*, 2003.

[Sha04] S. Sai Shankar *et al.*, Optimal packing of VoIP calls in an IEEE 802.11a/e WLAN in the presence of QoS constraints and channel errors, *Proc. IEEE Globecom*, 2004, pp. 2974–2980.

[Sha48] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27**, 379–423, 623–656 (1948).

[Sho03] G. Shor, TG3a-Wisair-CFP-Presentation, DS-UWB Proposal, IEEE P802.15 Working Group for Wireless Personal Area Networks, IEEE 802.15-03/151r3, May 2003.

[Sib02] G. Sibley, M. Rahimi, and G. Sukhatme, Robomote: a tiny mobile robot platform for large-scale ad hoc sensor networks,*Proc. International Conference on Robotics and Automation*, 2002.

[Sie00] T. Siep, I. Gifford, R. Braley, and R. Heile, Paving the way for personal area network standards: an over view of the IEEE P802.15 Working Group for Wireless Personal Area Networks, *IEEE Personal Commun.*, **7** (2), 37–43 (2000).

[Si100] R. D. Silverman,A cost-based security analysis of symmetric and asymmetric key lengths, RSA Bulletin Number 13, April 2000.

[Sim85] M. K. Simon *et al.*, *Spread Spectrum Communication*, Computer Science Press, 1985.

[Siv04] F. Sivrikaya and B. Yener, Time synchronization in sensor networks: a survey, *IEEE Network*, **18** (4), 45–50 (2004).

[Skl01] B. Sklar,Prentice Hall, Upper Saddle River, NJ, 2001.

[Sko08] D. Skordoulis *et al.*, IEEE 802.11n MAC frame aggregation mechanisms for next generation high throughput WLANs, *IEEE Wireless Commun.*, **15** (1), 40–47 (2008).

[Son94] H.-L. Song, Automatic vehicle location in cellular communication systems, *IEEE. Trans. Veh. Technol.*, **43** (4), 902–908 (1994).

[Srm04] C. Siva Rama Murthy and B. S. Manoj, *Ad Hoc Wireless Networks: Architectures and Protocols*, Pearson, 2004.

[Sta98] W. Stallings, *Cryptography and Network Security*, Prentice Hall, 1998.

[Sta00] W. Stallings, *Local and Metropolitan Area Networks*, 6th edn, Pearson, 2000.

[Sta03] W. Stallings, *Network Security Essentials*, 2nd edition, Prentice Hall, 2003.

[Sti95] D. Stinson, *Cryptography: Theory and Practice*, CRC Press, 1995.

[Sti02] D. Stinson, *Cryptography: Theory and Practice*, CRC Press, 2002.

[Sti03] D. R. Stinson, *Cryptography: Theory and Practice*, 3rd edn, Chapman & Hall/CRC Press, 2005.

[Sti05] D. Stinson, *Cryptography: Theory and Practice*, 3rd edn, Chapman & Hall/CRC Press, 2005.

[Sunspot] Sun Small Programmable Object Technology website: http://www.sunspotworld.com.

[Swa08] N. Swangmuang and P. Krishnamurthy, Location fingerprint analyses toward efficient indoor positioning, *Proc. Percom*, March 2008.

[Sze02a] R. Szewczyk *et al.*, An analysis of a large scale habitat monitoring application,*Proc. 3rd ACM Conference on Embedded Network Systems (SenSys)*, November 2004.

[Sze02b] R. Szewczyk *et al.*, Habitat monitoring with sensor networks, *Commun. ACM*, **47** (6), 34–40 (2004).

[Tak85] H. Takagi and L. Kleinrock, Throughput analysis for persistent CSMA systems, *IEEE Trans. Commun.*, **33**, 627–638 (1985).

[Tan97] A. S. Tanenbaum, *Computer Networks*, Prentice Hall, 1996.

[Tan03] A. S. Tanenbaum, *Computer Networks*, 4th edn, Prentice Hall, 2003.

[Tar98] V. Tarokh, N. Seshadri, and A. R. Calderbank, Space–time codes for high data rate wireless communications, *IEEE Trans. Inf. Theory*, **44** (2), 744–765 (1998).

[Tek98] S. Tekinay, E. Chao, and R. E. Richton, Performance benchmarking for wireless location systems, *IEEE Commun. Mag.*, **36** (4), 72–76 (1998).

[Tel95] E. Telatar, Capacity of multiantenna Gaussian channels, *AT&T Bell Labs Tech. Memo*, 1995.

[Tob75] F. A. Tobagi and L. Kleinrock, Packet switching in radio channels: Part II – the hidden terminal problem and the busy tone solution, *IEEE Trans. Commun.*, **23**, 1417–1433 (1975).

[Tob80] F. A. Tobagi, Multi-access protocol in packet communication systems, *IEEE Trans. Commun.*, **COM-28**, 468–488 (1980).

[Tho06] R. Thompson *et al.*, *The Physical Layer of Communication Systems*, Artech House, 2006.

[Tiny] The TinyOS Community Forum: http://www.tinyos.net/.

[Tol05] G. Tolle *et al.*, A macroscope in the redwoods,*4th ACM Conference on Embedded Network Systems (SenSys)*, November 2005.

[TR-45-02] TR-45, Enhanced Wireless 911 Phase 2, TIA/EIA-J-STD-036-A, Revision A, March 2002.

[Trev04] E. Trevisani and A. Vitaletti, Cell-ID location technique, limits, and benefits: an experimental study,*Proc. 6th IEEE Workshop on Mobile Computing Systems and Applications, WMCSA'04*, 2004.

[Tuc91] B. Tuch, An ISM band spread spectrum local area network: WaveLAN,*Proc. 1st IEEE Workshop on WLANs*, Worcester, MA, 1991, pp. 103–111.

[Unb02] M. Unbehaun,On the design and deployment of low-cost wireless infrastructure, Ph.D. Dissertation, Royal Institute of Technology, Sweden, 2002.

[Unb03] M. Unbehaun and M. Kamenetsky, On the deployment of picocellular wireless infrastructure, *IEEE Wireless Commun.*, **10** (6), 70–80 (2003).

[Ung87] G. Ungerboeck, Trellis coded modulation with redundant signal sets, *IEEE Commun. Mag.*, **25** (2), 5–21 (1987).

[Val98] R. T. Valadas, A. R. Tavares, A. M. deO. Duarte, A. C. Moreira, and C. T. Lomba, The infrared physical layer of the IEEE 802.11 standard for wireless local area networks, *IEEE Commun. Mag.*, **36** (12), 107–112 (1998).

[Vas05] D. Vassis, G. Kormentzas, A. Rouskas, and I. Maglogiannis, The IEEE 802.11g standard for high data rate WLANs, *IEEE Network*, **19** (3), 21–26 (2005).

[vNee99] R. van Nee *et al.*, New high-rate wireless LAN standards, *IEEE Commun. Mag.*, **37** (12), 82–88 (1999).

[WAC08] http://www.mobilewhack.com/3-million-iphone-3g-sold-estimated/.

[Wal99] B. H. Walke, *Mobile Radio Networks – Networking and Protocols*, John Wiley and Sons, Ltd, 1999.

[Wan05] W. Wang, S.C. Liew, V. O. K. Li, Solutions to performance problems in VoIP over a 802.11 wireless LAN, *IEEE Trans. Veh. Technol.*, **54** (1), 366–384 (2005).

[War03] J. Warrior, E. McHenry, and K. McGee, They know where you are, *IEEE Spectrum*, **40** (7), 20–25 (2003).

[War97] A. Ward, A. Jones, and A. Hopper, A new location technique for the active office, *IEEE Personal Commun.*, **4** (5), 42–47 (1997).

[Wel03] M. Welborn, M. McLaughlin, P. Ceva, and R. Kohno,DS-UWB Proposal, IEEE P802.15 Working Group for Wireless Personal Area Networks (WPANs), Document number: IEEE 802.15-03/334r3, September 2003.

[Wer98] J. Werb and C. Lanzl, Designing a positioning system for finding things and people indoors, *IEEE Spectrum*, **35** (9), 71–78 (1998).

[Wet01] M. Jakobsson and S. Wetzel, Security weaknesses in Bluetooth,*RSA Conference'01*, April 8–12, 2001.

[Whe07] A. Wheeler, Commercial applications of wireless sensor networks using ZigBee, *IEEE Commun. Mag.*, **45** (4), 70–77 (2007).

[Whi05] A. Whitaker and D. Newman, *Penetration Testing and Network Defense*, Cisco Press, 2005.

[Wil95] T. A. Wilkinson, T. Phipps, and S. K. Barton, A report on HIPERLAN standardization, *Int. J. Wireless Inf. Networks*, **2**, 99–120 (1995).

[Wil95b] J. E. Wilkes, Privacy and authentication needs of PCS, *IEEE Personal Commun.*, **2** (4), 11–15 (1995).

[Win00] M. Z. Win and R. A. Scholtz, Ultra-wide bandwidth time-hopping spread spectrum impulse radio for wireless multiple-access communications, *IEEE Trans. Commun.*, **48** (4), 679–691 (2000).

[Wol82] J. K. Wolf, A. M. Michelson, and A. H. Levesque, On the probability of undetected error for linear block codes, *IEEE Trans. Commun.*, **30**, 317–324 (1982).

[Woo03] A. Woo, T. Tong, D. Culler, Taming the underlying challenges of reliable multihop routing in sensor networks, *ACM Sensys*, 2003.

[Wor91] *First IEEE Workshop on WLANs*, Worcester, MA, 1991.

[Xbow] Crossbow website: http://www.xbow.com.

[Xia05] Y. Xiao, IEEE 802.11n: enhancements for higher throughputs in wireless LANs, *IEEE Wireless Commun.*, **12** (6), 82–91 (2005).

[Xu04] N. Xu *et al.*, A Wireless Sensor Network for Structural Monitoring,*Proc. 3rd ACM Conference on Embedded Network Systems (SenSys)*, November 2004.

[Yal02] R. Yallapragada, V. Kripalani, and A. Kripalani, EDGE: a technology assessment,*IEEE International Conference on Personal Wireless Communications*, December 2002.

[You06] O. Younis, M. Krunz, and S. Ramasubramanian, Node clustering in wireless sensor networks: recent developments and deployment challenges, *IEEE Network*, **20** (3), 20–25 (2006).

[Zah00] A. Zahedi and K. Pahlavan, Capacity of a wireless LAN with voice and data services, *IEEE Trans. Commun.*, **48** (7), 1160–1170 (2000).

[Zah97] A. Zahedi and K. Pahlavan, Terminal distribution and the impacts of natural hidden terminal, *Electron. Lett.*, **33** (9), 750–751 (1997).

[Zen00] M. Zeng, A. Annamalai, and V. K. Bhargava, Harmonization of global third-generation mobile systems, *IEEE Commun. Mag.*, **38** (12), 94–104 (2000).

[Zha90] K. Zhang and K. Pahlavan, An integrated voice/data system for mobile indoor radio networks, *IEEE Trans. Veh. Technol.*, **39**, 75–82 (1990).

[Zha92] K. Zhang and K. Pahlavan, Relation between transmission and throughput of slotted ALOHA local packet radio networks, *IEEE Trans. Commun.*, **40**, 577–583 (1992).

APPENDICES

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

APPENDIX A: WHAT IS DECIBEL?

Decibel or dB is usually the unit employed to compute the logarithmic measure of power and power ratios. The reason for using decibel is that all computation reduces to addition and subtraction rather than multiplication and division. Every link, node, repeater, or channel can be treated as a *black box* (see Fig. A.1*a*) with a particular decibel gain. The decibel gain of such a black box is given by

$$\text{dB gain} = 10\log\left(\frac{\text{power of output signal}}{\text{power of input signal}}\right) = 10\log\left(\frac{P_{\text{out}}}{P_{\text{in}}}\right) \qquad \text{(A.1)}$$

This corresponds to the *relative* output power with respect to the input power. The logarithm is always to the base 10. If the ratio in Eq. (A.1) is negative, then it is a decibel loss.

The decibel gain relative to an absolute power of 1 mW is denoted by dBm. For example, if the input power is 50 mW, relative to 1 mW, then the input power is 10 log(50 mW/1 mW) = 16.98 dBm. If this is followed by a link having a loss of 10 dB, then the absolute power at the output of the link will be 16.98 dBm – 10 dB = 6.98 dBm (see Fig. A.1*b*).[1]

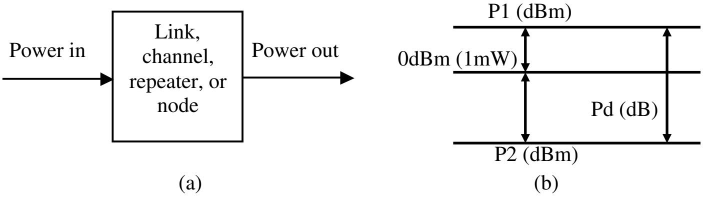


FIGURE A.1 Decibel: (a) Overall concept; (b) relation between dB and dBn.

[1] Antenna gains are represented similarly with respect to an isotropic antenna (which radiates with a gain of unity in all directions) or a dipole antenna. The former gain is in units of dBi and the latter in units of dBd. The units in dBi are 2.15 dB larger than the units in dBd.

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

APPENDIX B: STC FOR TWO TRANSMITTERS AND ONE RECEIVER

Figure B.1 shows the basic operation of the traditional MRC with one transmitter and two receiver antennas. The transmitted symbol s_0 in the constellation arrives at time T at two antennas with different random gains $\{h_0,\ h_1\}$. The received sampled signal from two antennas is given by

$$\begin{aligned} z_0(T) &= s_0h_0 + \eta_0 \\ z_1(T) &= s_0h_1 + \eta_1 \end{aligned}$$

where $\{\eta_0,\eta_1\}$ is the noise added by each channel. In MRC, each received signal is scaled with the signal strength of that antenna, which is accomplished by multiplying the first branch by the estimated value of h_0^* and the second branch by the estimated value of h_1^*. As shown in Fig. B.1, when we add the scaled branches of the diversity branches we will have the received signal

$$z(T) = s_0(|h_1|^2 + |h_0|^2) + h_1^*\eta_1 + h_0^*\eta_0 \tag{B.1}$$

The received points in the constellation are scaled by $|h_1|^2 + |h_0|^2$ and distorted by $h_1^*\eta_1 + h_0^*\eta_0$.[1]

To show the basic concept of STC we now describe the simple two transmit and one received antenna block coding system known as the Alamouti coded STC system [Ala98]. Figure B.2 illustrates the operation of the two-transmit and one-receive antenna system. Alamouti showed a simple transmission and detection scheme that reproduces Eq. (B.1) with one-transmit and two-received antenna systems. Alamouti's transmitted block code operates on a sequence of two symbols $\{s_0,s_1\}$, as shown in Fig. B.2, this sequence being time coded into two sequences $\{s_0, -s_1^*\}$ and $\{s_1, s_0^*\}$ to be then space coded by the first and second antennas respectively. Each of these two sequences has two symbols that are

[1] The variance of this noise is $(|h_1|^2 + |h_0|^2)\sigma_\eta^2$

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

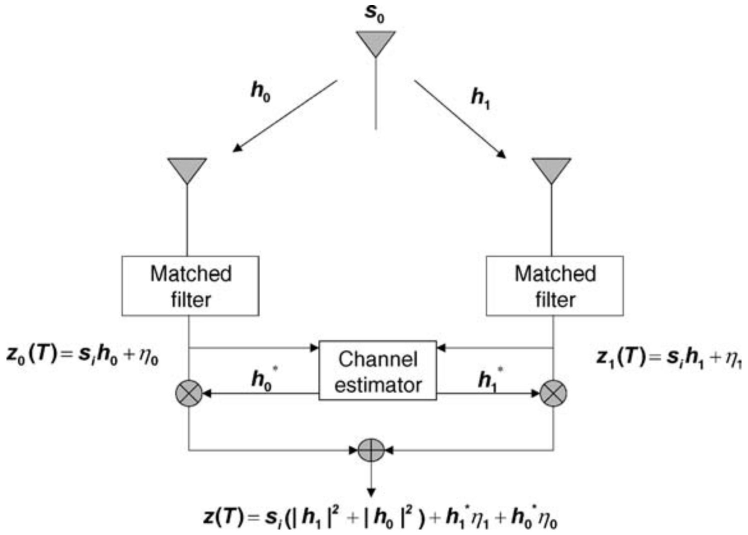

s_0
h_0
h_1
Matched filter
Matched filter
$z_0(T) = s_i h_0 + \eta_0$
Channel estimator
h_0^*
h_1^*
$z_1(T) = s_i h_1 + \eta_1$
$z(T) = s_i(|h_1|^2 + |h_0|^2) + h_1^* \eta_1 + h_0^* \eta_0$

FIGURE B.1 Implementation of the MRC for two transmit and one received antenna.

transmitted in two consecutive time slots of one of the antennas in parallel with transmission of the other sequence in the other antenna. Between the two transmitter antennas and one receiver antenna we have two channel gain factors h_0 and h_1 that are samples of two independent random processes representing the fading characteristics of the channel. Since all mobile channels are slow fading channels, the value of the channel gains during

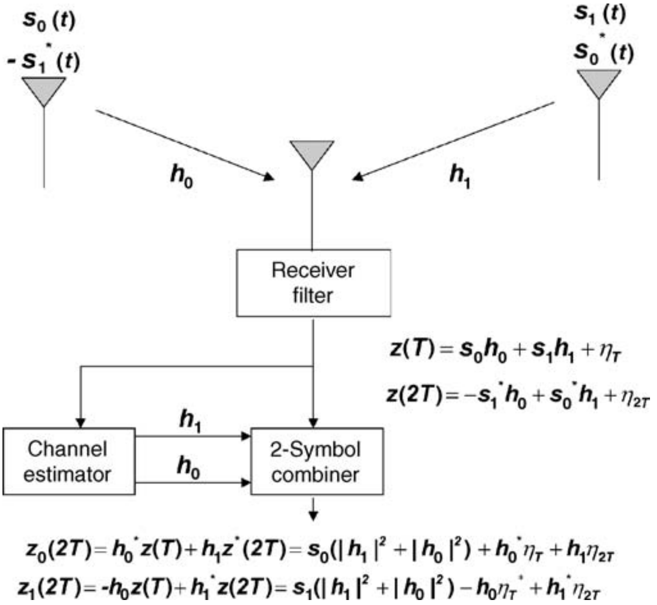

$s_0(t)$
$-s_1^*(t)$
$s_1(t)$
$s_0^*(t)$
h_0
h_1
Receiver filter
$z(T) = s_0 h_0 + s_1 h_1 + \eta_T$
$z(2T) = -s_1^* h_0 + s_0^* h_1 + \eta_{2T}$
Channel estimator
h_1
h_0
2-Symbol combiner
$z_0(2T) = h_0^* z(T) + h_1 z^*(2T) = s_0(|h_1|^2 + |h_0|^2) + h_0^* \eta_T + h_1 \eta_{2T}$
$z_1(2T) = -h_0 z(T) + h_1^* z(2T) = s_1(|h_1|^2 + |h_0|^2) - h_0 \eta_T^* + h_1^* \eta_{2T}$

FIGURE B.2 Simple Alamouti code for two transmit and are receive antennas.

transmission of two symbols remains the same. Therefore, the received signals after sampling at the receiver for the first and the second symbols are given by

$$\begin{aligned} z(T) &= s_0 h_0 + s_1 h_1 + \eta_T \\ z(2T) &= -s_1^* h_0 + s_0^* h_1 + \eta_{2T} \end{aligned}$$

To form the decision variables for the two transmitted symbols, Alamouti suggests the following transformation on the received two samples to form two new variables observed at the end the two intervals:

$$\begin{aligned} z_0(2T) &= h_0^* z(T) + h_1 z^*(2T) = s_0(|h_1|^2 + |h_0|^2) + h_0^* \eta_T + h_1 \eta_{2T} \\ z_1(2T) &= -h_0 z(T) + h_1^* z(2T) = s_1(|h_1|^2 + |h_0|^2) - h_0 \eta_T^* + h_1^* \eta_{2T} \end{aligned}$$

The statistical behavior of these decision variables is identical and they are the same as the statistical behavior of the decision variable of the traditional MRC receiver shown by Eq. (B.1). Consequently, and not surprisingly, the results of simulations for two transmit and one received antennas provided by Alamouti [Ala98] are identical to the results of the analysis of MRC for one transmit antenna and two receiver antennas we presented in Eq. (3.12).

APPENDIX C: SOURCE CODING

Basic source coding techniques are divided into two classes of lossless and lossy compression. As the names indicate, lossless source coding refers to techniques in which data is compressed such that no part of the original data is lost in the compression. On the other hand, lossy source coding refers to the compression of data where ‘‘truncating’’ is possible and data can be compressed more effectively. Lossless source coding is, in general, reversible, while received data after lossy source coding is not the same as that before coding. In this section we introduce several source coding schemes used for transmission of audio/video payloads over modern networks.

C.1 SOURCE CODING FOR VOICE

There are numerous source coding techniques used in telecommunication applications to transmit voice data. The term speech coding in general refers to describe this class of source coding. In this section we describe few examples of such source coding schemes for existing speech communication applications.

Linear Predictive Coding Voice Codes. Linear predictive coding (LPC) is an application of adaptive filters in speech coding techniques. For such coding techniques, a discrete model of human speech production is required. The fundamental idea in LPC is to combine the previous samples of a speech signal and predict the current sample at low bit rates. This seemingly simple problem is well discussed in the realm of adaptive filters and is one the most useful methods to encode good quality speech signals.

To model human speech production, the linear predictive coder considers the voiced speech sound signal constructed by a buzzer at the end of the glottal filter which is followed by a vocal tract filter. The coder then analyzes the speech signal by removing the effects of the vocal tract filter and glottal filter and estimating the amplitude and frequency of the remainder of the speech signal. The speech signal parameters and the residue of the speech signal can then be transmitted to the receiver side. At the receiver side, the decoder constructs the filter with the estimated parameters of the filters and runs the residue of the signal through the filters, which results in the original speech signal.

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

Linear predictive source coding is used in cellular networks as a form of voice compression. As an example, the GSM standard utilizes such a codec to encode a 3.1 kHz voice signal into 5.6 kb/s (half rate) and 13 kb/s (full rate) digitized signals.

G.711 and G.726. G.711 is an ITU-T standard for voice compression which is widely used in circuit-switched telephone networks. This source coding scheme represents logarithmic pulse code modulation samples.

The two logarithmic companders used in G.711 are known as the μ-law algorithm and the A-law algorithm, with the latter specifically designed for computer processing. A G.711 encoder constructs a 64 kb/s signal from a speech signal sampled at 8 kHz. The μ-law algorithm encodes a 14-bit pulse coded signal into 8-bit samples, while the A-law algorithm encodes 13-bit PCM signals into the same 8-bit samples.

Other ITU-T standards utilize different coding schemes to compress the voice signal. For example, the G.726 standard uses adaptive differential pulse code modulation to compress the signal. In adaptive pulse coding, the size of the quantization step changes from one sample to the next. This allows the speech signal to be mapped into 4-bit samples, which instantly doubles the capacity of the medium.

Example C.1: μ-Law Example G.711 uses the μ-law as its voice compression technique. The essence of such compression is to give more weight to samples that are close to zero. In other words, the intervals for classification of the input signals are more compact when the signal is close to zero. The actual formula used for the μ-law is

$$y = \operatorname{sgn}(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)}$$

The second class of coding techniques used for audio source coding is referred to as analysis by synthesis coding. This class of source coding schemes relies mainly on closed-loop analysis of the optimal parameters by synthesis of the speech signal and, hence, it includes a decoder inside the encoder. Principally, the closed loop forces the encoder to try all the possible combinations of the bit patterns and select the pattern with the highest quality of decoded speech. Using the LPC approach, the coefficients of the speech signal are computed and stored. Then various bit patterns are inserted to the model and their outputs are computed and compared with the original speech signal. The best matching bit pattern is then considered as the data to be transferred. Numerous applications of analysis by synthesis coding are applied for audio signal compression, which we will discuss in this section.

G.728. Similar to other ITU-T standards, G.728 specifies a set of regulations for audio source coding at 16 kb/s. The main source coding algorithm in G.728 is low-delay code-excited linear prediction. To attain low-delay characteristics of the original code-excited linear prediction approach, the fundamentals of analysis by synthesis are untouched, but other parts of the algorithms have been modified. G.728 achieves delay of only five samples (0.625 ms) and a 50th-order LPC filter is utilized to model the parameters of the speech signal. G.728 is used in modems and networks. In practice, the filter coefficients, the gain of the bit pattern, and the pointer to the location of the bit pattern are transmitted over the medium. The rate of compression is about four times of the rate of G.711.

G.723.1. Used in VoIP applications, G.723.1 achieves very low bandwidth requirements and high compression rates compared with other audio source coding schemes. The two different bit rates of G.723.1 are achieved by two different algorithms. The first bit rate, 6.3 kb/s, uses frames with a length of 24 bytes and utilizes multipulse LPC with a maximum likelihood quantization algorithm. The second bit rate, 5.3 kb/s, uses frames of length 20 bytes and applies an algebraic code-excited linear prediction algorithm. The bit patterns used in such an encoder's algorithm have algebraic structure applied and, hence, the algorithm can store larger sets of bit patterns. A similar standard, G.729, also utilizes the algebraic coding scheme with silence insertion description frames. G.729 has a delay of 15 ms and supports data rates of 8, 6.4 and 11.8 kb/s. The extension of G.729.B utilizes voice activity detection based on analysis of the speech signal and sends silence insertion frames when the channel is inactive.

IS.733. Used in CDMA cellular networks, Qualcomm code-excited linear prediction is another alteration of the original code-excited linear predictive approach. With such an encoder, the rate of compression is variable, but the two rates of 8 and 13 kb/s are the most common rates in practice. The encoder includes two main filters (including a 10th-order LPC filter) for adjusting to the parameters of the speech signal and also includes a gain control block. The signal is transmitted with different rates corresponding to different types of speech signal. For example, speech is encoded at 8 kb/s, while silence and background noise are encoded at lower rates.

GSM-Enhanced Full Rate and GSM-Adaptive Multirate. Adopted in GSM cellular systems, GSM-enhanced full rate (GSM-EFR) is the speech coding scheme used to improve the quality of GSM voice signals. This technique utilizes an algebraic algorithm. The overall rate of the signal in GSM-EFR coding technique is 12.2 kb/s with support of discontinuous transmission. In addition to GSM-EFR, GSM-adaptive multirate (GSM-AMR) is used in 3G cellular networks. Principally, the rate of transmission and type of coding is changed based on the quality of the link between the mobile node and BS, denoted by link adaptation. The main coding scheme used in GSM-AMR is also algebra based, allowing total of 14 different models. The GSM-AMR frames usually contain 160 samples in 20 ms. The bit rate of such a scheme varies between 12.2 and 4.75 kb/s according to type of channel and background noise. Such coding schemes are referred to as hybrid source coding schemes, as they adjust the coding technique according to quality of the link.

C.2 SOURCE CODING FOR IMAGES AND VIDEO

Like source coding techniques for audio, there exist numerous source coding techniques used in telecommunication applications to transmit video data. The main goal of these schemes is to compress the video data (video compression) to be sent over a network. The compression makes it bandwidth efficient in wireless networks, communication over cable, and satellite communications. Like audio source coding, video source coding can be classified as lossless and lossy.

JPEG. In its simplest form, video compression includes coding techniques to reduce the size of an image. The most well-known coding technique to compress an image is the Joint Photographic Experts Group (JPEG) standard. Being far superior to its ancestors in quality, JPEG is a lossy compression scheme. In principle, the image is first broken into 8 × 8 pixel blocks. The frequency-domain components of each block are then calculated using a *discrete cosine transform*. The amplitude and frequency components of such a transformation are then quantized with a nonlinear quantizer to weight the low-frequency components with more accuracy. The resulting blocks of 8×8 pixels transformed to frequency are then compressed more with a lossless variant of Huffman coding. Different realizations of JPEG utilize different downsampling techniques in the process without drastically altering the quality of the image. The numerous features make this source coding scheme popular for use on the Web. In general, high compression rates without losing most of the high-resolution information of the image will result in a 20 times reduction in the size of the image.

PNG. Another compression technique widely used in today's networks is the portable network graphics (PNG) format. PNG is a lossless image compression technique designed for the specific needs of the Internet. The basic encoding scheme takes place by combining a variant of the Lempel–Ziv–Welch (LZW) compression format and the Huffman coding scheme known as *deflate*.

MPEG. Moving Pictures Experts Group (MPEG)-based standards are widely employed audio/video source coding schemes and have been used in a variety of applications ranging from video CD (VCD) playback technique to various schemes for broadcasting video over networks. The first generation of MPEG coding schemes, MPEG-1, was primarily used in VCD devices with limited coding abilities, and before the introduction of the MPEG-2 it was used in satellite and cable TV as well. MPEG-1 supports multiple sample rates of 32, 44.2, and 48 kHz. It also includes four different stereo encoding types of mono, mono with duplicated tracks, stereo, and joint encoded stereo. Three separate levels of encoding are considered in the MPEG-1 scheme. The encoder in MPEG-1 consists of a polyphase filter bank, layer-specific frequency mapping, and layer-specific bit packing. The filter bank consists of 32 filters with a 512 sampling window with Hanning window postprocessing. Layer-1 of MPEG-1 is designed for fast compression, which results in inaccurate and low compression rates. It consists of 384 samples per frame which is passed to an FFT algorithm. The transformed frame is then analyzed to estimate its amplitude. Layer-2 and Layer-3 algorithms are more sophisticated algorithms designed for better compression rates. Each frame consists of 1152 bits with two 1024 size windows per frame. The FFT algorithm then converts the central part of the output of the windows and LPC is used to extract the signal parameters. While MPEG-1 could support data rates of up to 1.5 Mb/s, an extension of it, MPEG-2, is now used for streaming with higher data rates. DVD-quality video is supported with MPEG-2 coding schemes. Similar to MPEG-1, MPEG-2 is a lossy compression scheme, but it can achieve data rates of 3.5 Mb/s up to 6 Mb/s and it is used in digital TV via satellite and cable.

H.120 and H.26x. The first digital video standard introduced is known as H.120 by ITU-T. Working with a differential PCM coding scheme and scalar quantization with variable-length coding, H.120 achieved 1.5 Mb/s for NTSC and 2 Mb/s for PAL systems. However, it was shown that the quality of the output of such a scheme is poor. The next generation of H.120, known as H.261, was designed to support different data rates of 40 kb/s up to 2 Mb/s. The coding scheme in H.261 uses a macroblock consisting of matrices of the luma and chroma samples of each picture. After removing the redundancies of the adjacent frames, a motion vector is employed to illustrate the motion. Like the MPEG coding scheme, a discrete cosine transform is then used to convert the signal and extract its frequency information. The frame is quantized and the matrix coded and transformed into bits to be transmitted.

APPENDIX D: ACRONYMS

1G	first-generation
2G	second-generation
3G	third-generation

A

AAL	ATM adaptation layer
ACF	autocorrelation function
ACL	access control list
ACL	asynchronous connectionless
ADPCM	adaptive differential PCM
AES	advanced encryption standard
A-FLT	advanced forward link trilateration
AGPS	assisted GPS
AH	authentication header
AMI	alternate mark inversion
AMPS	Advanced Mobile Phone Services
ANSI	American National Standards Institute
AOA	angle of arrival
AODV	ad-hoc on-demand distance vector
AP	access point
ARIB	Association of Radio Industries and Businesses
ARP	address resolution protocol
ARQ	automatic repeat request
AS	autonomous system
ASK	amplitude-shift keying
ASN	access service network

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and
Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

ATIM	announcement TIM
ATM	asynchronous transfer mode
AuC	authentication center
AWGN	additive white Gaussian noise

B

BAN	body-area network
BCH	Bose–Chaudhuri–Hocquenghem
BER	bit-error rate
BGP	border gateway protocol
BI	back-off interval
BPDU	bridge protocol data unit
BPSK	binary phase shift keying
BRAN	broadband radio access network
BS	base station
BSA	basic service area
BSC	BS controller
BSS	basic service set
BSS	BS subsystem
BTMA	busy-tone multiple-access
BTS	base transceiver system

C

CAS	call-associated signaling
CBC	cipher-block-chaining
CCA	clear channel assessment
CCK	complementary code keying
CCMP	counter-mode cipher-clock-chaining MAC protocol
CDDI	copper distributed data interface
CDM	code-division multiplexing
CDMA	code-division multiple access
CDPD	cellular digital packet data
CEPT	Committee of the European Post and Telecommunications
CFI	canonical format indicator
CFP	contention-free period
CHAP	challenge handshake authentication protocol
CLP	cell loss priority
CM	connection management
COFDM	coded OFDM
COST	Co-operative for Scientific and Technical Research
CRC	cyclic redundancy check

CSMA	carrier sense multiple access
CSMA/CA	CSMA with collision avoidance
CSMA/CD	CSMA with collision detection
CSN	connectivity service network
CT-2	cordless telephone-2
CTS	clear-to-send
CW	contention window
CWINS	Center for Wireless Information Network Studies

D

DARPAnet	Defense Advanced Research Projects Agency Department Network
DBPSK	differential BPSK
DCF	distributed coordination function
DCLA	DC level adjustment
DDoS	distributed DoS
DDP	dominant direct path
DDS	digital data service
DECT	digital enhanced (*formerly* European) cordless telephony
DES	data encryption standard
DFIR	diffuse IR
DFS	dynamic frequency selection
DGPS	differential GPS
DH	data high
DHCP	dynamic host configuration protocol
DIFS	DCF inter-frame space
DL	discrete logarithm
DLL	data link layer
DLOS	direct LOS
DM	data medium
DMZ	demilitarized zone
DNS	domain name service
DOA	direction of arrival
DoS	denial of service
DPAM	demand priority access method
DQPSK	differential QPSK
DS	direct sequence
DSL	digital subscriber line
DSMA	digital/data sense multiple-access
DSS	digital signature standard
DSSS	direct sequence spread spectrum

DS-UWB	direct sequence UWB
DV	data voice
DVCC	digital verification color code

E

EAP	extensible authentication protocol
EDGE	Enhanced Data for Global Evolution
E-FLT	enhanced forward link trilateration
EGP	external gateway protocol
ELID	emergency location information delivery
E-OTD	enhanced observed time difference
ERP	extended rate physical
ERP	exterior router protocol
ESA	extended service area
ESME	emergency services message entity
ESN	emergency services network
ESNE	ESN entity
ESP	encapsulated security payload
ESS	extended service set
ETSI	European Telecommunication Standards Institute
EWC	Enhanced Wireless Consortium
EY-NPMA	elimination-yield non-preemptive multiple access

F

FCC	Federal Communication Commission
FCS	frame correction sequence
FDD	frequency-division duplexing
FDDI	fiber distributed data interface
FDM	frequency-division multiplexing
FDMA	frequency-division multiple access
FEC	forward error correction
FEXT	far end cross talk
FFD	full-function device
FFT	fast Fourier transform
FH	frequency hopping
FHS	frequency hop synchronization
FHSS	frequency-hopping spread spectrum
FM	frequency modulation
FMS	Fluhrer, Mantin, Shamir
FSK	frequency-shift keying
FSTC	Financial Services Technology Consortium

FTP	file transfer protocol
FTP	foiled twisted pair
G	
GBS	geolocation BS
GFC	generic flow control
GFSK	Gaussian frequency-shift keying
GGSN	gateway GPRS support node
GMLC	gateway mobile location center
GMSK	Gaussian minimum shift keying
GP	gap period
GPRS	general packet radio service
GPS	global positioning system
GSM	Global System for Mobile Communications
GSM-AMR	GSM-adaptive multirate
GSM-EFR	GSM-enhanced full rate
H	
HDLC	High-level data link control
HDR	High Data Rate
HFC	hybrid fiber cable
HLR	home location register
HPN	home phone networking
HPNA	Home Phone Network Alliance
HR	high rate
HTTP	hypertext transfer protocol
HV	high-quality voice
I	
IANA	Internet Assigned Numbers Authority
IAPP	inter-AP protocol
IBSS	independent BSS
ICMP	Internet control message protocol
IDS	intrusion detection system
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IFS	interframe spacing
IGP	interior gateway protocol
IHL	Internet header length
IIS	Internet Information Server (Microsoft)
IKE	Internet key exchange

IMSI	international mobile subscriber identity
IMT-2000	International Mobile Telecommunications beyond the year 2000
IP	Internet protocol
IPS	interpacket spacing
IPS	intrusion prevention system
IPX/SPX	internetwork packet exchange/sequenced packet exchange
IR	infrared
IRP	interior router protocol
ISDN	integrated service data network
ISI	intersymbol interference
IS-IS	intermediate system to intermediate system
ISM	industrial, scientific, and medical
ISO	International Standards Organization
ISP	Internet service provider
ITS	intelligent transportation system
ITU	International Telecommunications Union
IV	initialization vector

J

JPEG	Joint Photographic Experts Group
JTC	Joint Technical Committee

L

L2CAP	LLC and adaptation protocol
LAN	local-area network
LANE	LAN emulation
LAPD	link access protocol-D
LBT	listen-before-talk
LCS	location services
LED	light-emitting diode
LFSR	linear feedback shift register
LIF	Location Interoperability Forum
LLC	logical link control
LM	link manager
LMDS	local multipoint distribution system
LMP	link management protocol
LMU	location measurement unit
LOS	line-of-sight
LPC	linear predictive coding
LZW	Lempel–Ziv–Welch

M

MAC	medium access control
MAC	message authentication code
MAHO	mobile-assisted hand-off
MAN	metropolitan area network
MBOA	Multiband OFDM Alliance
MB-OFDM	multiband OFDM
M-BOK	multiple bi-orthogonal keying
MBWA	mobile broadband wireless access
MCM	multicarrier-modulation
MD	message digest
ME	mobile equipment
MIC	message integrity check
MIMO	multiple-input–multiple-output
MLME	MAC layer management entity
MLP	mobile location protocol
MM	mobility management
MMAC-PC	Multimedia Mobile Access Communications Promotion Council
MMF	multimode fiber
MOS	mean-opinion-score
MPC	mobile positioning center
MPDU	MAC protocol data unit
MPEG	Moving Pictures Experts Group
MPLS	multiprotocol label switching
MRC	maximal ratio combining
MS	mobile station
MSC	mobile switching center
MSK	minimum shift keying
MSS	mobile subscriber station
MTP	message transport protocol

N

NAMA	node activation multiple access
NAP	network access provider
NAV	network allocation vector
NB	normal burst
NCAS	non-call-associated signaling
NDDP	non-DDP
NEXT	near end crosstalk
NIC	network interface card

NIST	National Institute of Standards and Technology
NLOS	non-LOS
NMT	Nordic Mobile Telephone
NNI	network–network interface
NRZ	non-return to zero
NSP	network service provider
NSS	network and switching subsystem

O

OAM	operation, administration, and maintenance
OBEX	object exchange
OC-1	optical carrier 1
OFDM	orthogonal frequency division multiplexing
OFDMA	orthogonal frequency division multiple access
OMA	Open Mobile Alliance
OQPSK	offset QPSK
OSI	Open Systems Interconnection
OSPF	open shortest path first
OTDOA-IPDL	observed time difference of arrival – idle period on downlink

P

PAM	pulse amplitude modulation
PAN	personal-area network
PBCC	packet binary convolutional coding
PBX	private box exchange
PCF	point coordination function
PCM	pulse code modulation
PCMCIA	Personal Computer Memory Card International Association
PCS	personal communication service
PDA	personal digital assistant
PDE	position-determining entity
PDF	probability density function
PDN	public data network
PHP	personal handy phone
PHS	personal handy system
PIFS	PCF IFS
PIN	personal identification number
PLCP	physical layer convergence protocol
PLME	PHY layer management entity
PLR	packet loss rate

PLT	payload type
PLW	packet-length width
PMD	physical medium dependent
PN	pseudo noise
PNG	portable network graphics
POA	point of association
POTS	plain old telephone service
PPM	pulse position modulation
PPP	point-to-point protocol
PRF	pulse-rate frame
PRMA	packet reservation multiple access
PSAP	public safety answering point
PSDU	physical service data unit
PSF	packet-signaling field
PSTN	public switched telephone network

Q

QAM	quadrature amplitude modulation
QoS	quality of service
QPSK	quadrature phase-shift keying

R

RADIUS	remote authentication dial-in user service
R-ALOHA	reservation ALOHA
RF	radio-frequency
RFD	reduced-function device
RIP	routing information protocol
RMS	root-mean-square
RNC	radio network controller
RPC	remote procedure call
RRC	radio resource control
RRM	radio resource management
RS	Reed–Solomon
RSN	robust security network
RSS	received signal strength
RSVP	resource reservation protocol
RTMP	routing table maintenance protocol
RTP	real-time transport protocol
RTS	request-to-send
RZ	return to zero

S

SACCH	slow associated control channel
SAP	service AP
SCCP	signaling connection control part
SCI	sync capsule indicator
SCO	synchronous connection oriented
SDH	synchronous digital hierarchy
SDP	service discovery protocol
SFD	start of the frame delimiter
SGSN	serving GPRS support node
SIFS	short IFS
SIG	special interest group
SIM	subscriber identity module
SIMO	single-input–multiple-output
SISO	single input–single output
SMAC	sensor MAC
SME	station management entity
SMF	single-mode fiber
SMLC	serving mobile location center
SMS	short messaging system
SNR	signal-to-noise ratio
SOAP	simple object access protocol
S-OFDMA	scalable OFDMA
SOHO	small office and home office
SONET	synchronous optical network
SPIN	sensor protocol for information via negotiation
SSL	secure sockets layer
STA	spanning-tree algorithm
STC	space–time coding
STM-1	synchronous transport mode 1
STP	shielded twisted-pair
STS-1	synchronous transport signal 1

T

TB	tail bit
TCM	trellis-coded modulation
TCP	transmission control protocol
TCS	telephony control protocol specification
TDD	time-division duplexing
TDM	time-division multiplexing

TDMA	time-division multiple access
TDOA	time difference of arrival
TEEN	threshold-sensitive energy efficient sensor network
TF	time–frequency
TFMA	time–frequency multiple access
TIA	Telecommunications Industry Association
TIA/EIA	Telecommunication/Electronic Industry Association
TIM	traffic indication map
TKIP	temporal key integrity protocol
TLS	transport-layer security
TMSI	temporary mobile subscriber identity
TOA	time of arrival
TPC	transmit power control
TRAMA	traffic-adaptive medium access
TSN	transitional security network
U	
UDP	undetected direct path
UDP	user datagram protocol
UMTS	Universal Mobile Telecommunications System
UNI	user-network interface
U-NII	Unlicensed National Information Infrastructure
US-CERT	United States Computer Emergency Readiness Team
UTDOA	uplink TDOA
UTP	unshielded twisted-pair
UWB	ultra wideband
V	
VCD	video CD
VCI	virtual channel identifier
VG	voice-grade
VLAN	virtual LAN
VLR	visitor location registration
VoIP	voice-over-IP
VPI	virtual path identifier
VPN	virtual private network
W	
WAE	wireless application environment
WAN	wide area network
WAP	wireless access protocol
WARC	World Administrative Radio Conference

WAVE	wireless access for the vehicular environment
WCAN	wireless campus area network
W-CDMA	wideband CDMA
WDM	wavelength-division multiplexing
WEP	wired equivalent privacy
WLAN	wireless LAN
WMAN	wireless MAN
WPA	Wi-Fi protected access
WPAN	wireless personal area network
WPI	Worcester Polytechnic Institute
XML	extensible markup language

Z

ZDO	ZigBee device object

APPENDIX E: LIST OF VARIABLES

CHAPTER 2

f	frequency
l	length of the cable
$\mathrm{G_p}(f, l)$	path-gain function
$\lambda = c/f$	wavelength
c	speed of light in vacuum
τ	delay
d	distance between transmitter and the receiver
L_p	path-loss
L_0	path-loss in first meter distance
α	distance-power gradient
x	lognormal shadow fading
σ	variance of the shadow fading
F_σ	fade margin in dB
h_b	BS antenna height
h_m	height of the mobile antenna
C_M	city correction factor for PCS extension of Akumura–Hata model
K_r	correction factor for suburban areas
f_c	center frequency
K	Rician distribution constant factor
f_m	maximum Doppler shift
v_m	mobile velocity
$D(\lambda)$	Doppler spectrum
β_i	amplitude of the multipath component
$P_i = E\{\beta_i^2\}$	power of the multipath component
τ_i^-	delay of the multipath component

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0

φ_{ie}	phase of the multipath component
θ_i	angle of arrival

CHAPTER 3

R_b	data transmission rate
R_s	symbol transmission rate
M	number of symbols
T_s	symbol time duration
p_i	probability of occurrence of a symbol
m_i	number of bits per symbol
m	average number of bits per symbol
E_{s_i}	energy per symbol
E_s	average energy per symbol
E_b	energy per bit
N_0	variance of the noise
d	Minimum distance in the constellation
$\gamma_s = E_s/N_0$	SNR
$\gamma_b = E_b/N_0$	SNR ratio per bit
P_s	probability of symbol transmission error
P_b	probability of bit transmission error
P_e	probability of error
$\bar{P}_e$	average probability of error
η	bandwidth efficiency
W	transmission bandwidth
C	channel capacity or maximum achievable data transmission rate
$R(k)$	autocorrelation function of a PN sequence

CHAPTER 4

B	bandwidth per carrier
n	LENGTH of coded block of data
k	a block of data used for coding
$R_\mathrm{c} = k/n$	coding rate
R_b	data rate
P_FD	probability of false detection
t_max	maximum number of correctable bits
p	bit transmission error rate
$\mathbf{I}$	information matrix
$\mathbf{C}$	coded information matrix
$\mathbf{R}$	received code word

$\mathbf{G}$	code generation matrix
$\mathbf{P}$	parity matrix
$\mathbf{H}$	parity check matrix
$\mathbf{e}$	error matrix
$\mathbf{S}$	syndrome matrix
P_i	probability of making i-bits error
$G(x)$	generator polynomial
$I(x)$	information polynomial
$P(x)$	parity or remainder polynomial
$R(x)$	received code polynomial
$e(x)$	error polynomial
$R(k)$	autocorrelation function of a code
$\mathbf{H}$	Hadamard matrix
$\mathbf{W_i}$	Walsh codeword

CHAPTER 5

W	transmission bandwidth
N_{p}	processing gain or bandwidth expansion factor of the CDMA
M	number of simultaneous users
S_{r}	SNR
P	transmission power
R_{b}	information bandwidth or data rate per user
G_{A}	sectored antenna gain in CDMA
G_{v}	voice activity factor in CDMA
H_0	interference increase factor in CDMA
k	performance improvement factor
B	bandwidth per carrier
m	number of users per carrier
N_{f}	frequency reuse factor
N_{u}	number of user channels in Erlang equations
ρ	normalized call density
λ	call arrival rate
μ	call service rate
$B(n, \rho)$	probability of blockage
$P(\text{delay} > t)$	probability of delay greater than t
D	average delay
G	aggregated traffic
S	throughput
a	normalized propagation delay

T_{p}	propagation delay
τ	propagation delay
T_{I}	average length of time a channel is idle
T_{a}	Average active period
T_{b}	The mean rate of blockage period
T_{th}	threshold time for discarding voice packets

CHAPTER 7

D_{L}	distance between co-channel cells
R	radius of a cell
S_{r}	signal-to-interference ratio
P_{t}	power of a MS
d	spatial distance
N_{f}	frequency reuse factor
L_0	path loss in the first meter
$P_{\mathrm{r}}(d)$	received power at distance d
J_{s}	number of interfering cell sites
R	radius of a cell
H_0	additional interference from adjacent channel
G_{A}	gain due to sectored cell

CHAPTER 8

T_{p}	packet length
T_{I}	average idle interval
p	probability of successful transmission
N	number of terminals
$A(N, p)$	probability of successful transmission in any slot
$P(k)$	probability of successful transmission in k - slots
T_{s}	length of a slot
T_{P}	length of the data packet
$\bar{E}$	average energy in the constellation
d	minimum distance in the constellation
m	average number of bits per symbol in the constellation

CHAPTER 9

$\mathbf{C}$	codeword
R_{av}	average data rate

R_{n}	one of the available data rates
p_{n}	probability of occurrence of a data rate

CHAPTER 10

r	ring of interference
d	distance between terminals
S_{r}	signal-to-interference ratio
P_X	power of terminal X
N	processing gain
Px	probability of occurrence of X
L_{IE}	length of 802 packet
L_{BS}	length of Bluetooth system packet
P_x	probability of x
n	number of Bluetooth hops
M	number of symbols

CHAPTER 11

x	plaintext
y	cyphertext
k	secret key
N_{u}	number of users
α	base number for calculation
ρ	a large prime number

CHAPTER 12

(x, y)	coordinates of a location
d	spatial distance
τ	propagation time
N_{r}	number of reference points
θ_{s}	width of the angle of an antenna array
$\mathbf{R}$	vector of RSS measurements
$\mathbf{S}$	vector of RSS samples
L	length of the database
D_i	Euclidean distance

INDEX

Networking Fundamentals: Wide, Local and Personal Area Communications Kaveh Pahlavan and Prashant Krishnamurthy © 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-99289-0