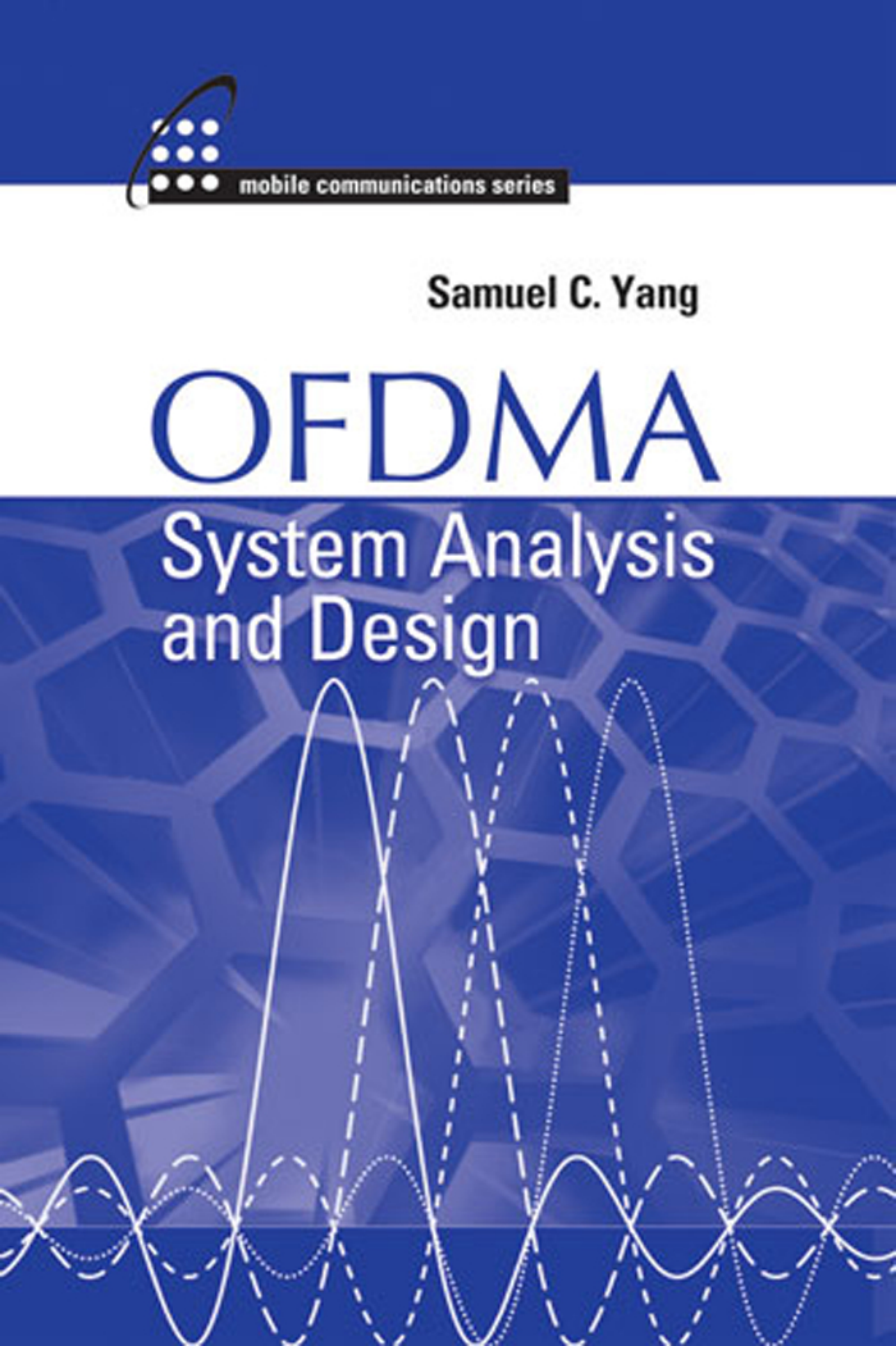

mobile communications series
Samuel C. Yang
OFDMA
System Analysis and Design

OFDMA System Analysis and Design

For a listing of recent titles in the
Artech House Mobile Communications Library,
turn to the back of this book.

OFDMA System Analysis and Design

Samuel C. Yang

ARTECH
HOUSE
BOSTON | LONDON
artechhouse.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the U.S. Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN-13: 978-1-60807-076-3

Cover design by Vicki Kane

685 Canton Street
Norwood, MA 02062

10 9 8 7 6 5 4 3 2 1

To my beautiful wife Jenny

Contents

CHAPTER 9

CHAPTER 10

Preface

In late 2001, Bill Gates was having dinner with a group of journalists on the night before Microsoft's launch of the Windows XP operating system. At that time, the dotcom crash was just unfolding, and during dinner, Bill Gates said (correctly) that the dotcom bubble was a distraction and caused a lot of money to be wasted on things that did not offer much innovation. Then Mr. Gates was asked if anything significant came out in the last few years.

"People will look back and say, 'Wow, at least they did 802.11,'" he was quoted as replying [1].

He is right. Most of us today cannot imagine working (and living) without 802.11-based wireless networks, popularly known as Wi-Fi. Traveling professionals who really needed to send an e-mail and who had to drive around town looking for a hot spot can attest to the technology's importance. At the end of the twentieth century and the beginning of the twenty-first century, people became used to high-speed wireless networking in the *local area* (i.e., hot spots). Similarly, as the twenty-first century progresses, we will become used to high-speed mobile networking in the *wide area* (i.e., everywhere).

This book is designed as a broad examination of orthogonal frequency division multiplexing (OFDM) and orthogonal frequency division multiple access (OFDMA), which are fast becoming the de facto methods of transmission at the physical layer in broadband mobile systems. The associated functions necessary to support OFDM and OFDMA at layer 2 are also addressed. This book focuses on system analysis, design, and engineering of an OFDMA-based system, and it deals with both the theory and the application of OFDMA in the context of a broadband mobile wireless network. To the extent possible, based on the analysis of OFDMA, this book develops and presents applicable design frameworks in different areas of treatment.

In addition, the book adopts the approach of focusing on key results from the literature where appropriate so as not to detract readers from the central theme of the book. Should readers wish to pursue a particular topic further, relevant references are provided. This way, those who are interested can research those areas that are of most interest to them. Moreover, this book uses the case of the IEEE 802.16 standard to exemplify the general concepts of OFDMA. It does not attempt to encompass all details of the standard. Rather, this book covers those salient points that are important to a system-level understanding of the technology. Given that

IEEE 802.16 is a well-understood implementation of OFDMA, it serves as a solid foundation from which to investigate the relevant subject matters.

Instead of making the chapters modular so that they can be considered individually, the author structured this book so that the best result can be obtained when a reader proceeds through the chapters sequentially. This organization is due to the intertwined nature of OFDMA and IEEE 802.16 themselves, so it is recommended that the reader reads this book in its arranged sequence of chapters.

Chapter 1 provides a general introduction to OFDM and OFDMA. To understand their benefits in a broadband wireless system, one should recognize the impairments introduced by a terrestrial mobile wireless channel, topics covered by Chapter 2. Chapter 3 deals with those aspects of digital communication and networking related to an OFDMA system. After going through these background materials, this book proceeds to Chapters 4 to 6, which deal with the physical layer (layer 1). Specifically, Chapter 4 presents the theoretical details of OFDMA, Chapter 5 addresses the OFDMA implementation from the perspectives of *time* and *frequency*, and Chapter 6 addresses the same topics from the perspective of *space* (i.e., multiple antennas).

Moving from layer 1 to layer 2, Chapters 7 through 9 examine how IEEE 802.16 implements layer 2. In particular, Chapter 7 analyzes functions related to data traffic transfer, and Chapters 8 and 9 examine functions related to control and management.

Chapter 10 discusses quality of service (QoS), which is an important topic, as broadband wireless systems are increasingly called upon to transport heterogeneous traffic. In addition, network security has always been critical in wireless systems. After Chapter 11 reviews the fundamentals of network security, Chapter 12 describes the actual security functions as implemented by the standard. Finally, Chapters 13 and 14 deal with system design, emphasizing the trade-off between coverage and capacity. Specifically, Chapter 13 addresses the design of an OFDMA-based broadband wireless system from the perspective of *coverage*, and Chapter 14 addresses system design from the perspective of *capacity*.

Reference

[1] Maney, K., "Gates Speaks on Bioterror, Passport and 802.11," *USA Today*, October 31, 2001, available at http://www.usatoday.com/tech/columnist/2001/10/31/money.htm, retrieved on April 29, 2010.

Acknowledgments

The publication of this book would not be possible without the help and support of many people around me. I will do my best to acknowledge them here. First of all, I would like to thank Mark Walsh at Artech House for shepherding the book proposal process and Lindsey Gendall at Artech House for overseeing the review of the manuscript. I also want to thank the anonymous reviewer of this book for their feedback and many valuable comments; the subsequent revisions based on them have made this book a better one. In addition, I would like to thank my colleagues. I would like to thank Barry Pasternack, who has supported my professional activities for many years and has always made himself available when I needed advice. I appreciate the wise words that Bhushan Kapoor gives me, and his positive attitude is an encouragement to me as well as to others. Paul Minh's door has always been open when I needed some time to talk and share, and I will always remember his counsel. Zvi Drezner's research work is an inspiration to me and to others, and his support to me is much appreciated. Zvi Goldstein is always willing to give me his help whenever I am in a jam. I would also like to thank Anil Puri, who tirelessly works for the benefit of the college and university, and whose support over the years has made much of my research possible. Moreover, I am thankful to the special group of faculty who work with me and who in different ways challenge me to do the best I can every day.

No acknowledgment is complete without mentioning my wife Jenny. I would like to thank her for enduring much of my absence from home while I raced to finish writing this book. Her quiet strength and gentle spirit keep our family going, and for her commitment and encouragement I shall forever be grateful. This book is dedicated to her. In addition, my son Daniel has been a joy and blessing in my life. His smile and optimistic demeanor refresh my spirit every day; I very much hope that time will be frozen in its place so there will be more days spent with him. Thus, with a sincere desire to have more people know about this wireless technology, I send this book on its way.

CHAPTER 1

Introduction to OFDM and OFDMA

1.1 Motivation

People are increasingly accustomed to communicate anywhere, anytime, in any way they want, and this pattern of communication is accompanied by the ever-increasing demand for mobile broadband wireless access. As demand increases, wireless network access providers and network service providers are deploying next-generation systems that can support high-speed data. The International Telecommunication Union—Radiocommunication Sector/International Mobile Telecommunications (ITU-R/IMT)-Advanced group specifies the requirements for next-generation systems in Report ITU-R M.2134. Among the requirements specified, the cell spectral efficiency η is perhaps the most important and is defined as the ratio of the aggregate throughput of all users to the product of bandwidth and the number of cells. More specifically, it is (for either the uplink or the downlink) [1]:

$$\eta = \frac{\sum_{i=1}^{N} \chi_i}{TWM} \tag{1.1}$$

where

- χ_i is the number of correctly received bits (contained in service data units delivered to layer 3) for user i;
- N is the number of users in the system;
- W is the bandwidth;
- T is the time over which the bits are received; and
- M is the number of cells.

The first three requirements listed by Report ITU-R M.2134 are as follows:

- The cell spectral efficiencies are 2.2 bps/Hz/cell on the downlink and 1.4 bps/Hz/cell on the uplink (base coverage urban). These values assume a configuration of four transmit antennas and two receive antennas (4×2) on the

downlink and two transmit antennas and four receive antennas (2 × 4) on the uplink.

- The peak (highest theoretical) spectral efficiencies are 15 bps/Hz on the downlink and 6.75 bps/Hz on the uplink. These values assume a configuration of four transmit antennas and four receive antennas (4 × 4) on the downlink and two transmit antennas and four receive antennas (2 × 4) on the uplink.
- The support of scalable bandwidth is up to and includes 40 MHz; the consideration of wider bandwidths such as 100 MHz (achievable through frequency aggregation, for example) is encouraged.

Given a bandwidth of 40 MHz, the first two requirements shown above become:

- The aggregate throughputs are 88 Mbps/cell on the downlink and 56 Mbps/cell on the uplink (base coverage urban); and
- The peak (highest theoretical) data rates are 600 Mbps on the downlink and 270 Mbps on the uplink.

As mobile wireless systems around the world evolve to those that can support these kinds of throughputs, the underlying technology is changing from ones based on direct-sequence spread spectrum (DSSS) to ones based on orthogonal frequency division multiplexing (OFDM) and orthogonal frequency division multiple access (OFDMA).[1] Third-generation (3G) systems are based mostly on DSSS, such as Evolution-Data Optimized (EV-DO) and High-Speed Packet Access (HSPA). Most fourth-generation (4G) systems use OFDM and OFDMA, including Mobile WiMAX[2] and Long Term Evolution (LTE).

This evolution in wireless wide area networks (WWANs) is not surprising given that we have already seen a similar shift in wireless local area networks (WLANs). Earlier, lower-speed versions of the IEEE 802.11 standards, such as IEEE 802.11b, use DSSS at the physical layer. Later, higher-speed versions, such as IEEE 802.11g and IEEE 802.11n, predominantly use OFDM. One reason for such a shift is that OFDM offers some intrinsic advantages in delivering high-speed data, especially in a multipath, frequency-selective fading environment.

Given that OFDM and its variants are rapidly becoming the technology of choice in broadband wireless communications, it is important for wireless professionals to have a working knowledge of this scheme. In addition, when discussing OFDM it is important to place OFDM in the context of a specific technology implementation. Because WiMAX is the first widely deployed broadband wireless network in the United States and around the world, this book discusses OFDM in the context of WiMAX and the IEEE 802.16 family of standards. It uses the case of WiMAX to demonstrate the general concepts of OFDM. Since WiMAX is a well-understood implementation of an OFDM-based broadband wireless system, it serves as a solid foundation from which to exposit OFDM. Later OFDM-based

1. One notable exception is ultra-wideband (UWB), which is outside the scope of this book.
2. WiMAX is an acronym for Worldwide Interoperability for Microwave Access.

mobile broadband access technologies make use of fundamental concepts that were originally used in Mobile WiMAX [2]. We start with a definition of OFDM:

> OFDM is a method of multiplexing by which a high-rate data stream is divided into multiple low-rate substreams, which are then simultaneously transmitted over multiple subcarriers at the same time, and data carried by the subcarriers are sent in such a way that they do not interfere with one another in frequency.

1.2 Conventional FDM

Given the definition just presented, OFDM sounds similar to the conventional frequency division multiplexing (FDM), but OFDM differs from conventional FDM in some important respects. Figure 1.1 shows the transmitter portion of a conventional digital FDM system. At the input of the transmitter, there is a single high-rate stream of baseband data symbols running at a rate of R_s symbols per second (sps). This high-rate stream of baseband data symbols consists of blocks of complex data symbols, and each block contains L complex data symbols.

A serial-to-parallel (S-to-P) converter converts the high-rate stream into K separate low-rate substreams. As a result, each low-rate substream has a rate of R_s/K sps. Also, the serial-to-parallel converter breaks the one large block containing L symbols into K smaller blocks in parallel, each containing L/K data symbols. Each low-rate substream goes through a digital-to-analog (D-to-A) converter and is then modulated by its own complex sinusoid $\exp(-j2\pi f_k t)$, where f_k is the subcarrier frequency assigned to each low-rate substream. After modulation, the K modulated subcarriers at K different frequencies are summed, and the composite signal is then transmitted over the air.

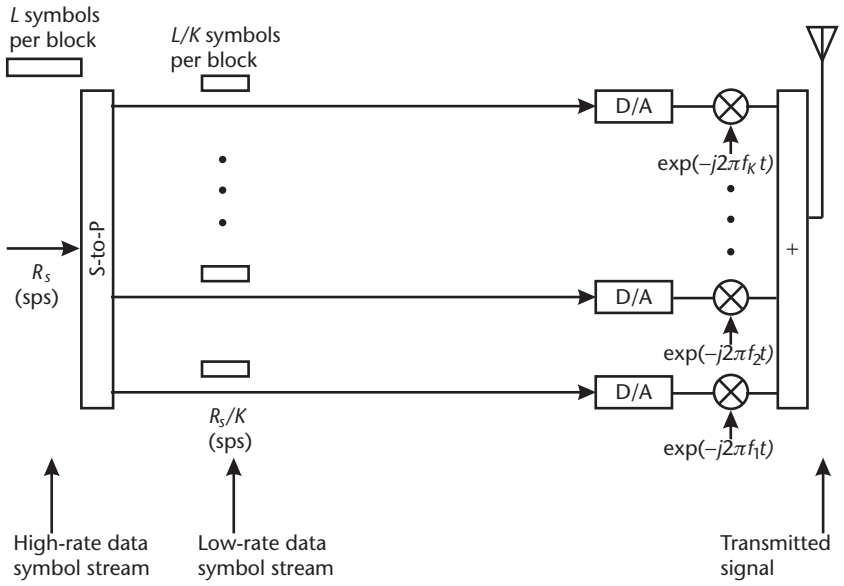


Figure 1.1 A conventional FDM transmitter.

Figure 1.2 depicts the spectrums of the high-rate stream and the K low-rate substreams. The spectrums shown are for the continuous-time equivalents of the symbol streams. While we do not go into details of computing the spectrums, it suffices for our purposes to state that the bandwidth of a stream of symbols is mostly confined to $1/T_s$, where T_s is the symbol time (duration) of a symbol in the stream. Since the symbol rate $R_s = 1/T_s$, for now we can assume that the bandwidth of a stream of symbols is limited to its symbol rate R_s [3].[3]

In this example, the high-rate stream has a symbol rate of R_s sps and a spectrum that is R_s Hz wide; each symbol lasts $1/R_s$ s. The serial-to-parallel converter divides the high-rate stream into K low-rate substreams, and each low-rate substream now has a lower rate of R_s/K sps and a narrower spectrum that is R_s/K Hz wide. Because of the lower rate, each symbol in the substream lengthens to K/R_s s. Then, each low-rate substream is modulated by a complex sinusoid $\exp(-j2\pi f_k t)$ to a different frequency f_k. After summation, the composite signal consists of K signals multiplexed in the frequency domain. As a result, K separate low-rate, narrowband subcarriers are used to transmit the original high-rate, wideband stream.

To minimize interference between subcarriers, a guard band is placed between two adjacent subcarriers. In Figure 1.2, the K subcarriers have frequencies $f_1, f_2, \ldots f_K$ that are sufficiently spaced to minimize interference between subcarriers. Because of the guard bands between subcarriers, the total bandwidth occupied by the K subcarriers is greater than simply K times the bandwidth of each subcarrier.

Although Figure 1.2 shows that there are K different subcarriers, in FDM all K subcarriers are used to carry data for one user only.

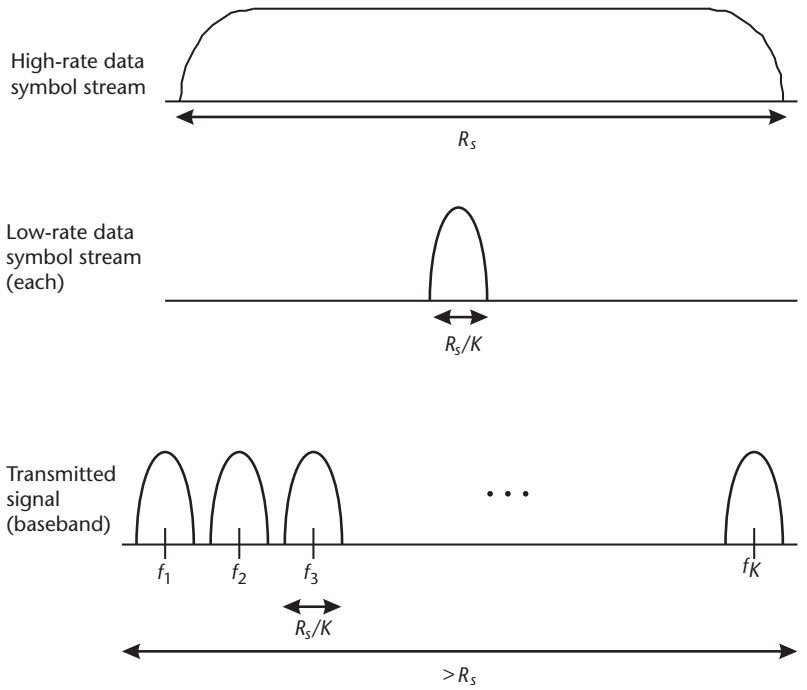


Figure 1.2 The spectrums of the high-rate data symbol stream, the low-rate data symbol substreams, and the transmitted FDM signal.

3. For an excellent treatment of power spectra of different discrete-time waveforms, consult [4].

1.3 Advantages of FDM

At this point, the reader may ask, "Why bother with dividing the high-rate stream of data symbols into K separate low-rate substreams?" From a systems perspective, transmitting a high-rate stream through K separate low-rate substreams has the following advantages (for a multicarrier system):

- It is effective at combating intersymbol interference (ISI) and multipath fading;
- It can adjust modulation and coding for each subcarrier; and
- It has simple equalization.

1.3.1 Intersymbol Interference (ISI) and Multipath Fading

Transmitting K separate, narrowband subcarriers is effective in combating ISI and multipath fading. In the *time* domain, multipath leads to the "spreading out" of the arrival time of received signals due to multiple propagation paths through which signals travel. This dispersion of arrival time is called channel delay spread τ.

In a high-speed wireless system, the symbol rate R_s is high, hence the symbol time T_s is low. As the symbol rate R_s becomes higher and the symbol time T_s becomes shorter, eventually T_s can become much shorter than the channel delay spread τ (i.e., $T_s << \tau$) for a given channel. When the symbol time becomes small as compared to the channel delay spread, the delayed versions of one symbol start to leak into and interfere with the subsequent symbol (see Figure 1.3). This phenomenon, called ISI, turns out to be a major impediment to high-speed wireless systems.

Transmitting narrowband subcarriers addresses the problem of ISI by artificially lengthening the symbol time. The symbol time is lengthened by reducing the symbol rate, and the symbol rate is reduced by dividing the high-rate symbol stream into many low-rate symbol substreams, each with a lower symbol rate R_s/K.

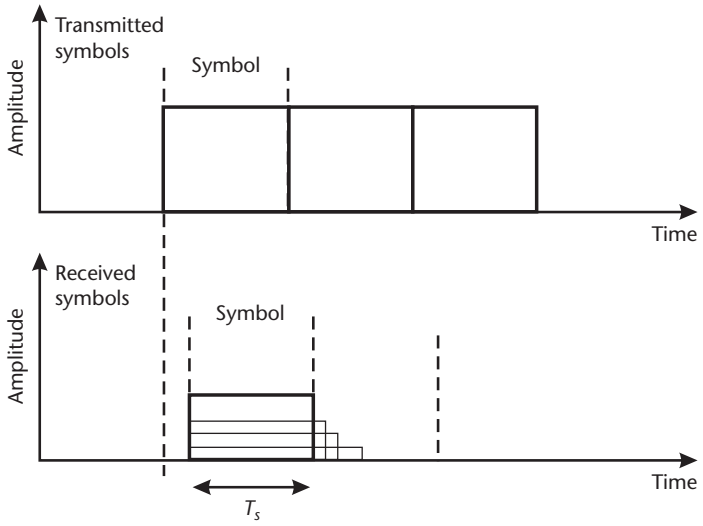


Figure 1.3 Intersymbol interference (ISI).

As a result, the symbol time of each low-rate symbol substream becomes T_sK. Thus, from the perspective of each low-rate substream of data symbols, each symbol experiences little ISI. This is so because when the symbol time of a symbol in the substream becomes large relative to the channel delay spread τ (i.e., $T_sK >> \tau$), delayed versions of a symbol have little effect on the next symbol.

The same issue can be examined in the *frequency* domain. In the frequency domain, multipath leads to "nulls" in the frequency response of the channel in frequency (see Chapter 2). Thus, multipath fading is also known as *frequency-selective* fading. A frequency-selective channel is characterized by the coherence bandwidth W_c, which is the bandwidth over which the channel appears relatively flat and unvarying.

Transmitting narrowband subcarriers addresses the problem of frequency-selective fading by artificially dividing a wideband carrier into smaller narrowband subcarriers. As a result, each narrowband subcarrier has a much smaller bandwidth (than the original wideband carrier). Each narrowband subcarrier undergoes its own fade. If the bandwidth of each narrowband subcarrier is sufficiently small (usually much smaller than the coherence bandwidth), then each narrowband subcarrier can be said to undergo flat fading. Thus, from the perspective of each narrowband subcarrier, the subcarrier experiences little frequency-selective fading. In other words, when the bandwidth R_s/K of a subcarrier becomes small relative to the channel coherence bandwidth W_c (i.e., $R_s/K << W_c$), frequency-selective fading is substantially reduced.

The robustness against ISI and multipath fading is the key advantage of using multiple narrowband subcarriers, and this advantage is carried into OFDM and OFDMA discussed later. Dealing with the effect of multipath fading experienced by wideband signals is the main reason why many high-speed technologies are adopting OFDM and OFDMA. ISI and other channel impairments are discussed in more detail in Chapter 2.

1.3.2 Modulation and Coding per Subcarrier

When the system puts data symbols across multiple subcarriers all over the band, at any given time some subcarriers experience fades but other subcarriers experience no fades. For those subcarriers that experience fades, they can fall back to a more robust modulation (e.g., quadrature phase shift keying or QPSK) and/or lower-rate error correction code (e.g., rate 1/3 convolutional code). Doing so increases the chance that data symbols will be received without errors but reduces the effective bit rate. For those subcarriers that experience little fades, they can take advantage of a more efficient modulation scheme (e.g., 16-quadrature amplitude modulation or 16-QAM) and/or higher-rate error correction code (e.g., rate 3/4 convolutional code). Doing so increases the effective bit rate without sacrificing the error rate. By adapting modulation and coding for each subcarrier, the system can achieve the best possible overall capacity and performance.

1.3.3 Simple Equalization

The receiver needs an equalization function to invert (i.e., equalize) the channel response. When a subcarrier is narrow, the required equalization function for that

subcarrier in the receiver is simpler. This is because a narrow subcarrier in frequency means a long transmission symbol in time. Note that in a receiver, a channel equalizer is still required for each subcarrier. Thus a total of K (albeit simpler) equalizers are needed in the receiver.

1.4 Disadvantages of FDM

FDM in its conventional form has two disadvantages (from the perspective of a multicarrier system). First, the transmitter needs to have K separate D-to-A converters and K separate radio frequency (RF) modulators. Second, FDM is not bandwidth efficient. The extra guard bands necessarily add to the total bandwidth requirement. So is there a way to address these disadvantages and, at the same time, retain the advantage of transmitting multiple narrowband subcarriers? The answer is yes—in the form of OFDM.

1.5 Basics of OFDM

In OFDM, the objective is still to transmit a high-rate stream using multiple subcarriers. OFDM overcomes the problem of the large bandwidth requirement imposed by guard bands. Instead of using K local oscillators (LOs) and K multipliers in modulation, OFDM uses a mathematical technique called discrete Fourier transform (DFT)[4] to generate the subcarriers. The subcarriers generated this way do not need additional guard bands and can be placed closer together in the frequency domain. The subcarriers are also orthogonal to each other over a set duration (i.e., over the duration of an OFDM symbol). In addition, DFT and its inverse can be efficiently computed, eliminating the need for separate RF components for separate subcarriers.

Figure 1.4 depicts a simplified OFDM transmitter matching the example presented in the FDM section. The high-rate stream of data symbols is still running at a rate of R_s sps, and each data symbol lasts $1/R_s$ s. This high-rate stream of data symbols consists of blocks of complex data symbols, and each block contains K complex data symbols.

Since K subcarriers are to be generated, the serial-to-parallel converter converts the high-rate stream into K separate low-rate substreams; each low-rate substream has a rate of R_s/K sps. In doing so, the serial-to-parallel converter assigns successive data symbols (at its input) to K separate substreams (at its outputs). So at any given time at the output of the serial-to-parallel converter, there is a set of K data symbols in parallel.

The set of K data symbols in parallel pass through the inverse DFT (IDFT) function, which transforms the K data symbols. After IDFT, the K transformed symbols in the K substreams then pass through the parallel-to-serial (P-to-S) converter that puts the K transformed symbols in series. This block of K transformed symbols in series constitutes a single block or an *OFDM symbol*. Successive OFDM symbols

4. In actuality, *inverse* DFT (IDFT) is used to generate the subcarriers at the transmitter, and DFT is used to recover the data symbols at the receiver.

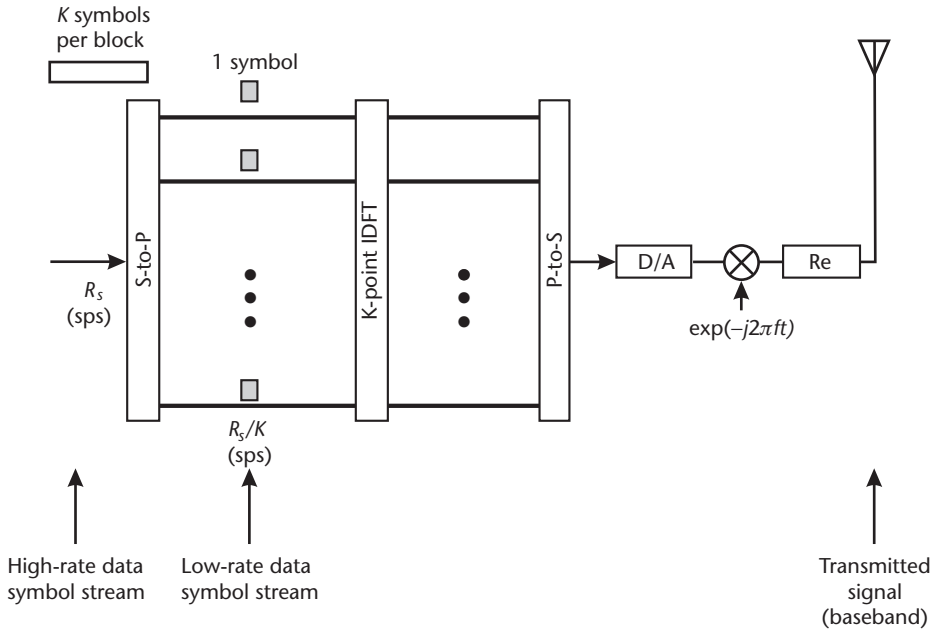


Figure 1.4 An OFDM transmitter.

at the output of the parallel-to-serial converter are running at a rate of R_s/K OFDM symbols per second, and each OFDM symbol lasts K/R_s s. Note that an OFDM symbol is different from a *data symbol*, which encodes one or more user bits and is the input to the serial-to-parallel converter.

Figure 1.5 shows the spectrums of the high-rate stream, the low-rate substreams, and the transmitted signal. In particular, the spectrum of the transmitted OFDM signal is shown over one block or one OFDM symbol. In the transmitted OFDM signal, the subcarriers are separated such that they physically overlap in frequency, but the first zero crossings of one subcarrier fall on the peaks of the two adjacent subcarriers. In fact, all zero crossings of a subcarrier fall on the peaks of all adjacent subcarriers. Because OFDM recovers the data symbol at the peak of each subcarrier, the subcarriers are orthogonal to each other and there is no interference—hence the term *orthogonal* FDM (OFDM). Chapter 4 discusses in more detail the OFDM operation that produces the overlapping subcarriers.

Although Figure 1.5 shows that there are K different subcarriers, all K subcarriers in a block (an OFDM symbol) are assigned to only one user. In other words, only one user transmits in a block (an OFDM symbol).

1.6 Advantages of OFDM

Because OFDM also transmits using multiple narrowband subcarriers, it is robust against ISI and multipath fading, can adjust modulation and coding for each subcarrier, and has simple equalizers. In addition, OFDM offers two more important advantages:

- OFDM has low-complexity modulation; and

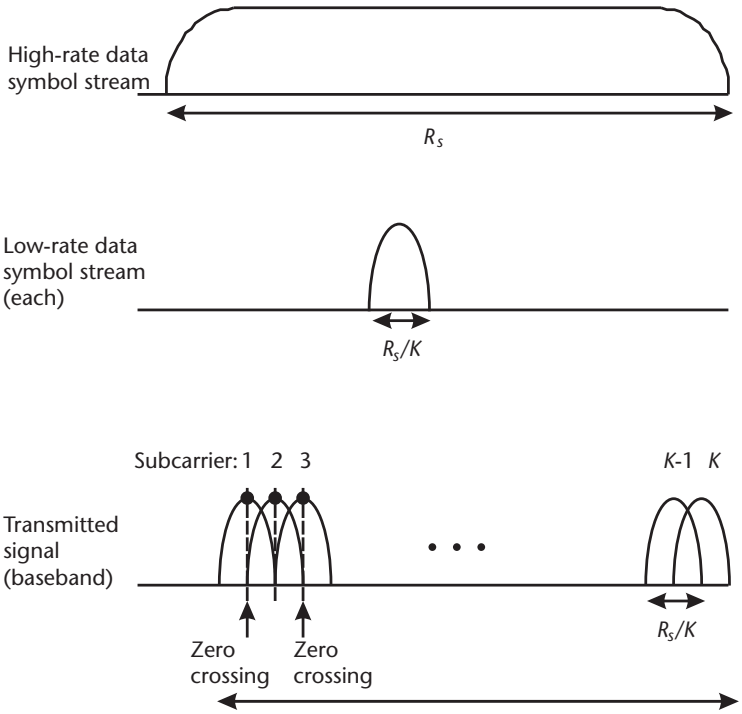


Figure 1.5 The spectrums of the high-rate data symbol stream, the low-rate data symbol substreams, and the transmitted OFDM signal. The spectrum of the transmitted OFDM signal is shown for the duration of one OFDM symbol. Note that the first zero crossings of subcarrier 2 fall on the peaks of adjacent subcarriers 1 and 3.

- OFDM achieves a better spectral efficiency than conventional FDM.

1.6.1 Low-Complexity Modulation

OFDM does not require K separate D-to-A converters and K separate RF modulators in the transmitter (and corresponding components in the receiver). Instead, OFDM modulation can be simply performed by a digital IDFT function in the transmitter and a digital DFT function in the receiver. In fact, mathematically equivalent (and more efficient) ways of computing IDFT and DFT, called inverse fast Fourier transform (IFFT) and fast Fourier transform (FFT), are used to rapidly compute the K-point transforms.

1.6.2 Spectral Efficiency

Compared with conventional FDM, OFDM achieves a better spectral efficiency. This is because whereas conventional FDM requires a guard band between adjacent carriers, OFDM do not. As an example, Figure 1.6 shows four subcarriers at passband *over the duration of one OFDM symbol*. The subcarriers are arranged in such a way that the zero crossings of one subcarrier coincide with the peaks of all other subcarriers.

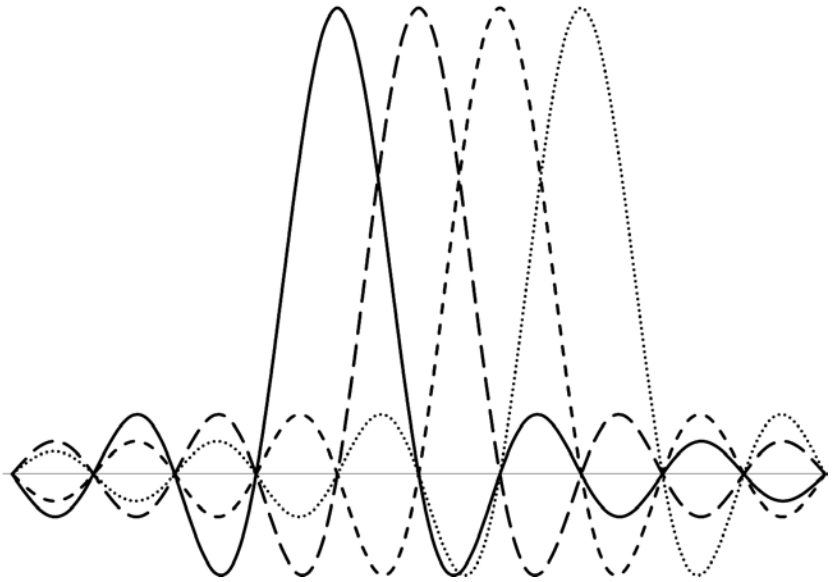


Figure 1.6 The spectrum of an OFDM signal consisting of four subcarriers. The spectrum is shown over the duration of one OFDM symbol.

After an OFDM receiver demodulates the subcarriers to baseband, it ultimately recovers the data symbol at the peak (i.e., center frequency) of a corresponding subcarrier. Because the peak of the corresponding subcarrier is where the zeros of all other subcarriers are, subcarriers do not interfere with one another over the period of one OFDM symbol. Thus in OFDM, subcarriers do overlap in frequency, but as far as the recovery of data symbols is concerned, they do not interfere with one another. Such overlapping subcarriers are what enable OFDM to have a total bandwidth less than that of conventional FDM.

1.7 Basics of OFDMA

Whereas OFDM assigns one block (in *time*) to one user, OFDMA is a method that assigns different groups of subcarriers (in *frequency*) to different users. This way, more than one user can access the air interface at the same time. Recall that in OFDM all K subcarriers are used to carry data for one user only. OFDM assigns all subcarriers to a single user at the same time, and only one user can transmit at a time. If multiple users want to transmit using OFDM, then those users have to take their turns in time. For example, in OFDM each user can be assigned one OFDM symbol in time, and OFDM symbols are assigned to their respective users before OFDM symbols enter the OFDM transmitter (Figure 1.4).

In OFDMA, instead of sequentially assigning OFDM symbols in time to different users, the system directly assigns subcarriers in frequency to different users. Figure 1.7 shows a simplified OFDMA transmitter. The high-rate stream of baseband data symbols is still running at a rate of R_s sps, and each data symbol lasts $1/R_s$ s. This high-rate stream consists of J groups of complex data symbols; each group contains L complex data symbols, and each group (of L data symbols) is

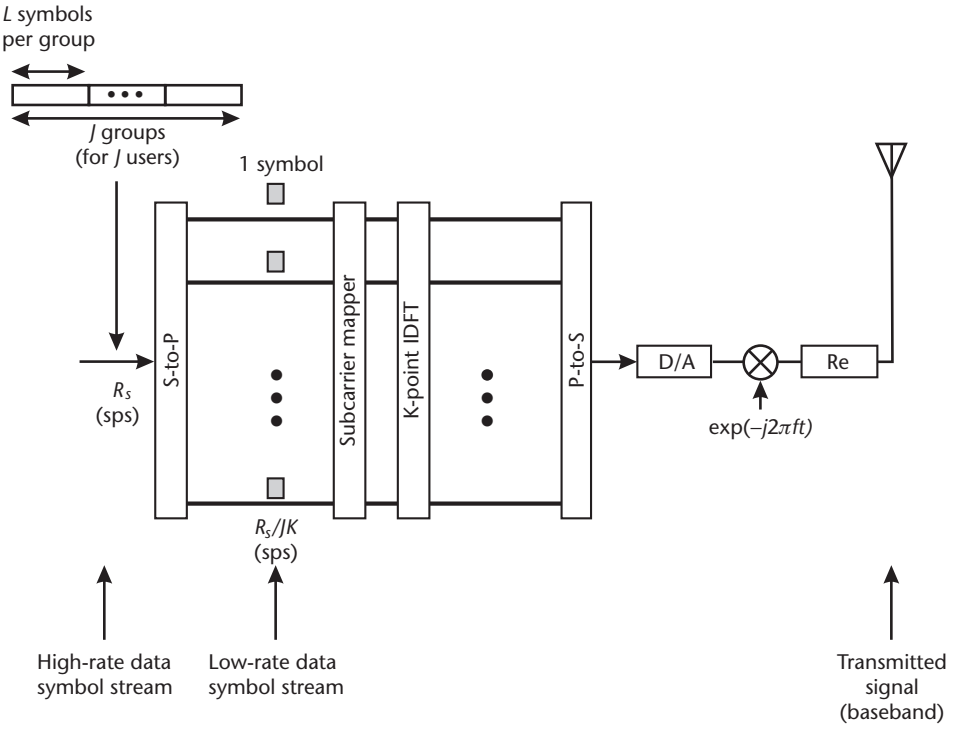


Figure 1.7 An OFDMA transmitter.

later assigned to a different user. So there are a total of JL complex data symbols in J groups.

The serial-to-parallel converter assigns the high-rate stream into JL separate low-rate substreams; each low-rate substream has a rate of R_s/JL sps. In doing so, the serial-to-parallel converter assigns successive data symbols (at its input) to JL separate low-rate substreams (at its outputs). So at any given time at the output of the serial-to-parallel converter, there is a set of JL data symbols in parallel.

The subcarrier mapper maps JL data symbols to their respective subcarriers (which are assigned to different users). Specifically, the subcarrier mapper assigns J groups of data symbols to J users in frequency. In effect, the subcarrier mapper reorders the parallel data symbols according to the particular subcarriers assigned to each user.

The set of mapped $JL(= K)$ data symbols in parallel pass through the IDFT function, which transforms the K data symbols. The K transformed symbols in K substreams then pass through the parallel-to-serial converter that puts the K transformed symbols in series. This block of K transformed symbols in series constitutes a single *OFDM symbol*. Successive OFDM symbols at the output of the parallel-to-serial converter are running at a rate of R_s/K OFDM symbols per second, and each OFDM symbol lasts K/R_s s.

Figure 1.8 shows the spectrums of the high-rate stream, the low-rate substreams, and the transmitted signal. In particular, the spectrum of the transmitted signal is shown over J groups of data symbols, or one OFDM symbol, only. The subcarriers are spaced so that they overlap in frequency but are orthogonal because

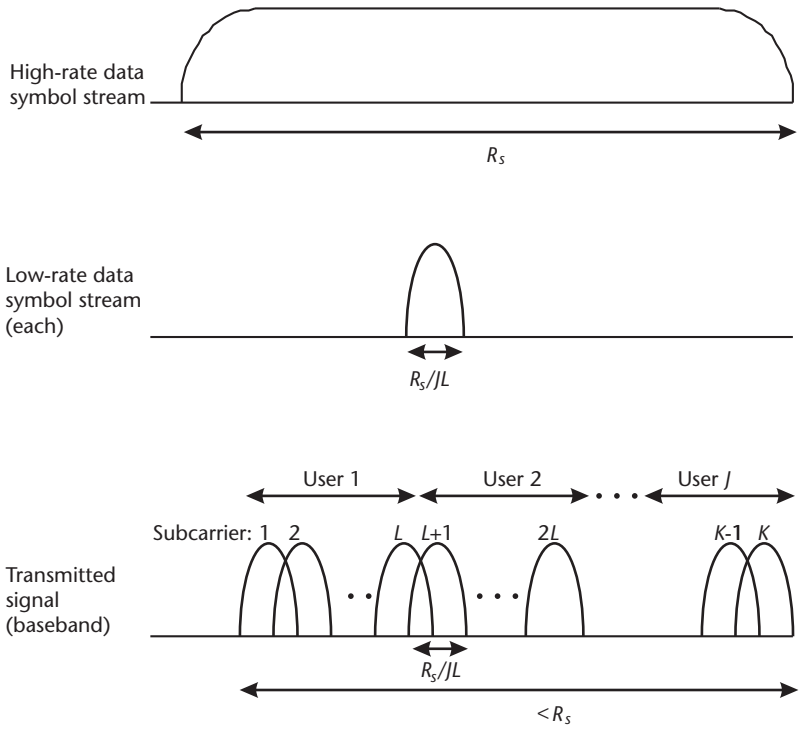


Figure 1.8 The spectrums of the high-rate data symbol stream, the low-rate data symbol substreams, and the transmitted OFDMA signal. The spectrum of the transmitted OFDMA signal is shown for the duration of an OFDM symbol (*J* groups). The figure shows that data symbols belonging to a user are carried by contiguous subcarriers.

a data symbol is recovered at the peak of a subcarrier. Note that, in this case, data symbols belonging to a user are carried by contiguous subcarriers.

In general, there are two ways to assign users' data symbols to subcarriers: distributed and contiguous. In a *distributed subcarriers* arrangement, subcarriers are assigned pseudorandomly to users. In a *contiguous subcarriers* arrangement, subcarriers are assigned to users in continuous sets (this is the scheme shown in Figure 1.8).

1.8 Advantages of OFDMA

In addition to possessing low-complexity modulation and better spectral efficiency, OFDMA affords two further advantages:

- OFDMA can take advantage of *frequency diversity* through distributed subcarriers for a single user. Distributing a user's subcarriers pseudorandomly throughout the band means some of the user's subcarriers likely would not experience fades while some of the user's other subcarriers likely would.
- OFDMA can take advantage of *multiuser diversity* through contiguous subcarriers. Multiuser diversity occurs because different users at different locations would likely experience different channel responses, thus the system can

improve a particular user's link by assigning to that user a set of contiguous subcarriers that experience the best channel condition.

Therefore, in summary, an OFDMA system has the following advantages:

- It is effective at combating ISI and multipath fading;
- It can adjust modulation and coding for each subcarrier;
- It has simple equalization;
- It has low-complexity modulation that can be implemented using IDFT/DFT (and more efficiently using IFFT/FFT);
- It has better spectral efficiency;
- It can take advantage of frequency diversity through distributed subcarriers; and
- It can take advantage of multiuser diversity through contiguous subcarriers.

1.9 Some Practical Issues of OFDM and OFDMA

The OFDM and OFDMA examples shown thus far serve to illustrate the basic principles. They show how a transmitter can send a single high-rate stream using multiple, narrowband orthogonal subcarriers. However, there are practical issues, especially in a mobile environment, that can degrade the performance of such systems. In particular, there are two important issues with implementing working OFDM and OFDMA systems: interblock interference (IBI) and intercarrier interference (ICI).

1.9.1 Time Domain: Interblock Interference

Dividing a high-rate stream into multiple low-rate substreams results in a lower R_s/K and hence a longer T_sK for each substream. Since the resulting longer T_sK is large relative to the channel delay spread τ (i.e., $T_sK >> \tau$), ISI is substantially reduced. However, although this technique reduces ISI between adjacent data symbols within an OFDM symbol, it does not reduce IBI between adjacent OFDM symbols (i.e., adjacent blocks).

To reduce interference between successive OFDM symbols transmitted through the channel, an extra "guard time" is provisioned at the end of each OFDM symbol to further prevent delayed versions of itself from interfering with the next OFDM symbol. This guard time is referred to as the *cyclic prefix* and is discussed in more detail in Chapter 4. In short, if the guard time is larger than the delay spread τ, then successive OFDM symbols would not interfere with each other, and IBI is reduced.

1.9.2 Frequency Domain: Intercarrier Interference

The second issue has to do with frequency synchronization. This problem comes about when there is a difference between the carrier frequency at the transmitter

and the carrier frequency at the receiver. Looking at Figure 1.6, readers can easily see that frequency synchronization is very important in OFDM. If the center frequency of a subcarrier is shifted just a little bit, then that subcarrier is no longer orthogonal to its neighboring subcarriers, and intercarrier interference (ICI) will result. The issue of frequency synchronization in OFDM is discussed in more detail in Chapter 4.

1.10 OFDM and DSSS

In a traditional DSSS system, the (interference-rejection) performance and capacity depend on the *processing gain* G_p, which is defined as

$$G_p = \frac{W}{R_b} \tag{1.2}$$

where W is the final spread bandwidth and R_b is the baseband bit rate (see Figure 1.9) [3]. As we know, users continue to demand broadband mobile applications that require a higher and higher bit rate R_b. If R_b increases, W necessarily has to increase as well to maintain the same G_p. If R_b increases but W is fixed (e.g., due to a specific spectrum allocation), then G_p decreases. In addition, the number of code channels that a traditional DSSS system can support depends on G_p [3, 5], so G_p relates to the capacity of the system also.

OFDM is a technique that belongs to a larger class of modulation schemes called *multicarrier modulation*. In multicarrier modulation, one high-rate stream is divided into many low-rate substreams, which are then transmitted over parallel frequency subcarriers. Initial works on multicarrier modulation were done in the 1960s in the United States, and in the 1970s it was demonstrated that DFT can be used to implement an FDM system [6]. What makes OFDM unique is that in OFDM the subcarriers are arranged in such a way that they are orthogonal to each other.

1.11 Overview of the Book

This chapter serves as a brief introduction to OFDM and OFDMA and their advantages. However, the simplified transmitter structures shown in this chapter do not take into account some practical implementation issues. To help understand these

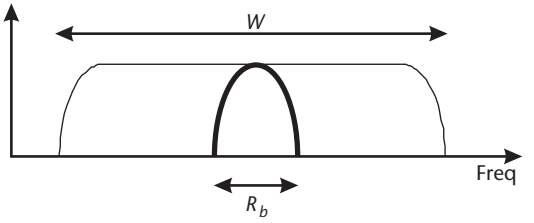


Figure 1.9 Processing gain G_p in a traditional DSSS system.

issues, Chapter 2 characterizes the types of impairments that a terrestrial mobile wireless channel imparts, and Chapter 3 covers the fundamentals of digital communication and networking relevant to the understanding of an OFDM system.

After discussing these backgrounds, we then consider more implementation details of OFDM and OFDMA. Chapters 4–6 cover the physical layer (layer 1). Chapter 4 goes through those aspects of OFDM that are conducive to transmitting and receiving high-rate streams over a wireless channel, especially one that is time dispersive. After going through the general remedy that OFDM offers, the book proceeds to Chapter 5, which covers the specific format of the physical layer as implemented in IEEE 802.16. Specifically, Chapter 5 addresses the physical layer from the perspective of channel structure (i.e., time and frequency), whereas Chapter 6 addresses the physical layer from the perspective of multiple antennas (i.e., space).

Broadband wireless networks cannot only rely on the increase in raw bit rate provided by the enhanced physical layer. To be sure, any increase in throughput as perceived by a higher-layer application also depends on how effectively the medium access control layer (layer 2) utilizes the physical layer. Thus, any increase in throughput relies on enhancements made in layer 2. Chapters 7–9 use IEEE 802.16 as a case and examine how layer 2 is implemented by this standard. Specifically, Chapter 7 discusses layer 2 with respect to the data traffic it transports, whereas Chapters 8 and 9 discuss layer 2 from the perspective of its control and supervisory functions.

Chapters 10–12 discuss two special topics that are particularly relevant to operating modern and future broadband wireless networks: quality of service (QoS) and security. Chapter 10 presents QoS. The use of QoS has enabled different networks (both wireless and wireline) to deliver heterogeneous traffic based on the needs of individual users. Given that broadband wireless networks are inherently capacity-constrained, the application of QoS is an important topic to address. In addition, a broadband wireless network carries with it unique security issues. Chapter 11 first reviews the fundamentals of network security, then Chapter 12 shows how important security features work in an actual network (IEEE 802.16). The book concludes with Chapters 13 and 14, which cover topics in the design of an OFDMA-based broadband wireless system. Since a broadband wireless system effectively trades off *coverage* against *capacity*, the design aspects are presented in a coverage part (Chapter 13) and a capacity part (Chapter 14).

References

[1] ITU-R M.2134, "Report ITU-R M.2134: Requirements Related to Technical Performance for IMT-Advanced Radio Interface(s)," ITU, Nov. 2008.

[2] Ahmadi, S., "An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Communications*, Vol. 47, No. 6, 2009, pp. 84–98.

[3] Yang, S.C., *CDMA RF System Engineering*, Norwood, MA: Artech, 1998.

[4] Carlson, B.A., *Communication Systems*, New York: McGraw-Hill, 1986.

[5] Yang, S.C., *3G CDMA2000 Wireless System Engineering*, Norwood, MA: Artech, 2004.

[6] Weinstein, S.B., Ebert, P.M., "Data Transmission by Frequency-Division Multiplexing Using the Discrete Fourier Transform," *IEEE Transactions on Communication Technology*, Vol. 19, No. 5, 1971, pp. 628–634.

Selected Bibliography

Prasad, R., *OFDM for Wireless Communication Systems*, Boston: Artech, 2004.

CHAPTER 2

Characterization of the Mobile Wireless Channel

2.1 Introduction

The characterization of the mobile wireless channel is important in understanding why OFDM has become a popular technology of choice in broadband mobile systems.[1] Here the characterization of the channel takes the form of models, which are *analytical* or *empirical* conceptualizations of a real-world phenomenon. Before going into the various impairments introduced by the mobile wireless channel, we need to emphasize that some channel impairments can be corrected by increasing the transmit power (e.g., path loss and shadowing loss), while other impairments cannot be compensated by merely increasing the transmit power (e.g., delay spread and Doppler spread). In a broadband mobile system, it is often the latter impairments that present the greatest challenge to system designers. In fact, the evolution of wireless communication can be likened to humankind's effort to combat the effects of channel impairments in the quest for higher bit rates.

To organize the discussion of different channel impairments, this chapter categorizes the impairments into three categories:

- *Distance:* Path loss, shadowing loss, and multipath fading (Section 2.3);
- *Time:* Multipath delay spread and time dispersion (Section 2.4);
- *Frequency:* Doppler spread and frequency dispersion (Section 2.5).

We first start with an overview of the link equation in Section 2.2 and then proceed to discuss the different categories of impairments.

2.2 Link Analysis

In any communication system, we are concerned with an important parameter called the signal-to-noise ratio (SNR), or *S/N*, at the receiver. The parameter defines

1. Some parts of this chapter are adopted from Chapter 2 of [1] with substantial enhancements and revisions tailored to those issues relevant to a high-speed OFDM system.

how much received signal power there is as compared to the noise power at the receiver; therefore, the SNR can be considered a figure of merit for the communication system.

The classical *link equation* is a formula that calculates the SNR using several other parameters of the communication system.

$$\frac{S}{N} = \frac{(S_T L_c G_T) L_p G_R}{N} \tag{2.1}$$

where:

- S_T is the signal power at the output of the transmitter power amplifier;
- L_c is the cable loss between the power amplifier and the transmit antenna;
- G_T is the gain of the transmit antenna;
- L_p is the propagation loss introduced by the channel;
- G_R is the gain of the receive antenna;
- N is the thermal noise power.

The product $(S_T L_c G_T)$ is also known as the *effective isotropic radiated power* (EIRP) from the transmit antenna. The thermal noise power N is defined as

$$N = kTW \tag{2.2}$$

where k is the Boltzman's constant (1.38×10^{-23} W/Hz-K or –228.6 dBW/Hz-K), T is the noise temperature of the receiver, and W is the bandwidth of the system.

A parameter that is more comprehensive than the SNR is the signal-to-interference plus noise ratio (SINR), or $S/(I+N)$. The SINR differs from the SNR in that the denominator of SINR includes not only thermal noise power, but also interference powers from other sources, such as adjacent (sub)carriers in frequency or nearby cells in space. SINR is a more descriptive figure-of-merit because it takes into account other interference effects, which are common in a terrestrial broadband wireless system.

Another parameter that is often used in terrestrial systems is the signal-to-interference ratio (SIR) or S/I. The SIR is sometimes used to approximate SINR, especially in an interference-limited system where other interference effects (e.g., interference from other cells) are more prominent than thermal noise.

As one can see from (2.1), the link quality is dependent on parameters such as gains of transmit and receive antennas, transmitter power, and receiver noise temperature. These parameters are largely within the control of the system designer and can be changed to optimize system performance. One parameter, however, in (2.1) is not entirely within the control of the system designer. This parameter is the propagation loss, which is discussed in the next section.

2.3 Distance Dimension: Propagation Loss

The propagation loss refers to the attenuation that the signal suffers en route from the transmitter to the receiver. In the context of (2.1), the propagation loss includes three components: path loss, shadowing loss, and multipath fading. These are the types of attenuations that can be compensated by increasing the transmit power. All these losses are present in a terrestrial wireless system where the mobile is either stationary (i.e., fixed at one location) or moving (i.e., traveling at some speed).

2.3.1 Path Loss

Path loss is the mean power loss that the signal experiences between the transmitter and the receiver. There are many models used to describe and predict path loss, and they can be *analytical* or *empirical* in nature. Although the models differ in their methodologies, all have distance between the transmitter and the receiver as a critical parameter. Other effects may also come into play in addition to distance. For example, a model may have a parameter that describes the clutter of the channel being modeled, with a value for "dense urban" that produces a higher loss and another value for "rural" that produces a lower loss. Several models for calculating path loss are examined next.

2.3.1.1 Free-Space Model

This model is an analytical model that describes the power loss in free space. In free space, electromagnetic waves diminish as a function of inverse square, or $1/d^2$, where d is the distance between the transmitter and the receiver. Physically, as the signal radiates away from a point source in space, the spherical wavefront expands and diminishes in intensity.

In its linear form, the free-space path loss is

$$L_p = \frac{\lambda^2}{(4\pi d)^2} \text{ or } L_p = \frac{c^2}{(4\pi)^2 f^2 d^2} \tag{2.3}$$

where d is the distance (in meters), λ is the wavelength of the signal (in meters), f is the frequency of the signal (in hertz), and c is the speed of light (in meters/second); the speed of light is a product of frequency and wavelength (i.e., $c = \lambda f$). Equation (2.3) can also be rewritten in a decibel (dB) form:

$$L_p = 147.56 - 20\log(f) - 20\log(d) \tag{2.4}$$

where path loss L_p is in decibels. Note that once the carrier frequency f of the signal is known, the second term of (2.4) is effectively a constant, and L_p varies strictly as a function of d in the third term. If we plot (2.4) on a log-log graph, then the *slope* of the curve would be –20 dB/decade.

The free-space model is mostly used in satellite and deep-space communication systems where the signal truly travels through free space. In a mobile communication system where additional losses are introduced by ground reflections, terrestrial

obstacles, and other impairments, alternative models are needed to accurately predict path loss.

2.3.1.2 Plane-Earth Model

In a terrestrial environment, the path loss experienced is worse than that in free space. The most significant difference between terrestrial and free space is the presence of ground (and ground reflections) in a terrestrial environment. In addition, there are often obstacles between the base station and the mobile.[2] As a result, the received signal can be made up of signals traveling via direct and indirect paths. Signals traveling via direct paths are those in line of sight (LOS), whereas signals traveling via indirect paths are those in nonline of sight (NLOS) and involve refraction and reflection from objects (e.g., buildings, trees, and hills) between the transmitter and the receiver. Therefore, the path loss in the terrestrial environment is higher than that in free space, and the extent of the loss is even more strongly influenced by the distance between the transmitter and the receiver.

In the terrestrial environment, an analytical model known as the plane-Earth model is

$$L_p = \frac{h_t^2 h_r^2}{d^4} \tag{2.5}$$

where d is the distance (in meters) between the transmitter and the receiver, h_t is the height (in meters) of the transmit antenna, and h_r is the height (in meters) of the receive antenna. The model analytically calculates path loss by taking into account the phase difference between two paths, the direct path and the ground-reflected path, and by assuming that $d >> h_t$ and $d >> h_r$ (see Figure 2.1).

This model is known as the plane-Earth model because it considers reflection from the ground that is flat, not curved. Note that in this case, the path loss varies as an inverse power of 4 in contrast to an inverse power of 2 in free space. In other

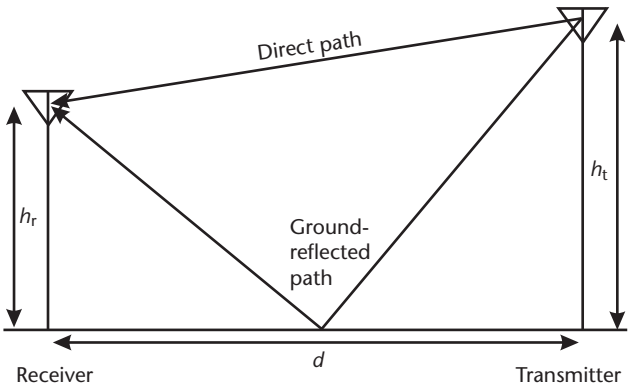


Figure 2.1 There are two paths, the direct path and the ground-reflected path, in the plane-Earth model.

2. This book uses the term "mobile" as a mobile user device that communicates with the base station. Other similar terms in the literature include the mobile station (MS) and the subscriber station (SS).

words, the path loss encountered in the terrestrial environment is worse than that seen in free space. In addition, the modeled loss becomes less as the antenna heights increase.

In practice, a correction factor a that depends on the frequency of the carrier is added to yield

$$L_p = a\frac{h_t^2 h_r^2}{d^4} \tag{2.6}$$

Converting (2.6) into decibel form produces

$$L_p = 10\log(a) + 20\log(h_t) + 20\log(h_r) - 40\log(d) \tag{2.7}$$

where the path loss L_p is in decibels. Note that in (2.7) the path loss slope is –40 dB/decade.

2.3.1.3 Simple Empirical Model

Although the two analytical models presented thus far have strong theoretical underpinnings (one based on expanding wavefronts in free space and one based on two paths on plane Earth), in terrestrial environments the power loss rarely varies by exactly $1/d^2$ or $1/d^4$. Path loss models often have to be augmented by actual measurements of the propagation environment. A simple empirical model of path loss that can be tuned by measurements is:

$$L_p = \left(\frac{\lambda}{4\pi d_{\text{ref}}}\right)^2 \left(\frac{d_{\text{ref}}}{d}\right)^{\alpha} \text{ or } L_p = \left(\frac{c}{4\pi f d_{\text{ref}}}\right)^2 \left(\frac{d_{\text{ref}}}{d}\right)^{\alpha} \tag{2.8}$$

where d is the distance between the transmitter and the receiver (in meters), and d_{ref} is a reference distance, commonly taken to be some fixed distance from the transmitter. The path loss exponent α depends on the actual propagation environment, often ranging from 3 in rural areas (i.e., low loss) to 5 in dense urban areas (i.e., high loss).

One can recognize that the first factor of (2.8) is the free-space model, so an interpretation of (2.8) is that it is the free-space loss (at d_{ref}) corrected by $1/d^{\alpha}$ (at d). The similarity can also be seen when (2.8) is rewritten in the decibel form:

$$L_p = 147.56 - 20\log(f) - 20\log(d_{\text{ref}}) + 10\alpha\log(d_{\text{ref}}) - 10\alpha\log(d) \tag{2.9}$$

and

$$L_p = 147.56 - 20\log(f) - 10\alpha\log(d) \tag{2.10}$$

for $d_{\text{ref}} = 1$m. The slope of this equation is -10α dB/decade. Note that this equation differs from the decibel form of the free-space model only in the coefficient of the last term.

2.3.1.4 Erceg Model

The Erceg model is an empirical model that is more complex. A useful path loss model should be a function of different parameters that describe the different characteristics of the propagation environment. The Erceg model illustrates a more complicated path loss model that depends on parameters such as frequency, height of the base station, height of the mobile, and terrestrial conditions. The model is quite powerful and uses these different parameters to model path loss. It is based on propagation measurements taken at 1.9 GHz [2], but has since been modified for higher frequencies. The median path loss is given by (in decibels):

$$L_p = 147.56 - 20\log(f) - 20\log(d_{\text{ref}}) + 10\alpha\log(d_{\text{ref}}) - 10\alpha\log(d) + C_f + C_h \quad (2.11)$$

where f is the frequency (in hertz), d is the distance between the base station and the mobile (in meters), α is the path loss exponent, d_{ref} is the reference distance (= 100m), C_f is the frequency correction factor, and C_h is the correction factor for the height of the mobile. One can recognize that the first five terms of (2.11) constitute the simple empirical model, so an interpretation of (2.11) is that it is the simple empirical model (at d) further corrected by the correction factors C_f and C_h.

The path loss exponent is:

$$\alpha = a - bh_b + \frac{c}{h_b} \quad (2.12)$$

where h_b is the antenna height of the base station in meters. h_b needs to be between 10m and 80m.

The parameters in the path loss exponent expression depend on three types of terrain ranging from high loss to low loss: terrain A (hilly/moderate to heavy density of trees), terrain B (hilly/light density of trees or flat/moderate to heavy density of trees), and terrain C (flat/light density of trees). The parameters are:

- For terrain A: $a = 4.6$, $b = 0.0075$, and $c = 12.6$;
- For terrain B: $a = 4$, $b = 0.0065$, and $c = 17.1$;
- For terrain C: $a = 3.6$, $b = 0.005$, and $c = 20$.

The frequency correction factor C_f is

$$C_f = -6\log\left(\frac{f}{2\times 10^9}\right) \quad (2.13)$$

Thus, the higher the frequency, the higher the loss.

The correction factor for the height of the mobile C_h is

$$C_h = 10.8\log\left(\frac{h_m}{2}\right) \text{ for terrain A and terrain B} \quad (2.14)$$

$$C_h = 20\log\left(\frac{h_m}{2}\right) \text{ for terrain C} \tag{2.15}$$

where h_m is the antenna height of the mobile in meters. h_m needs to be between 2m and 10m. The higher the antenna height of the mobile, the lower the loss.

2.3.1.5 Okumura-Hata Model

The Okumura-Hata model is another empirically based path loss model. It was based on extensive empirical measurements taken in urban environments by Okumura and others in 1968 [3]. Then in 1980, Hata simplified the work of Okumura and provided an expression for path loss that later became known as the Okumura-Hata model [4]. The path loss expression is:

$$\begin{aligned} L_p = &-69.55 - 26.16\log(f) + 13.82\log(h_b) + \\ &a(h_m) - \left[44.9 - 6.55\log(h_b)\right]\log(d) \end{aligned} \tag{2.16}$$

where f is the carrier frequency (in megahertz), h_b is the antenna height (in meters) of the base station, h_m is the antenna height (in meters) of the mobile, and d is the distance (in kilometers) between the base station and the mobile user. The terms $a(h_m)$ is a correction factor that depends on the height of the mobile antenna. The model is valid for frequencies from 150 MHz to 1.5 GHz, and for many years, it has been a widely used model for predicting path loss in wireless systems, especially at the cellular frequencies.

2.3.1.6 COST-231 Hata Model

Although the Okumura-Hata model has been used in wireless communications for a long time, the model is only valid up to 1.5 GHz. The European Cooperation in the field of Scientific and Research (COST) extended the Okumura-Hata model to the COST-231 Hata model [5], which is an empirical model that is valid up to 2 GHz. The path loss predicted by the COST-231 Hata model is:

$$\begin{aligned} L_p = &-46.3 - 33.9\log(f) + 13.82\log(h_b) \\ &+a(h_m) - \left[44.9 - 6.55\log(h_b)\right]\log(d) - K_0 \end{aligned} \tag{2.17}$$

where f is the carrier frequency (in megahertz), h_b is the antenna height (in meters) of the base station, h_m is the mobile antenna height (in meters), and d is the distance (in kilometers) between the base station and the mobile user. For these parameters, there are only certain ranges in which the model is valid (i.e., h_b should only be between 30m to 200m, h_m should be between 1m to 10m, and d should be between 1 km to 20 km). Note that the slope of (2.17) is $-[44.9 - 6.55\log(h_b)]$ dB/decade.

The terms $a(h_m)$ and K_0 are used to account for whether the propagation takes place in an "urban" or a "dense urban" environment. In particular,

$$a(h_m) = \left[1.1\log(f) - 0.7\right]h_m - \left[1.56\log(f) - 0.8\right] \text{ for "urban"} \tag{2.18}$$

$$a(h_m) = 3.2\left[\log(11.75h_m)\right]^2 - 4.97 \text{ for "dense urban"} \tag{2.19}$$

and

- $K_0 = 0$ dB for urban;
- $K_0 = 3$ dB for dense urban.

The COST-231 Hata model is a popular model for predicting path loss. In fact, the WiMAX Forum makes use of the COST-231 Hata model in its published documents [6, 7].

2.3.1.7 Observations

The path loss models presented thus far can be written in a general form of straight-line equation (in decibels):

$$L_p = -L_0 - \gamma\log(d) \tag{2.20}$$

where L_0 is the *intercept* and γ is the *slope*. The slope is a factor showing how severe the loss becomes as a function of distance. For illustration purposes, Figure 2.2 shows a comparison among three path loss models: free-space, simple empirical, and COST-231 Hata. Note that the slopes for these models are, respectively, –20 dB/decade, –45 dB/decade, and –35.2 dB/decade—all for a base station height of 30m.

These models have their limitations when used to predict path loss in terrestrial environments. The accuracy of these models typically varies between 6 dB to 8 dB when compared to field measurements. The accuracy can be increased, however, by integrating field measurement results with the model. For example, it is a common industry practice to take field measurements and custom-calculate the model *slope* that is used over certain distances from the base station. See [8] for a comparison of empirical models of path loss.

Another limitation is that these models are not as well suited in microcell regions. The microcell regions refer to those distances that are very close to the base station. Other propagation phenomena dominate when one attempts to predict path loss very close to the base station; hence, specialized microcell models are needed to predict losses in these regions. A popular model for predicting path loss in microcell regions is the Walfisch-Ikegami model [9]. See also [10] for a good discussion of microcell path loss models.

2.3.2 Shadowing Loss

As discussed previously, the median path loss increases when the distance between the transmitter and the receiver increases. However, at different locations from the transmitter, different obstacles (such as trees, buildings, and moving trucks) would

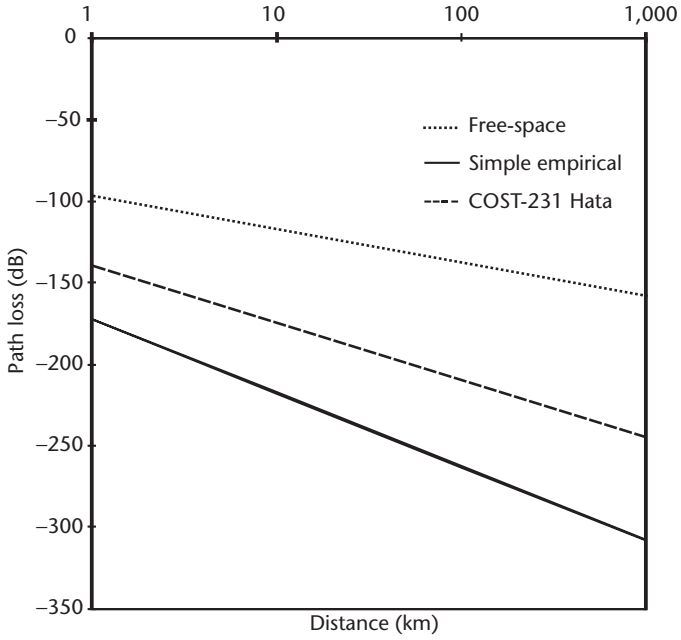


Figure 2.2 As an illustration, the graph shows the path loss versus distance for three different path loss models: free-space, simple empirical, and COST-231 Hata. The antenna height and carrier frequency are 30m and 1.8 GHz, respectively. α is taken to be 4.5 for the simple empirical model. For the COST-231 Hata model, a mobile antenna height of 1.5m and a "dense urban" scenario are used.

move in the way and out of the way of the signal path, and such distribution of obstacles causes occasional extra rises and falls in power loss. This variation in power loss occurs over relatively large distances (on the order of tens of meters) and is thus called *slow fading*.

Slow fading is usually modeled by a *log-normal* distribution with a mean and a standard deviation expressed in decibels. In log-normal distribution, the variation of path loss is distributed as $10^{\xi/10}$, where ξ is a normal (Gaussian) random variable. The standard deviation in a cellular environment can vary between 5 dB and 12 dB. Some references recommend using 10 dB for modeling shadowing loss in macrocell environments [11].

The log-normal distribution has the following probability density function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left[\frac{(x-m)^2}{2\sigma^2}\right]} \tag{2.21}$$

where x is the random variable describing the path loss (in decibels), m is the mean path loss (in decibels), and σ is the standard deviation of the path loss (in decibels).

We know that the median path loss increases as the distance increases between the transmitter and the receiver. One way to visualize slow fading is to picture that there is a slow loss variation (occurring over relatively large distances) on top of the median path loss, and that variation can be described by a log-normal probability distribution (see Figure 2.3).

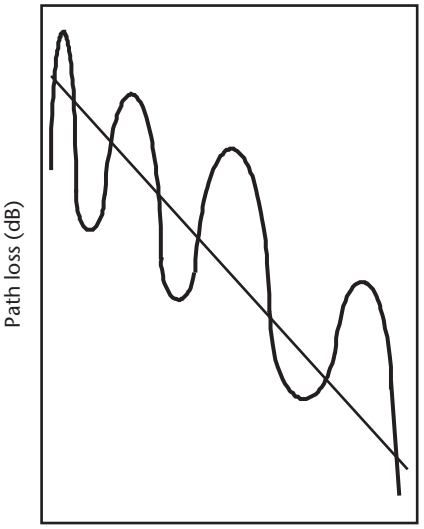


Figure 2.3 As an illustration, the graph shows shadowing loss superimposed on top of the path loss; both are functions of the distance between the transmitter and the receiver. The distribution of variation in shadowing loss is log-normal.

The reason for the log-normally distributed slow fading is that the received signal is the result of the transmitted signal passing through or reflecting off many different objects, such as trees and buildings. Each object attenuates the signal to some extent, and the final received signal power is the product of transmission or reflection factors of all these objects. As a consequence, the *logarithm* of the received signal equates to the sum of a large number of logarithms of transmission/reflection factors, each of which can then be expressed in decibels. As the number of transmission/reflection factors becomes large, the central limit theorem states that the distribution of the sum approaches Gaussian, even if the individual terms are not Gaussian [12].

2.3.3 Multipath Fading

There are times when a receiver is completely out of sight of the transmitter (i.e., there is no signal path traveling to the receiver via LOS). In this case, the received signals are made up of a group of reflections from objects, and none of the reflected paths is any more dominant than the other ones. The different reflected signal paths arrive at slightly different times, with different amplitudes, and with different phases. These paths superimpose constructively and destructively to cause the received power to go up and down. These variations in loss occur over very small distances (on the order of a wavelength), so it is called *fast fading*.

It was verified, both theoretically and experimentally, that the envelope of a received carrier signal over small distances is Rayleigh distributed [13]. Therefore, this type of fading is also called *Rayleigh fading*.[3] The theoretical model makes the

3. Strictly speaking, the *Rayleigh* distribution describes the variations in received power when there is no LOS between the transmitter and the receiver; the *Rician* distribution describes the variations in received power when there is LOS.

use of the fact that there are many (N) reflected signal paths from different directions with different phases. The received signal envelope **s** in complex form is:

$$\mathbf{s} = \sum_{n=1}^{N} s_n e^{j\theta_n} \tag{2.22}$$

where s_n is the amplitude of the nth signal path, θ_n is the phase of the nth signal path uniformly distributed between 0 and 2π, and **s** is a composite of N signal paths. If one assumes that individual N signal paths are independent and identically-distributed (i.i.d.) random variables, then the central limit theorem can be invoked. The central limit theorem states that if N is large, the sum of i.i.d. random variables becomes a zero-mean Gaussian random variable. Thus, if N is large, **s** becomes a zero-mean complex Gaussian random variable.

Because **s** is complex, it can be broken down into a real part (s_r) and an imaginary part (s_i):

$$\mathbf{s} = s_r + js_i \tag{2.23}$$

where it can be shown that s_r and s_i are real Gaussian random variables. The amplitude of **s** then is:

$$s = \sqrt{s_r^2 + s_i^2} \tag{2.24}$$

which is a random variable that has a Rayleigh distribution. The Raleigh distribution has the following probability density function:

$$p(s) = \frac{s}{\sigma} e^{\frac{-s^2}{2\sigma^2}} \quad \text{for} \quad s \geq 0 \tag{2.25}$$

and $p(s) = 0$ for $s < 0$. One way to visualize this type of fading is to picture a transmitter sending an unmodulated carrier with a constant envelope. As the receiver measures the received envelope at different locations, the measurements (at different locations) would be different and be distributed according to a Rayleigh distribution.

This type of fading occurs in addition to median path loss and shadowing loss. Figure 2.4 shows that there is an additional loss variation (occurring over very small distances) on top of both median path loss and shadowing loss, and that variation can be described by a Rayleigh probability distribution.

Constructive and destructive interferences result because there are many different signal paths. Thus, another way to visualize this particular fading phenomenon is to picture electromagnetic fields radiated by a transmitter combining constructively and destructively, forming a standing-wave pattern in the surrounding area. As a receiver changes place to different locations in the pattern, it would experience ups and downs in amplitudes; a severe drop in received power is sometimes called a *deep fade* (see Figure 2.5). The distance and spacing between each fade are depen-

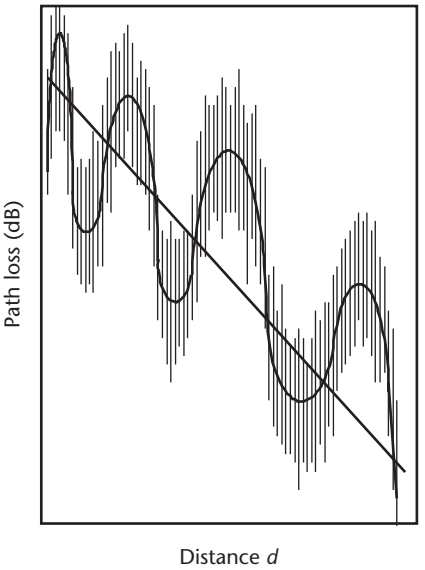


Figure 2.4 As an illustration, the graph shows multipath fading superimposed on top of both path loss and shadowing loss. The distribution of variation in multipath fading is Rayleigh.

dent on the carrier frequency. In a wireless environment, the amplitude variation due to this fading phenomenon can be as much as –20 dB.

Example 2.1

Compare the distance between fades between a carrier frequency of 2.5 GHz and a carrier frequency of 700 MHz. Use Figure 2.5 as the propagation environment.

The wavelengths at 2.5 GHz and 700 MHz are:

$$\lambda_{2.5\text{ GHz}} = \frac{c}{f_{2.5\text{ GHz}}} = \frac{3\times10^8}{2.5\times10^9} = 0.12\text{m}$$

$$\lambda_{700\text{ MHz}} = \frac{c}{f_{700\text{ MHz}}} = \frac{3\times10^8}{700\times10^6} = 0.43\text{m}$$

The half wavelengths at 2.5 GHz and 700 MHz are then:

$$\frac{\lambda_{2.5\text{ GHz}}}{2} = \frac{0.12\text{m}}{2} = 0.06\text{m}$$

$$\frac{\lambda_{700\text{ MHz}}}{2} = \frac{0.43\text{m}}{2} = 0.22\text{m}$$

Thus, a lower-frequency carrier has fades that are farther apart from each other than a higher-frequency carrier.

It is important to note that the term "fast fading" does not necessarily mean that the receiver (or the transmitter) is moving. It simply means that the distance

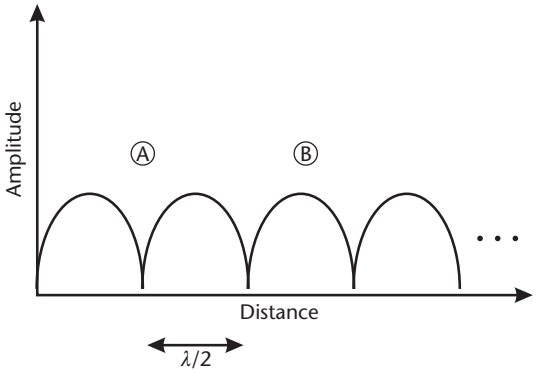


Figure 2.5 In the standing-wave pattern around the transmitter, a receiver at location A suffers a deep fade as compared to another receiver at location B. In this case, fades occur once every half wavelength. Note that the standing-wave pattern shown is a simple example resulting from the addition of two equally strong waves that are 180° out of phase.

between successive fades is small due to constructive and destructive interferences of multiple signal paths. If one takes a number of loss measurements at a number of stationary positions (separated by small distances) around a transmitter, then those measurements are Rayleigh distributed (provided the measurements are taken at locations with no LOS to the transmitter). Of course, if the receiver or the transmitter is moving, then the time between successive fades will be short, and successive fades occur very quickly.

2.3.4 Concluding Remarks

The three types of impairments discussed so far—path loss, shadowing loss, and multipath fading—occur along the distance dimension. They introduce losses into the channel and degrade the received power. These impairments can be compensated by adjusting the transmit power and adding enough link margin. Specifically, path loss is compensated by including it in the link budget during system design. Shadowing loss is compensated by including an additional margin for shadowing loss in the link budget, and the exact margin to include depends on the availability that the system designer desires. In practice, shadowing loss can also be compensated in real time by power control.

Because power loss due to multipath fading can be large (e.g., –20 dB), system designers typically do not include a margin for multipath fading in the link budget. Instead, it is expected that channel coding (i.e., forward error correction) applied in the physical layer can reduce the amount of transmit power required. Whatever additional amount of power is still required is then compensated by power control. In short, multipath fading requires that modulation and coding be carefully designed and chosen to ensure that a sufficient bit error rate is maintained at a given SINR. For power control, because multipath fades occur over such short distances (and over such short intervals if there is relative motion between the transmitter and the receiver), power control that compensates for multipath fading typically has to be faster than power control that compensates shadowing loss.

In the distance dimension, using power control to combat shadowing loss and multipath fading actually leverages a form of diversity called *multiuser diversity*.

The idea is that in the coverage area of a base station, there is a multitude of mobiles, and these mobiles are at different locations. As a result, it is unlikely that these mobiles would all experience fades at the same time. It is more likely that, at any given moment, some of them would experience fades (thus requiring more power from the base station) and some of them would not experience fades (thus requiring less power from the base station). This is how a power amplifier in the base station, with a fixed transmit power budget, can serve many mobiles. Of course, if all mobiles served by a base station are experiencing fades at the same time, then there would not be enough transmit power available at the base station for all the mobiles; as a result, some links would experience high bit error rates, and some links would drop. Chapter 8 discusses power control as implemented by IEEE 802.16e.

2.4 Time Dimension: Multipath Delay Spread

Contrary to different types of propagation loss described in the previous section, multipath delay spread cannot be compensated by only adjusting the transmit power at the system level. Multipath occurs when signals arrive at the receiver via different paths due to reflection by or transmission through objects. The amount of signal reflection depends on factors such as angle of arrival, carrier frequency, and polarization of incident wave. As discussed previously, the multipath phenomenon causes multipath fading, which occurs over short distances due to constructive and destructive interferences of signals of multiple paths. This impairment takes place in the *distance dimension*. The same multipath phenomenon also causes another type of impairment in the *time dimension*. Multipath delay spread occurring in the time dimension is also referred to as *time dispersion*.

2.4.1 Delay Spread

In multipath, because path lengths are different among different paths, different signal paths could arrive at the receiver over different distances at different times. Figure 2.6 illustrates the concept. An impulse is transmitted at time 0; assuming that there is a multitude of reflected paths present, a receiver approximately 1-km away should detect a series of pulses, or *delay spread*. Due to multipath, delay spread originates from a single impulse transmitted in time.

If the time difference τ_{MAX} between the earliest arriving pulse and the latest arriving pulse is significant compared to one symbol period, *intersymbol interference* (ISI) can occur.[4] In other words, symbols arriving significantly earlier or later than its own symbol period can corrupt preceding or trailing symbols. Given fixed path differences and a specific delay spread, a higher symbol-rate system is more likely to suffer ISI. Given a fixed symbol-rate system, an environment with longer path differences (and thus a higher delay spread) is more likely to result in ISI.

4. In practice, the root-mean square delay spread τ_{RMS} is typically used to characterize the length of a significant spread.

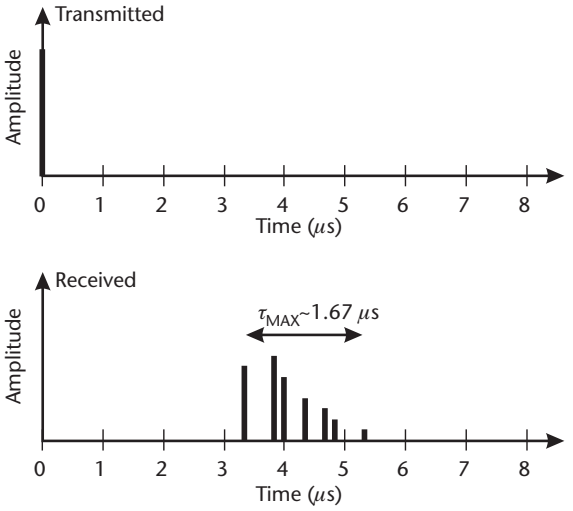


Figure 2.6 An example of delay spread. Here the delay spread is the result of a single impulse transmitted in time.

Example 2.2

Determine if the delay spread profile shown in Figure 2.6 will cause ISI to a mobile communication system using 270.833 ksps as its symbol rate.

$$R_s = 270.833 \text{ ksps}$$
$$T_s = \frac{1}{R_s} = \frac{1}{270.833 \times 10^3 \text{ sps}} = 3.69\,\mu\text{s}$$

Since the symbol period is only a little more than twice the delay spread (1.67 μs) shown in Figure 2.6, ISI could occur in this situation without any use of equalization. Note that the GSM system uses a symbol rate of 270.833 ksps.

Example 2.3

Determine if the delay spread profile shown in Figure 2.6 will cause ISI to a mobile communication system using 1.2288 Msps as its symbol rate.[5]

$$R_s = 1.2288 \text{ Msps}$$
$$T_s = \frac{1}{R_s} = \frac{1}{1.2288 \times 10^6 \text{ sps}} = 0.81\,\mu\text{s}$$

Since in this case the symbol period is less than the delay spread (1.67 μs), ISI would normally occur. Note that the IS-95 CDMA system uses a symbol rate of 1.2288 Msps.

5. In IS-95, it is also known as the *chip rate* because symbols are used to transmit chips in IS-95.

The way the IS-95 CDMA system deals with ISI is to use a special form of time diversity to recover the signal. The system uses a *rake receiver* to lock onto different multipath components. If a time reference is provided, then different multipath components can be separately identified as distinct echoes of the signal separated in time. These separately identified components of the received signal can then be brought in phase and combined to yield a final composite received signal [14].

However, the IS-95 CDMA system cannot separately identify or resolve multipath components that are less than 0.81 μs apart. In a dense urban environment such as New York City, where base stations are very close to each other and each base station is operating at low power, multipath components may arrive at a spread of less than 0.81 μs with very low power. In this case, IS-95 CDMA would not be able to resolve the components and combine their powers to yield a usable signal. This is one of the reasons UMTS (also called wideband CDMA or WCDMA) uses a higher symbol rate of 3.84 Msps, and as such it can theoretically resolve multipath components that are 0.26 μs apart.

To deal with ISI, OFDM artificially decreases the symbol rate (and hence decreases the bandwidth) via the use of subcarriers. See Section 2.4.3 for more details.

2.4.2 Coherence Bandwidth

Although multipath delay spread is best visualized in the time domain, it has a physical manifestation in the frequency domain as well. The delay spread that occurs in the time domain is directly related to a quantity called the *coherence bandwidth* W_c in the frequency domain. The coherence bandwidth is the range of frequency over which the transfer function $H(f)$ of the channel varies little, that is,

$$H(f) \approx H(f + \Delta f) \text{ where } |\Delta f| \leq W_c \tag{2.26}$$

$$H(f) \neq H(f + \Delta f) \text{ where } |\Delta f| > W_c \tag{2.27}$$

In other words, delay spread τ_{MAX} in the time domain characterizes the duration of time over which the channel impulse response $h(t)$ lasts, whereas coherence bandwidth W_c specifies the range of frequency over which the channel transfer function $H(f)$ is approximately the same. Another way to look at it is that coherence bandwidth W_c is the range of frequency over which the channel is correlated. In fact, the two quantities, coherence bandwidth and delay spread, are inversely proportional to each other, that is,

$$W_c \approx \frac{1}{\tau_{MAX}} \tag{2.28}$$

If one defines the coherence bandwidth as the range of frequency over which the correlation is greater than 0.5, then the coherence bandwidth is [15]:

$$W_c \approx \frac{1}{5\tau_{RMS}} \tag{2.29}$$

The physical manifestation of delay spread in the time domain is frequency selectivity in the frequency domain. In other words, *delay spread* in the time domain translates directly into a *frequency selective* channel in the frequency domain. This makes sense because if there is no delay spread, then the impulse response of the channel is a single impulse (admittedly at a smaller amplitude). The frequency response of the channel in the frequency domain is the Fourier transform of the impulse response in the time domain. Given that the Fourier transform of an impulse is a constant (flat), the frequency response of a channel with no delay spread is flat, or constant over all frequencies. In this case, all components at all frequencies would experience the same attenuation.

In general, if the bandwidth-delay spread product of the channel is greater than or equal to 0.1, that is,

$$W\tau_{MAX} \geq 0.1 \tag{2.30}$$

then the channel is seen as being frequency selective. If the bandwidth-delay spread product is less than 0.1, that is,

$$W\tau_{MAX} < 0.1 \tag{2.31}$$

then the channel is treated as frequency flat [16, 17].

Let's use a simple two-path model to illustrate how a channel is no longer flat if there are multipaths. Assume that there are two multipaths having the same amplitude A as shown in Figure 2.7. One multipath is delayed by τ_{MAX} relative to the other multipath.

The received signal is

$$r(t) = As(t) + As(t - \tau_{MAX}) \tag{2.32}$$

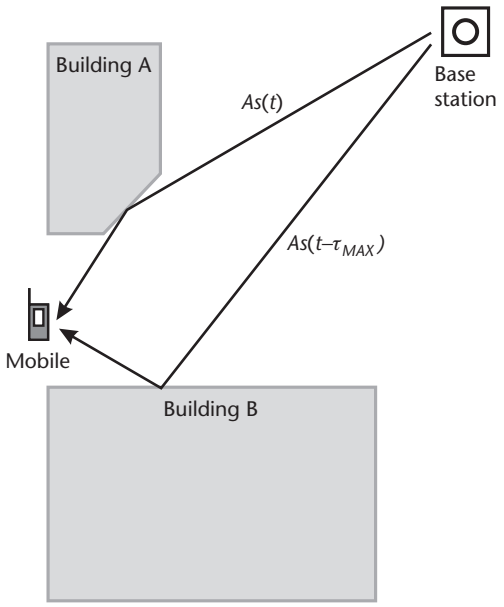


Figure 2.7 Two multipath components separated by time τ_{MAX}.

By taking the Fourier transform, we arrive at the spectrum of $r(t)$

$$R(f) = AS(f) + AS(f)e^{-j2\pi f\tau_{MAX}} \tag{2.33}$$

which can be rewritten as

$$R(f) = AS(f)\left[1 + e^{-j2\pi f\tau_{MAX}}\right] = AS(f)H(f) \tag{2.34}$$

Here $H(f)$ is effectively the transfer function of the channel that transforms the original signal $AS(f)$. $H(f)$ can also be written as

$$\begin{aligned} H(f) &= 1 + e^{-j2\pi f\tau_{MAX}} = e^{-j2\pi f\left(\tau_{MAX}/2\right)}\left[e^{j2\pi f\left(\tau_{MAX}/2\right)} + e^{-j2\pi f\left(\tau_{MAX}/2\right)}\right] \\ &= 2e^{-j2\pi f\left(\tau_{MAX}/2\right)}\cos\left(2\pi f\left(\tau_{MAX}/2\right)\right) \end{aligned} \tag{2.35}$$

and the magnitude spectrum $|H(f)|$ is

$$|H(f)| = 2\cos\left(2\pi f\left(\tau_{MAX}/2\right)\right) \tag{2.36}$$

$|H(f)|^2$ is shown in Figure 2.8. The frequency-selective fading is evident in the nulls of the magnitude spectrum as a result of multipath delay. The channel is said to be a frequency selective channel, where different components at different frequencies experience different degradations.

Note that in this simple two-path model, the delay spread is τ_{MAX}, thus, the coherence bandwidth W_c is:

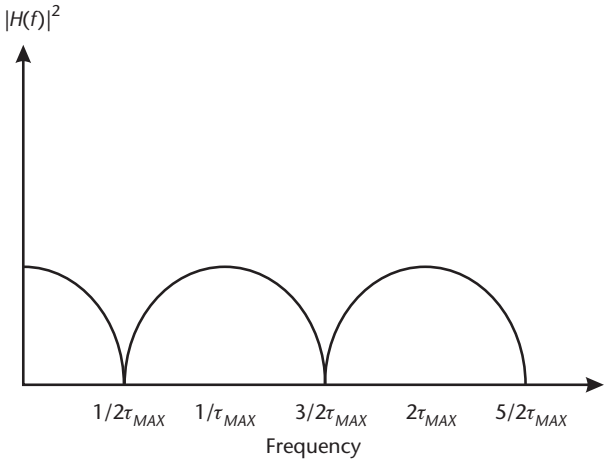


Figure 2.8 Frequency-selective fading is evident in the nulls of the transfer function.

$$W_c \approx \frac{1}{\tau_{MAX}}$$

which happens to be the width between successive nulls.

2.4.3 Implications for OFDM

The delay spread (and consequently the coherence bandwidth) turns out to be an important parameter in OFDM. Recall in Chapter 1 that the bandwidth of a random discrete-time waveform is mostly confined to $R_s = 1/T_s$, where R_s is the symbol rate and T_s is the symbol time or the duration of a symbol. By using narrower subcarriers, OFDM makes R_s small, and a small R_s leads to a large T_s. When symbol time T_s becomes much larger than the channel delay spread τ (i.e., $T_s >> \tau$), a symbol suffers little from delay spread, or delay spread has little effect on a symbol. This is how OFDM combats ISI in the time domain.

In the frequency domain, dividing a carrier into narrower subcarriers (as in OFDM) has the following advantage: each narrow subcarrier is said to be experiencing a flat channel. To see this, let's substitute (2.28) into (2.26) and (2.27). The substitutions yield:

$$H(f) \approx H(f + \Delta f) \text{ where } |\Delta f| \leq \frac{1}{\tau_{MAX}} \tag{2.37}$$

$$H(f) \neq H(f + \Delta f) \text{ where } |\Delta f| > \frac{1}{\tau_{MAX}} \tag{2.38}$$

As τ_{MAX} becomes smaller, the range of frequency over which $H(f)$ does not change (i.e., is flat) becomes larger. Of course, in the limit when $\tau = 0$ (e.g., in free space where there is no delay spread), $H(f)$ is identical for all frequencies. When there is no delay spread in the time domain, then there is a frequency flat channel in the frequency domain.

In the terrestrial environment, a true frequency flat channel is not possible because there is always some delay spread. However, the effect of frequency flatness can be approximated by keeping R_s and the bandwidth of each subcarrier less than the coherence bandwidth. If the bandwidth R_s of each subcarrier is much less than the coherence bandwidth, that is,

$$R_s << W_c \text{ or } R_s << \frac{1}{\tau_{MAX}} \tag{2.39}$$

then according to expression (2.37) each subcarrier can be said to be experiencing an approximately flat channel (see Figure 2.9). Note that a frequency flat channel can be easily corrected by using an equalizer at the receiver.

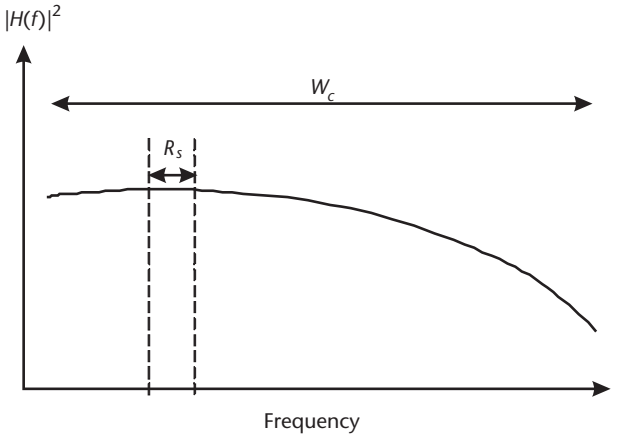


Figure 2.9 By keeping the bandwidth of a subcarrier much less than the coherence bandwidth, the subcarrier is experiencing an approximately flat channel.

2.5 Frequency Dimension: Doppler Spread

The two categories of impairment discussed previously (i.e., path loss and multipath delay spread) are applicable to all wireless environments, including stationary, nomadic, and mobile. A third category of impairment exists only in a *mobile* wireless channel, which is applicable when either the transmitter or the receiver is moving relative to each other,[6] or one or more scattering objects between the transmitter and the receiver are moving. This impairment is called *Doppler spread* and is often modeled first in the frequency domain; Doppler spread is also referred to as *frequency dispersion*. This section examines Doppler spread first from the perspective of the frequency domain and then from the perspective of the time domain.

2.5.1 Doppler Spread

We are all familiar with the phenomenon of the changing pitch of a train whistle as it moves toward us or away from us. When a train moves toward us, the frequency of its whistle sounds higher. When the train moves away, the frequency of its whistle sounds lower. The same phenomenon exists for electromagnetic waves emitted by a wireless transmitter. This impairment occurs in the *frequency dimension*. When there is relative motion between the transmitter and the receiver, the change in frequency or Doppler shift is

$$f_D = -v\frac{f}{c}\cos\phi \tag{2.40}$$

where f_D is the Doppler shift, v is the relative velocity between the transmitter and the receiver, f is the carrier frequency of the signal, c is the speed of light, and ϕ is the angle between the velocity of the relative motion and the direction of the transmis-

6. In a mesh network, both the transmitter and the receiver can move at the same time.

sion (see Figure 2.10). If the receiver is moving directly toward the transmitter, then ϕ is 180° and the Doppler shift is at a maximum; the maximum Doppler shift is

$$f_D = v\frac{f}{c} \tag{2.41}$$

Example 2.4

A base station is located next to the freeway and has an omnidirectional transmit antenna. A receiver in a car moves toward the base station at 90 km/hour. Compare the maximum Doppler shifts experienced by two carrier signals: one at 2.5 GHz and another at 700 MHz. The receiver has an omnidirectional receive antenna.

$$v = 90\frac{\text{km}}{\text{hour}} = 90\frac{\text{km}}{\text{hour}} \times \left(\frac{1{,}000\text{m}}{\text{km}}\right) \times \left(\frac{\text{hour}}{3{,}600}\right) = 25\frac{\text{m}}{\text{s}}$$

The maximum Doppler shifts experienced at 2.5 GHz and 700 MHz are:

$$f_{D,2.5\,\text{GHz}} = v\frac{f}{c} = (25\,\text{m/s})\left(\frac{2.5\times10^9\,\text{Hz}}{3\times10^8\,\text{m/s}}\right) = 208.33\,\text{Hz}$$

$$f_{D,700\,\text{MHz}} = v\frac{f}{c} = (25\,\text{m/s})\left(\frac{700\times10^6\,\text{Hz}}{3\times10^8\,\text{m/s}}\right) = 58.33\,\text{Hz}$$

Thus, a lower-frequency carrier experiences a smaller Doppler shift than a higher-frequency carrier.

If there is relative motion between the transmitter and the receiver, a carrier with a carrier frequency f received at the receiver is shifted to $f + f_D$. When the transmitter and the receiver are moving directly toward each other, the relative velocity v is positive. Thus, a carrier with carrier frequency f is shifted to $f + |f_D|$. When they are moving away from each other, the relative velocity is negative. Thus, the carrier frequency f at the receiver is shifted to $f - |f_D|$. Figure 2.11 shows what happens to a single carrier transmitted in free space when there is relative motion.

Figure 2.11 shows what happens to a single carrier in frequency when there is no reflection (as in free space). But what if there are many reflected paths (as in a terrestrial environment)? To model this phenomenon on the ground, we proceed to modify (2.22) to incorporate the Doppler shift. Because there is now a relative motion between the transmitter and the receiver, the N reflected paths now experience a Doppler shift. In particular, each path n has an amplitude of s_n, and each

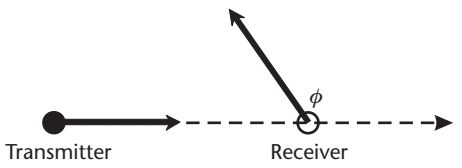


Figure 2.10 Relative motion between a transmitter and a receiver.

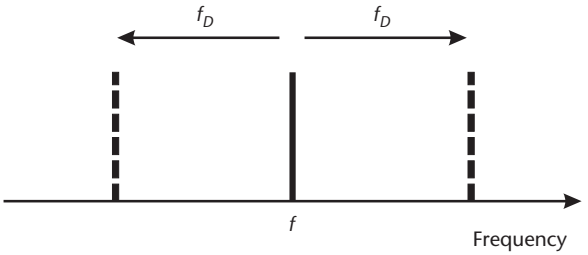


Figure 2.11 The relative motion between the transmitter and the receiver causes the received frequency of a single carrier to shift. Here the Doppler shift comes from a single carrier transmitted in free space.

path experiences a frequency shift $f_{D,n}$ due to relative motion. The received signal envelope **s** in complex form is:

$$\mathbf{s} = \sum_{n=1}^{N} s_n e^{j\left(\theta_n + 2\pi f_{D,n} t\right)} \tag{2.42}$$

where θ_n is the phase of the nth signal path uniformly distributed between 0 and 2π, and **s** is a composite of N signal paths. By definition, the autocorrelation function of **s** is

$$R(\Delta t) = E\left[\mathbf{s}(t)\mathbf{s}^*(t+\Delta t)\right] \tag{2.43}$$

where **s*** is the complex conjugate of **s** and E is the expectation operator. It is well known that the power spectrum is the Fourier transform F of the autocorrelation function, so the power spectrum is given by

$$S(f) = F\left\{R(\Delta t)\right\} \tag{2.44}$$

which is also known as the Doppler spectrum. It describes the power spectrum of the received signal at different frequencies. It has been shown that, under certain assumptions, (2.44) can be analytically derived to be [13, 18]:

$$\begin{aligned} S(f) &= \frac{s_0}{f_D\sqrt{1-\left(\dfrac{f-f_c}{f_D}\right)^2}} \quad \text{for } |f-f_c| < f_D \\ S(f) &= 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned} \tag{2.45}$$

where f_c is the carrier frequency. $S(f)$ is shown in Figure 2.12.

If one transmits a single carrier (at a single frequency), then due to relative motion, reflections, and Doppler spread, the received signal would have a power spectrum that is spread (as shown in Figure 2.12). In a real terrestrial environment, the actual Doppler spectrum shows that power is distributed toward $+f_D$ and $-f_D$. Just as multipath causes a single impulse transmitted in time to spread out in time, relative motion causes a single impulse transmitted in frequency to spread out in frequency.

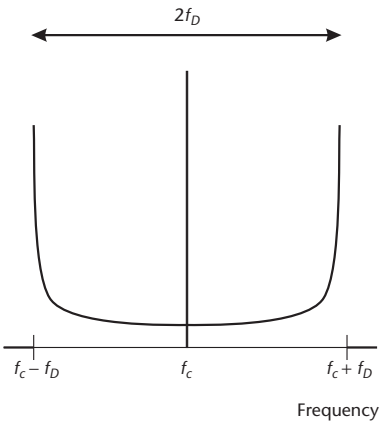


Figure 2.12 The Doppler spectrum when there are reflections. As an illustration, f_D is 208.33 Hz, which is the Doppler shift experienced by a 2.5-GHz carrier when the relative speed is 25 m/s. s_0 = 1 in the graph.

2.5.2 Coherence Time

The Doppler spread $2f_D$ that occurs in the frequency domain is directly related to a quantity called *coherence time* T_c in the time domain. The coherence time is the period of time over which the impulse response $h(t)$ of the channel varies little, that is,

$$h(t) \approx h(t + \Delta t) \text{ where } |\Delta t| \leq T_c \tag{2.46}$$

$$h(t) \neq h(t + \Delta t) \text{ where } |\Delta t| > T_c \tag{2.47}$$

In other words, the Doppler spread $2f_D$ in the frequency domain specifies the range of frequency over which the channel transfer function $H(f)$ is spread (due to motion), whereas the coherence time T_c specifies the period of time over which the channel impulse response $h(t)$ is approximately the same. Another way to look at it is that the coherence time T_c is the period of time over which the channel is correlated. In fact, the two quantities, coherence time and Doppler spread, are inversely proportional to each other, that is,

$$T_c \approx \frac{1}{2f_D} \tag{2.48}$$

Another expression for coherence time is given by [15]:

$$T_c \approx \frac{3}{4\sqrt{\pi} f_D} \tag{2.49}$$

The physical manifestation of Doppler spread in the frequency domain is time selectivity in the time domain. In other words, *Doppler spread* in the frequency domain translates directly into a *time selective* channel in the time domain. A time selective channel means that the channel is time-varying (i.e., changing with time). This makes sense because if there is relative motion between the transmitter and the receiver, then the channel cannot be expected to remain constant over time. One

simple example is a mobile moving through the coverage area of a base station and undergoing multipath fading as a function of time.

Let's examine a simple case of a mobile undergoing a changing channel because of multipath fading. Assume that there are two equally strong signals that are 180° out of phase. Then a standing-wave pattern would form with fades occurring once every half wavelength (see Figure 2.5). The distance Δd between two fades is then:

$$\Delta d = \frac{c}{2f} \tag{2.50}$$

where c is the speed of light and f is the carrier frequency. Also assume that the mobile is moving at a velocity v directly toward a stationary transmitter. This means that the time Δt between fades experienced by the mobile is

$$\Delta t = \frac{\Delta d}{v} = \frac{c}{2fv} \tag{2.51}$$

which is the same as coherence time T_c, that is,

$$T_c \approx \frac{1}{2f_D} = \frac{1}{2\left(v\frac{f}{c}\right)} = \frac{c}{2fv} \tag{2.52}$$

In other words, in this simple two-path model, coherence time is the same as the period of time between two consecutive fades experienced by a mobile moving at a velocity v.

2.5.3 Implications for OFDM

A mobile wireless channel implies that the mobile is moving; hence, the channel between the base station and the mobile is by definition dynamic and changing with time. How often the channel changes is stipulated by (2.48), which states that coherence time T_c over which the channel is approximately constant is inversely proportional to Doppler spread $2f_D$. It follows, therefore, that if the symbol time T_s is much less than the coherence time T_c (i.e., $T_s << T_c$), then each symbol should suffer little from Doppler spread. To see this, let's substitute (2.48) into expressions (2.46) and (2.47). The substitutions yield:

$$h(t) \approx h(t + \Delta t) \text{ where } |\Delta t| \leq \frac{1}{2f_D} \tag{2.53}$$

$$h(t) \neq h(t + \Delta t) \text{ where } |\Delta t| > \frac{1}{2f_D} \tag{2.54}$$

As f_D becomes smaller, the time duration over which $h(t)$ does not change (i.e., is flat) becomes longer. In the limit when $f_D = 0$ (i.e., if there is no relative motion between the transmitter and the receiver), then $h(t)$ is the same for all time. In other words, when there is no Doppler spread in the frequency domain, then there is a time flat (or time invariant) channel in the time domain.

In a broadband mobile system, Doppler shift has two important implications for an OFDM system. First, the Doppler shift degrades the orthogonality between subcarriers. In OFDM, subcarriers are carefully placed and centered at equally spaced subcarrier frequencies. For each subcarrier, its zero crossings are precisely at where the peaks of adjacent subcarriers are. Because data samples are taken at the peaks of subcarriers, there is no interference among subcarriers, and orthogonality among subcarriers is maintained. However, this orthogonality is degraded if subcarriers are misaligned due to Doppler shift, and the problem is exacerbated if f_D becomes significant as compared to the subcarrier bandwidth R_s. Chapter 13 examines the Doppler shift in OFDM system design.

Second, a high Doppler spread requires more frequent channel feedback. For example, if a receiver is moving at 90 km/hour (25 m/s) toward the transmitter operating at 2.5 GHz, then the coherence time T_c is 2.4 ms according to (2.48). This means that the channel changes once every 2.4 ms—very quickly. The link can be improved if the receiver gives feedback on the channel back to the transmitter. Power control feedback sent by the receiver is one such channel feedback that characterizes only the amplitude response of the channel. More sophisticated feedback involves the receiver sending information on both amplitude and phase responses of the channel, which can be utilized, for example, by multiple-antenna techniques. Power control is further discussed in Chapter 8. Multiple-antenna techniques are discussed in Chapter 6.

2.6 Conclusions

This chapter addresses the different kinds of impairments that are introduced by a wireless channel. The path loss, shadowing loss, and multipath fading occur in the *distance dimension* in that they primarily attenuate the signal power as a function of distance.

The delay spread occurs in the *time dimension* and is caused by multipath. Figure 2.13 summarizes the impairments that have a physical cause of multipath. The phenomenon of multipath results in the delay spread τ, also known as time dispersion in the time domain. The delay spread τ results in a frequency selective channel in the frequency domain, and that channel is characterized by coherence bandwidth W_c. The delay spread introduces ISI in time and results in a frequency selective channel in frequency. Of course, if there is no multipath, there is no delay spread in the time domain, and then the channel in frequency is a frequency flat channel.

The Doppler spread occurs in the frequency dimension and is present only if there is relative motion between the transmitter and the receiver. Figure 2.14 summarizes the impairments that have a physical cause of relative motion between the transmitter and the receiver. Such a motion results in the Doppler spread $2f_D$, also known as frequency dispersion in the frequency domain. Doppler spread results

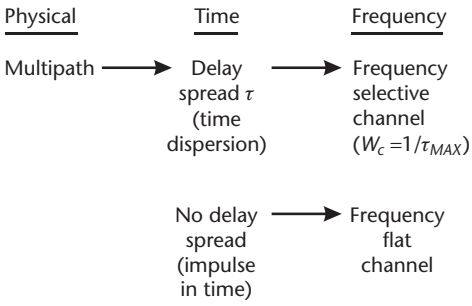


Figure 2.13 Manifestations of multipath.

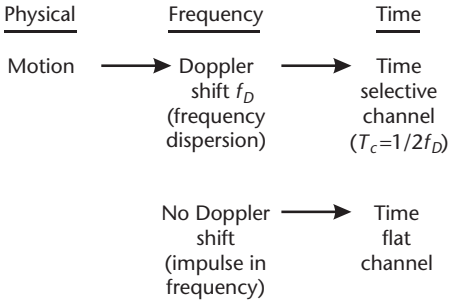


Figure 2.14 Manifestations of motion between the transmitter and the receiver.

in a time selective (time varying) channel in the time domain, and that channel is characterized by coherence time T_c. Doppler spread degrades orthogonality among OFDM subcarriers and requires frequent feedback on the channel from the receiver. If there is no motion, then there is no Doppler spread in the frequency domain, and the channel in time is a time flat (time invariant) channel.

Figure 2.15 depicts a 2×2 framework of the different types of wireless channels.

Given these different types of wireless channels, Figure 2.16 presents a system design framework that shows the generic countermeasures appropriate for the different wireless channels.

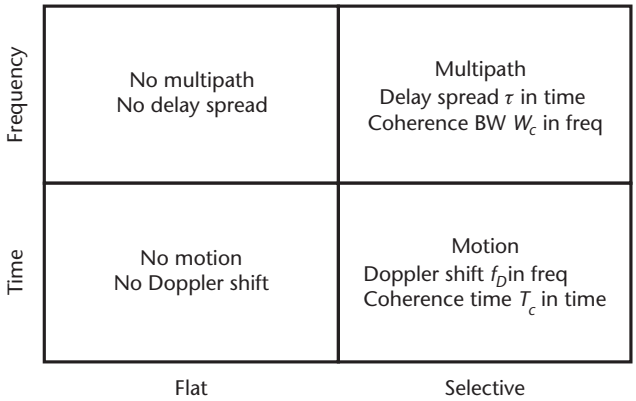


Figure 2.15 Four different types of wireless channels.

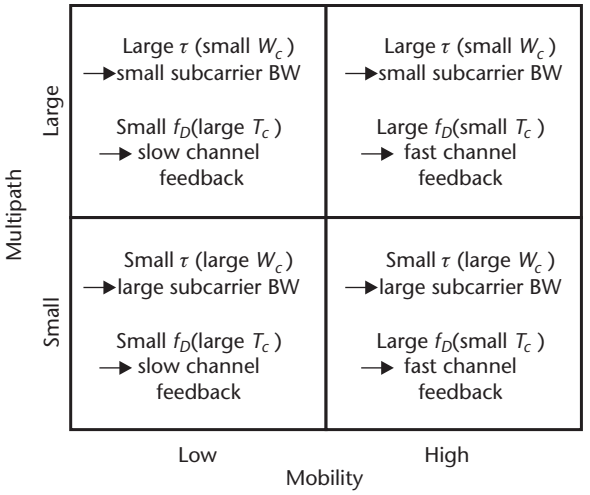


Figure 2.16 A system design framework for the different wireless channels. Note that the IEEE 802.16e standard fixes the subcarrier spacing and does not allow subcarrier bandwidth to vary.

References

[1] Yang, S. C., *CDMA RF System Engineering*, Norwood, MA: Artech House, 1998.

[2] Erceg, V., et al., "An Empirically Based Path Loss Model for Wireless Channels in Suburban Environments," *IEEE Journal on Selected Areas of Communications*, Vol. 17, No. 7, 1999, pp. 1205–1211.

[3] Okumura, Y., "Field Strength and Its Variability in UHF and VHF Land-Mobile Radio Service," *Review Electrical Communication Laboratory*, Vol. 16, No. 9–10, 1968, pp. 825–873.

[4] Hata, M., "Empirical Formula for Propagation Loss in Land Mobile Radio Services," *IEEE Trans. on Vehicular Technology*, Vol. 29, No. 3, 1980, pp. 317–325.

[5] COST Telecom Secretariat, *Digital Mobile Radio Towards Future Generation Systems: COST 231 Final Report*, Brussels: European COST Action 231, 1999.

[6] "Mobile WiMAX—Part I: A Technical Overview and Performance Evaluation," *WiMAX Forum*, August 2006.

[7] MWG/AWG, "A Comparative Analysis of Spectrum Alternatives for WiMAX Networks with Deployment Scenarios Based on the U.S. 700 MHz Band," *WiMAX Forum*, June 2008.

[8] Abhayawardhana, V. S., et al., "Comparison of Empirical Propagation Path Loss Models for Fixed Wireless Access Systems," *Proc. 61st IEEE Vehicular Technology Conference, Vol. 1*, Spring, May 30–June 1, 2005, pp. 73–77.

[9] Walfisch, J., and H. L. Bertoni, "A Theoretical Model of UHF Propagation in Urban Environments," *IEEE Trans. on Antennas and Propagation*, Vol. 36, No. 12, 1988, pp. 1788–1796.

[10] Lee, W. C. Y., *Mobile Communications Design Fundamentals*, New York: John Wiley & Sons, 1993.

[11] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

[12] Hess, G. C., *Land-Mobile Radio System Engineering*, Norwood, MA: Artech House, 1993.

[13] Jakes, W. C., *Microwave Mobile Communications*, New York: Wiley-IEEE Press, 1994.

{14} Mehrotra, A., *Cellular Radio Performance Engineering*, Norwood, MA: Artech House, 1994.

[15] Rappaport, T. S., *Wireless Communications: Principles and Practice*, Upper Saddle River, NJ: Prentice-Hall, 2002.

[16] Paulraj, A. J., et al., "An Overview of MIMO Communications—A Key to Gigabit Wireless," *Proceedings of the IEEE*, Vol. 92, No. 2, 2004, pp. 198-218.

[17] Goldsmith, A., *Wireless Communications*, Cambridge, U.K.: Cambridge University Press, 2005.

[18] Clarke, R. H., "A Statistical Theory of Mobile Radio Reception," *Bell Systems Technical Journal*, Vol. 47, No. 6, 1968, pp. 957–1000.

Selected Bibliography

Lee, W. C. Y., *Mobile Cellular Telecommunications: Analog and Digital Systems*, New York: McGraw-Hill, 1995.

Lee, W. C. Y., *Mobile Communications Engineering*, New York: McGraw-Hill, 1997.

Rappaport, T. S., *Wireless Communications: Principles and Practice*, Englewood Cliffs, NJ: Prentice-Hall, 1995.

Yacoub, M. D., *Foundations of Mobile Radio Engineering,* Boca Raton, FL: CRC Press, 1993.

CHAPTER 3

Fundamentals of Digital Communications and Networking

3.1 Introduction

In this chapter we address some of the fundamental issues in digital communication and networking, particularly as applied to modern broadband wireless networks.[1] In particular, this chapter considers some key functions of the medium access control (MAC) layer and the physical layer and also describes how these layers handle errors. Figure 3.1 shows the two layers, where an upper layer at the sender wants to send the data to the corresponding upper layer at the receiver. As such, the upper layer first hands its data to the MAC layer (layer 2) below it.

At the sender, the MAC layer organizes the data into a MAC protocol data unit (PDU). The MAC layer then hands the MAC PDU to be sent to the physical layer.

The physical layer (layer 1) is responsible for actually transmitting the bits (i.e., 0 or 1) in the MAC PDU over the physical medium. To do so, the physical layer uses signals (i.e., sinusoidal waveforms) to physically transmit the bits. This chapter examines the way that the physical layer transmits the bits and also describes how the MAC layer handles error in a MAC PDU.

3.2 Basic Functions of a Transceiver

At the physical layer, discussions of the transmitter and the receiver thus far have been at the level of data symbols. In other words, the inputs at the left side of Figure 1.4 (OFDM transmitter) and Figure 1.7 (OFDMA transmitter) are data symbols. Before going into details of an actual system (i.e., IEEE 802.16e), it is necessary to examine how data symbols themselves are constructed and how they relate to data bits. In short, a data symbol is different from a data bit because one symbol can carry one or more bits, depending on the specific set of data symbols (i.e., "constellation") used.

1. Some parts of this chapter are adopted from Chapter 3 of [1] with enhancements and revisions tailored to those issues relevant to a high-speed OFDM system.

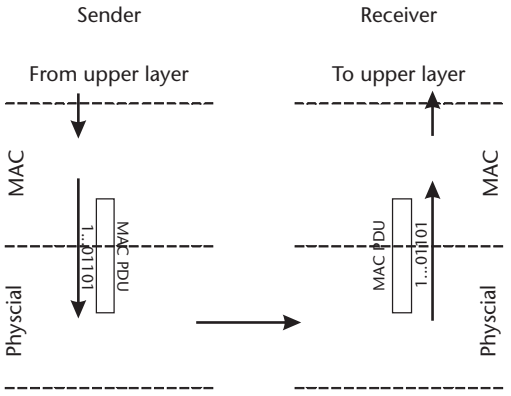


Figure 3.1 The MAC layer organizes data into a MAC PDU and passes it to the physical layer, which physically transmits the bits.

Figure 3.2 shows the functional block diagram of a typical digital wireless transmission and receiving system [2]. The data to be sent come from the MAC layer. The *channel coding* function codes the bits for the purpose of combating various degrading effects of the channel. Then the *symbol mapping* function maps the data bits to data symbols (drawn from a predetermined set or "constellation"). The *modulation* function converts the signals from baseband to bandpass waveforms that can be transmitted. The RF function performs the necessary filtering and power amplification functions before actual transmission.

On the receive side, the received bandpass waveform is intercepted by the antenna and goes through low-noise amplification and filtering in the RF function. The demodulation function converts the signals from RF to baseband. The symbol mapping function maps the data symbols back to data bits. Then the channel decoding function attempts to correct the errors in the data bits that have been introduced by the channel.

The subsequent sections discuss the channel coding, symbol mapping, and modulation functions shown in Figure 3.2.

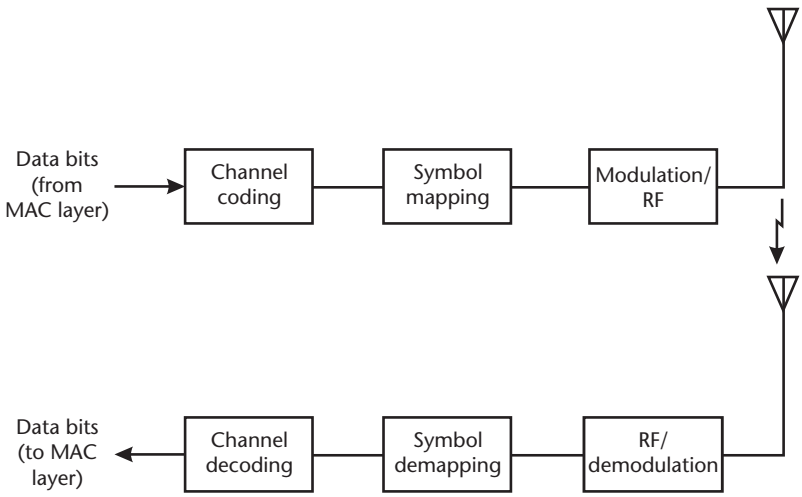


Figure 3.2 Main components of a digital transceiver.

3.3 Channel Coding

On a wireless channel, *redundancy* needs to be added to the information bits. This is done to improve performance by enabling the subsequent signals to better withstand the effects of channel impairments, such as interference and fading. The goal of channel coding is that, given a desired probability of bit error, the required energy per bit E_b over noise power density N_0 (E_b/N_0) is reduced; alternatively, given an achievable E_b/N_0, the probability of bit error is reduced. The cost of these two goals is more bandwidth or more redundancy bits that the system has to transmit [2].

This section specifically deals with error-correcting codes, which when applied to channel coding improve the error performance of the system. The purpose is to add extra bits to the information bits so that errors may be found and corrected at the receiver. In other words, a sequence of bits is represented by a longer sequence of bits with enough redundancy to protect the data [3]. For example, the simplest code is to repeat the information bits. Suppose there is a bit that I wish to send and error-protect. The bit can simply be repeated three times (i.e., if I have "1," I will send "111"). This way, I will improve the chance that the receiver correctly receives a "1" in case any one of the transmitted bits is flipped to "0" during transmission. In this case, the receiver will use *majority decoding*. Namely, the receiver will only decide a "1" if a majority of the three bits are received as 1s. This code is known as a (3, 1) code. (n, k) refers to a code where n is the length of the coded sequence and k is the length of the information sequence. A code can also be described by its rate. The code rate R of a code is defined as

$$R = \frac{k}{n} \tag{3.1}$$

Two popular classes of error-correcting codes are *block* codes and *convolutional* codes. Block codes, as the name implies, code information sequences one block at a time. Convolutional codes, on the other hand, have a memory property. The memory depends on the *constraint length* K of the convolutional code. The n-tuple output of a convolutional coder is not only a function of the current input k-tuple, but also a function of the previous $K - 1$ input k-tuples [2].

3.3.1 Linear Block Codes

Linear block codes are a class of codes that can be used for the purpose of error detection or error correction. A linear block code can be characterized by the (n, k) notation, and for a given code, the coder transforms a *block* of k information bits into a longer block of n code bits [4]. The code bits are only a function of the current block of information bits. For example, we can define a (7, 4) linear block code where a block of seven code bits is used to represent a block of four information bits. Given the four information bits (i_1, i_2, i_3, i_4), the three extra *redundancy* bits (r_1, r_2, r_3) are appended using the following functions [3]:

$$r_1 = i_1 + i_2 + i_3 \tag{3.2}$$

$$r_2 = i_2 + i_3 + i_4 \tag{3.3}$$

$$r_3 = i_1 + i_2 + i_4 \tag{3.4}$$

where + represents the modulo-2 addition. For example, if the information bits are (1, 0, 1, 0) corresponding to (i_1, i_2, i_3, i_4), then the extra redundancy bits are

$$r_1 = 1 + 0 + 1 = 0$$
$$r_2 = 0 + 1 + 0 = 1$$
$$r_3 = 1 + 0 + 0 = 1$$

and the code word (1, 0, 1, 0, 0, 1, 1) is used to represent the four information bits. Table 3.1 is a complete enumeration of this (7, 4) linear block code. This simple (7, 4) linear block code is also known as the (7, 4) *Hamming code*, and the redundancy bits are also known as the *parity* bits.

It is intuitive that the extra redundancy bits improve the error performance of the system. To quantify this performance, we introduce the concept of *Hamming distance*. The Hamming distance between any two code words is the number of places in which the two code words differ. For example, the Hamming distance between (1, 1, 1, 1, 1, 1, 1) and (1, 1, 1, 0, 1, 0, 0) is 3.

Table 3.1 (7, 4) Hamming Code

Information Bit Sequence (i_1, i_2, i_3, i_4)	*Redundancy Bits* (r_1, r_2, r_3)	*Code Sequence* (i_1, i_2, i_3, i_4, r_1, r_2, r_3)
0 0 0 0	0 0 0	0 0 0 0 0 0 0
0 0 0 1	0 1 1	0 0 0 1 0 1 1
0 0 1 0	1 1 0	0 0 1 0 1 1 0
0 0 1 1	1 0 1	0 0 1 1 1 0 1
0 1 0 0	1 1 1	0 1 0 0 1 1 1
0 1 0 1	1 0 0	0 1 0 1 1 0 0
0 1 1 0	0 0 1	0 1 1 0 0 0 1
0 1 1 1	0 1 0	0 1 1 1 0 1 0
1 0 0 0	1 0 1	1 0 0 0 1 0 1
1 0 0 1	1 1 0	1 0 0 1 1 1 0
1 0 1 0	0 1 1	1 0 1 0 0 1 1
1 0 1 1	0 0 0	1 0 1 1 0 0 0
1 1 0 0	0 1 0	1 1 0 0 0 1 0
1 1 0 1	0 0 1	1 1 0 1 0 0 1
1 1 1 0	1 0 0	1 1 1 0 1 0 0
1 1 1 1	1 1 1	1 1 1 1 1 1 1

The *minimum distance* d^* of a code is the Hamming distance of a pair of code words that have the smallest Hamming distance. For the Hamming code shown above, d^* is 3, which is the smallest Hamming distance for all possible pairs of code words. The minimum distance turns out to be a critical parameter that specifies the performance of a particular code. If t errors occur during the transmission of a code word, and if the (Hamming) distance between the received word and every other code word is larger than t, then the decoder will properly correct the errors if it assumes that the closest code word to the received word was actually transmitted [3]. In other words,

$$d^* \geq 2t + 1 \tag{3.5}$$

If (3.5) holds for a code, then this code is capable of *correcting* t errors. On the other hand, (3.6) summarizes the error detection capability q of a code.

$$d^* \geq q + 1 \tag{3.6}$$

If (3.6) holds for a code, then the code is capable of *detecting* q errors. Thus, given that d^* of the (7, 4) Hamming code is 3, the (7, 4) Hamming code is capable of correcting $t = 1$ error and detecting $q = 2$ errors.

As mentioned earlier, in order to decode a received code word, the decoder assumes that the closest code word to the received code word was actually transmitted. For example, suppose that a received code word is (0, 0, 0, 1, 1, 1, 1). Since this received code word is not one of the specified code words in the (7, 4) Hamming code, an error (or errors) must have occurred. Assuming that the closest code word to the received code word was actually transmitted, the decoder decides that the code word (0, 0, 0, 1, 0, 1, 1) was actually sent by the transmitter. This function can be easily implemented using a digital logic circuit.

In practice, block codes are popular for use at the MAC layer because they can be readily applied to a packet (block) of bits at the MAC layer. One popular code is the cyclic redundancy check (CRC), which is used to detect error in a packet received by the MAC layer at the receiver. If the CRC detects an error in a packet, then the receiver can request retransmission of that packet, thus correcting the error. The CRC is discussed later in this chapter.

3.3.2 Convolutional Codes

In addition to using block codes (e.g., CRC) at the MAC layer, broadband wireless systems also use forward error correction (FEC) at the physical layer. Convolutional codes are a popular way of implementing FEC. FEC codes not only can detect error but also can correct error without the need for retransmission.

The block codes are said to be *memoryless*, which means that the code word or the additional redundancy bits are only a function of the current block. The convolutional codes, on the other hand, do have memory. For convolutional codes, the coded bits are functions of information bits and the constraint length. Specifically, every coded bit (at the output of the convolutional coder) is a linear combination of some previous information bits. For example, an IEEE 802.16e system requires the use of convolutional codes. In particular, a rate 1/2, constraint length $K = 7$

convolutional coder is used as the baseline. Figure 3.3 shows the coding structure of this rate 1/2 convolutional coder [5].

Initially, all the registers are initialized to zero. As the information message bits m_i are clocked in from the left, bits are tapped off different stages of the delay line and summed in a modulo-2 adder. The summation is the output of the convolutional coder. Since this is a rate 1/2 coder, two bits are generated for each clock cycle. A commutator switch toggles through two output points (of the two adders) for every input clock cycle, hence the number of output bits is effectively twice the number of input bits. The generator function for the two output bits y_i' and y_i'' (shown in Figure 3.3) can also be written as

$$g'(x) = x^6 + x^3 + x^2 + x + 1 \tag{3.7}$$

$$g''(x) = x^6 + x^5 + x^3 + x^2 + 1 \tag{3.8}$$

To derive convolutional codes that have rates other than 1/2, the coder can first generate the bit sequence $y_i'y_i''$ using the rate 1/2 coder shown, then puncture out (delete) selected bits in the sequence. For example, to derive a rate 2/3 convolutional code, the coder would first generate the following bit sequence using the original rate 1/2 coder:

$$y_1'y_1''y_2'y_2''y_3'y_3''y_4'y_4''\cdots$$

Then the coder would puncture one out of every 4 bits (or include 3 out of every 4 bits):

$$y_1'y_1''y_2''y_3'y_3''y_4''\cdots$$

With puncturing, the effective code rate R becomes:

$$R = \frac{R'}{1 - r_p} \tag{3.9}$$

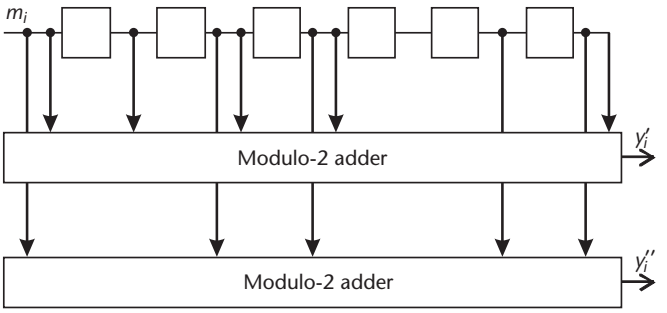


Figure 3.3 Convolutional coding (rate 1/2) in the IEEE 802.16e standard.

where R' is the original code rate and r_p is the puncturing ratio. For the above example, the effective code rate would then be

$$R = \frac{R'}{1 - r_p} = \frac{1/2}{1 - \frac{1}{4}} = \frac{1/2}{3/4} = \frac{2}{3}$$

Note that if there is no puncturing, then $R = R'$.

The decoding mechanism for convolutional codes is beyond the scope of this book. It suffices to mention that convolutional decoding uses a tree-search algorithm through a "trellis." The algorithm is a variant of *linear dynamic programming*. See [6] for a good discussion of convolutional decoding.

3.4 Symbol Mapping and Modulation

In a wireless communication system, the physical layer includes the wireless medium, so the physical layer uses *analog* signals to carry the bits. In other words, sinusoidal waveforms are used to transmit coded bits (0s and 1s) over the air.

In a digital communication system, the sinusoidal waveforms are "discrete time" in nature in that a single sinusoidal waveform (representing a bit or a group of bits) has a fixed duration with a well-defined start time and end time. Figure 3.4 shows such an arrangement. Here, the coded bits to be transmitted are 1101, and the following scheme is used to physically transmit the bits: if the data bit is a 1, then transmit a positive cosine waveform; if the data bit is a 0, then transmit a negative cosine waveform.

Figure 3.5 shows the physical implementation of this example. The baseband bit stream of 1101 goes into a symbol mapper, which maps each bit to a baseband *symbol*. The baseband symbol is +1 if the corresponding bit is 1 and –1 if

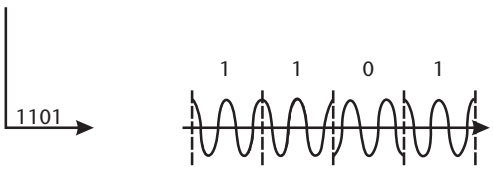


Figure 3.4 A simple signaling scheme (BPSK).

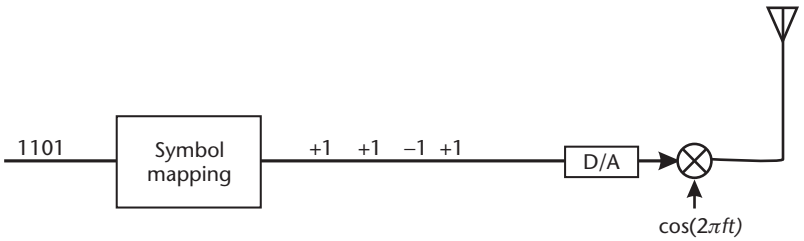


Figure 3.5 A simple digital transmitter (BPSK).

the corresponding bit is 0. The baseband symbol is then multiplied by a carrier $\cos(2\pi ft)$. The resulting passband waveforms $s(t)$ is then transmitted over the air.

Figure 3.6 shows the symbol constellation for this scheme. There are only two possible symbols {+1, –1}. In addition, these possible symbols are real numbers, have only real components, and lie on the real axis only; there are no imaginary components in these symbols. Also note that these two symbols differ only in their phase, not in their magnitude. The +1 symbol and the –1 symbol both have a magnitude of 1, but they have a phase difference of 180°. This scheme is known as the binary phase shift keying (BPSK).

In the example shown in Figures 3.4 to 3.6, the transmitted signals only make use of the real component. There are two symbols in the symbol constellation. As such, each symbol can only carry 1 bit of baseband data. However, if one adds two additional symbols along the imaginary axis,[2] then there would be a total of four symbols in the constellation. Figure 3.7 shows these four symbols, which are {+1,

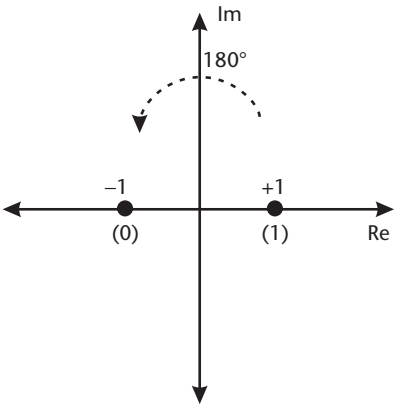


Figure 3.6 A simple symbol constellation (BPSK).

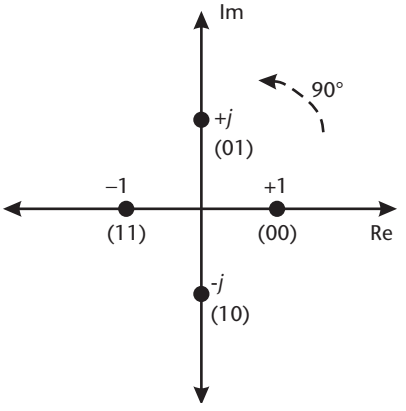


Figure 3.7 A symbol constellation consisting of four symbols (QPSK).

2. In digital communications, the real dimension (axis) is also known as the "in-phase" dimension, and the imaginary dimension (axis) is also known as the "quadrature" dimension.

-1, $+j$, $-j$}. These four symbols are real numbers and imaginary numbers. Note that the four symbols still differ only in their phase, not in their magnitude. All symbols have a magnitude of 1, but adjacent symbols now have a phase difference of 90°. This scheme is known as quadrature phase shift keying (QPSK).

Because there are now four symbols in the constellation, each symbol can carry two data bits at a time. Figure 3.7 also shows a possible way to assign groups of 2 bits to the symbols. For example, the +1 symbol carries the bits 00, and the $+j$ symbol carries the bits 01. The bit assignment shown in Figure 3.7 is an example of *Gray coding*, which assigns bits in such a way that adjacent symbols carry groups of bits that differ by 1 bit. This is done to minimize bit error rate. If, at the receiver, the +1 symbol got misinterpreted into a $+j$ symbol, then there would be at most one bit error (as opposed to two bit errors) as a result of this symbol error.

In general, the number of bits b carried by a symbol is

$$b = \log_2 M \tag{3.10}$$

where M is the total number of symbols in the constellation. In the previous example, there are four symbols ($M = 4$) in the constellation, so each symbol can carry 2 bits ($b = 2$) at a time.

For the same bit sequence 1101, Figure 3.8 shows the corresponding signaling scheme. The first two bits are 11, so a negative cosine (a cosine shifted by 180°) waveform is transmitted. The second two bits are 01, so a positive sine (a cosine shifted by 90°) waveform is transmitted.

Figure 3.9 shows the physical implementation of this example. The baseband bit stream of 1101 goes into a symbol mapper, which maps a group of two bits to a baseband symbol. The baseband symbol is –1 if the corresponding bits are 11, and the baseband symbol is $+j$ if the corresponding bits are 01. To show the actual implementation, Figure 3.9 depicts that the symbol mapper actually have two physical outputs: real and imaginary. Mapping the bits 11 into the symbol –1, the symbol mapper outputs –1 on the real line and 0 on the imaginary line during the corresponding symbol period. This is because the symbol –1 is more completely written as $-1 + 0j$ in complex form. Mapping the bits 01 into the symbol $+j$, the symbol mapper outputs 0 on the real line and +1 on the imaginary line during the next symbol period. Similarly, this is because the symbol $+j$ is more completely written as $0 + j$ in complex form.

In general, the real part (a_i) of the complex symbol is multiplied by a cosinusoidal carrier, and the imaginary part (b_i) of the complex symbol is multiplied by a sinusoidal carrier. The reason why the carriers are $\sqrt{2/T}\cos(2\pi ft)$ and $\sqrt{2/T}\sin(2\pi ft)$ is because the following integrals become unity,

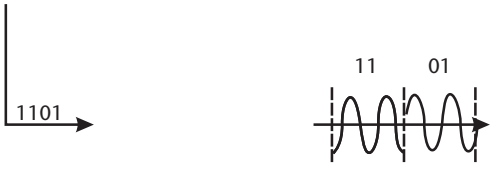


Figure 3.8 A more complex signaling scheme (QPSK).

$$\int_t^{t+T}\left(\sqrt{\frac{2}{T}}\cos 2\pi ft\right)^2 dt = \frac{2}{T}\int_t^{t+T}\cos^2(2\pi ft)\,dt = \frac{2}{T}\left(\frac{T}{2}\right) = 1 \tag{3.11}$$

where $f = k/T$ for $1, 2, \ldots$

$$\int_t^{t+T}\left(\sqrt{\frac{2}{T}}\sin 2\pi ft\right)^2 dt = \frac{2}{T}\int_t^{t+T}\sin^2(2\pi ft)\,dt = \frac{2}{T}\left(\frac{T}{2}\right) = 1 \tag{3.12}$$

where $f = k/T$ for $1, 2, \ldots$

which facilitate the mathematical modeling of the demodulation process later on. Note that T is the time duration of a symbol, and the integrations are over one symbol period.

To produce the resulting passband waveforms $s(t)$, the summer adds the real part and the imaginary part:

$$s(t) = a_i\sqrt{\frac{2}{T}}\cos 2\pi ft + b_i\sqrt{\frac{2}{T}}\sin 2\pi ft \tag{3.13}$$

For the sake of brevity Figure 3.9 and similar diagrams are often depicted as that shown in Figure 3.10 in complex form, where the complex multiplication is simply shown as the multiplication by $\exp(-j2\pi ft)$.

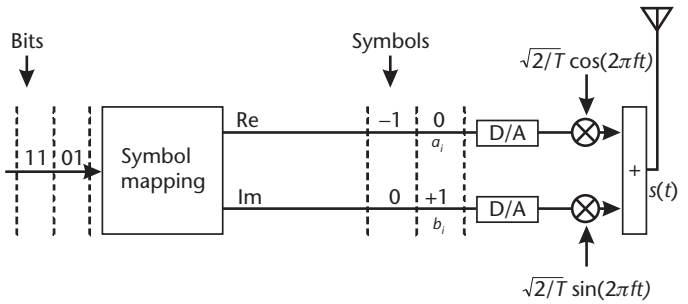


Figure 3.9 A more complex digital transmitter.

Figure 3.10 A complex equivalent of the more complex digital transmitter.

In general, data symbols are complex. Using complex symbols allows the system designer to exploit the imaginary dimension to add more symbols, and more symbols (i.e., higher M) means that more bits (i.e., higher b) can be carried in a symbol duration [see (3.10)]. Because the imaginary dimension (sine) is by definition orthogonal to the real dimension (cosine), using both dimensions to add more symbols is a rare free lunch in digital communication. For example, in the original IS-95 [7], the physical layer used BPSK (similar to Figure 3.6). The subsequent CDMA2000-1x [8] used QPSK (similar to Figure 3.7). As a result, part of the capacity gain from IS-95 to CDMA2000 came from the exploitation of the imaginary dimension in the physical layer.

The ability of one symbol to carry more than 1 bit of data is important because the RF bandwidth consumed by a transmitted signal is a function of the symbol rate, not the bit rate. In fact, a rule of thumb is that the RF bandwidth is taken to be approximately the symbol rate. Therefore, using increasingly higher order constellations (QPSK, 16-QAM, and so forth) enables more bits to be transmitted using a given symbol.

The IEEE 802.16e standard calls for the mandatory support of QPSK and 16-QAM in its OFDM and OFDMA options. These modulation schemes are supported on both uplink and downlink and are shown in Figure 3.11. Note that the QPSK constellation is rotated by 45° as compared to that shown in Figure 3.7. In

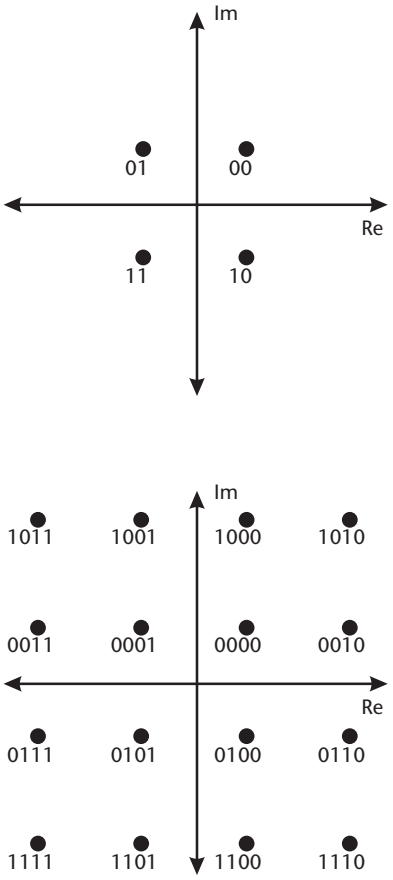


Figure 3.11 QPSK and 16-QAM supported by IEEE 802.16e.

addition, instead of using a magnitude of unity, the QPSK constellation shown has a magnitude of $\sqrt{E}$. This is because squaring the magnitude $\sqrt{E}$ yields E, which is the energy per symbol.

IEEE 802.16e can also optionally support 64-QAM as shown in Figure 3.12. The 64-QAM modulation may be supported on both downlink and uplink. Packing six data bits into each symbol, 64-QAM achieves the highest bandwidth efficiency at the cost of higher transmitter power. Also note that for all these modulation schemes (QPSK, 16-QAM, and 64-QAM) the bits assigned to a symbol differ from those of an adjacent symbol by at most 1 bit (i.e., Gray coding).

3.5 Demodulation

3.5.1 Matched Filter

This section looks at the demodulation function. One implementation of the demodulator is the *matched filter* approach shown in Figure 3.13. The received signal $r(t)$ consists of $s(t)$, which is the original transmitted signal, and $n(t)$, which is the noise introduced by the channel. In other words, $r(t) = s(t) + n(t)$. For simplicity, fading experienced by $s(t)$ is ignored for now. $r(t)$ is multiplied by both a cosinusoid and a sinusoid and then integrated to extract the real and imaginary components of the symbol.

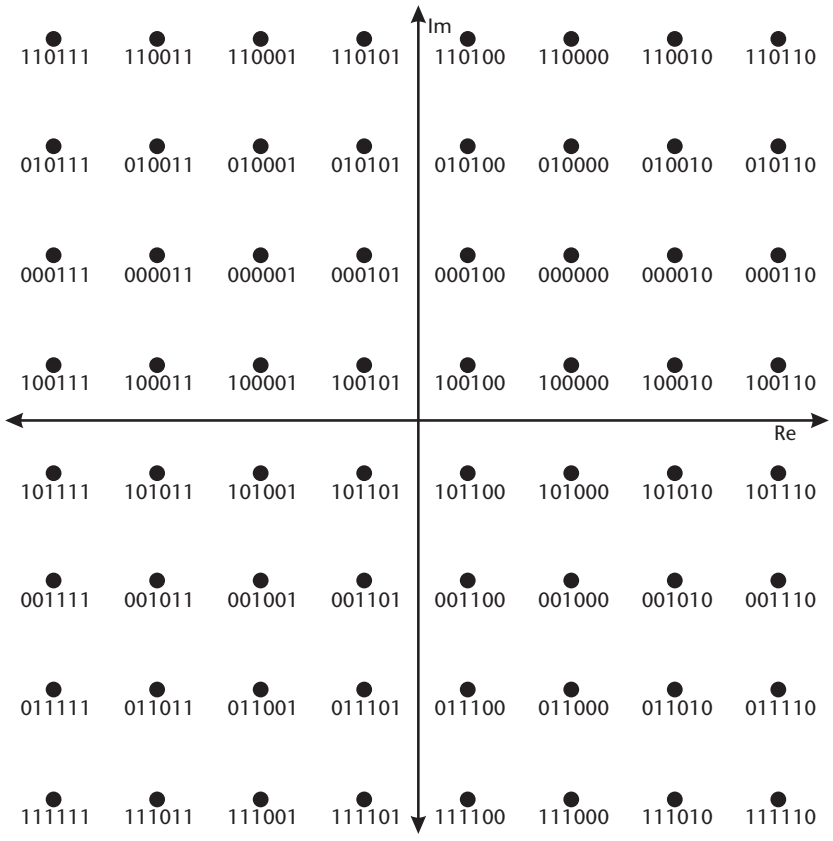


Figure 3.12 64-QAM supported by IEEE 802.16e.

Here, QPSK is used to illustrate the demodulation function. The QPSK constellation shown in Figure 3.11 is used where

$$a_i \in \left\{+\sqrt{\frac{E}{2}}, -\sqrt{\frac{E}{2}}\right\}$$
$$b_i \in \left\{+\sqrt{\frac{E}{2}}, -\sqrt{\frac{E}{2}}\right\}$$

for the purpose of scaling the energy E of each symbol.

After the integrators, the real path yields:

$$\begin{aligned} y_{\text{Re}} &= \int_t^{t+T} r(t)\sqrt{\frac{2}{T}}\cos(2\pi ft)\,dt \\ &= \frac{2}{T}\int_t^{t+T} a_i \cos^2(2\pi ft)\,dt + \frac{2}{T}\int_t^{t+T} b_i \sin(2\pi ft)\cos(2\pi ft)\,dt + \sqrt{\frac{2}{T}}\int_t^{t+T} n(t)\cos(2\pi ft)\,dt \\ &= a_i \frac{2}{T}\left(\frac{T}{2}\right) + b_i \frac{2}{T}(0) + \sqrt{\frac{2}{T}}\int_t^{t+T} n(t)\cos(2\pi ft)\,dt \\ &= a_i + \sqrt{\frac{2}{T}}\int_t^{t+T} n(t)\cos(2\pi ft) \\ &= a_i + n_{\text{Re}} \end{aligned} \tag{3.14}$$

where the first term a_i is the recovered real part of the symbol, and the second term n_{Re} is the noise contribution along the real dimension.

The integrated result is fed into the decision threshold. The decision threshold decides that a_i sent is $+\sqrt{E/2}$ if y_{Re} is greater than 0; the decision threshold decides that a_i sent is $-\sqrt{E/2}$ if y_{Re} is less than 0. In the absence of noise, only the first term remains at the integrator output, and the decision threshold can easily decide if a_i sent was $+\sqrt{E/2}$ or $-\sqrt{E/2}$.

Correspondingly, the imaginary path yields:

$$\begin{aligned} y_{\text{Im}} &= \int_t^{t+T} r(t)\sqrt{\frac{2}{T}}\sin(2\pi ft)\,dt \\ &= \frac{2}{T}\int_t^{t+T} a_i \cos(2\pi ft)\sin(2\pi ft)\,dt + \frac{2}{T}\int_t^{t+T} b_i \sin^2(2\pi ft)\,dt + \sqrt{\frac{2}{T}}\int_t^{t+T} n(t)\sin(2\pi ft)\,dt \\ &= a_i \frac{2}{T}(0) + b_i \frac{2}{T}\left(\frac{T}{2}\right) + \sqrt{\frac{2}{T}}\int_t^{t+T} n(t)\sin(2\pi ft)\,dt \\ &= b_i + \sqrt{\frac{2}{T}}\int_t^{t+T} n(t)\sin(2\pi ft)\,dt \\ &= b_i + n_{\text{Im}} \end{aligned} \tag{3.15}$$

where the first term b_i is the recovered imaginary part of the symbol, and the second term n_{Im} is the noise contribution along the imaginary axis.

On the imaginary path, the integrated result is also fed into the decision threshold. The decision threshold decides that b_i sent is $+\sqrt{E/2}$ if y_{Im} is greater than 0, or that b_i sent is $-\sqrt{E/2}$ if y_{Im} is less than 0. Again, if there is no noise, then the decision threshold can easily decide if b_i sent was $+\sqrt{E/2}$ or $-\sqrt{E/2}$.

Given a_i' and b_i' decided by the decision thresholds, the symbol mapper then outputs the detected symbols, specifically,

- If $a_i' = +\sqrt{E/2}$ and $b_i' = +\sqrt{E/2}$, then the transmitted bits are 00.
- If $a_i' = -\sqrt{E/2}$ and $b_i' = +\sqrt{E/2}$, then the transmitted bits are 01.
- If $a_i' = -\sqrt{E/2}$ and $b_i' = -\sqrt{E/2}$, then the transmitted bits are 11.
- If $a_i' = +\sqrt{E/2}$ and $b_i' = -\sqrt{E/2}$, then the transmitted bits are 10.

In this *maximum likelihood detector*, it is assumed that the probability of sending any one of the (four) symbols is identical.

The demodulation process again demonstrates why the quadrature component may be added without interfering with the in-phase component. This is because the two are orthogonal to each other; that is,

$$\int_0^{k/f} \cos(2\pi ft)\sin(2\pi ft)\,dt = 0 \quad \text{for } k = 0,1,2\ldots \tag{3.16}$$

For the above demodulation process to work, it is assumed that the demodulator is coherent, meaning that the phases of the carriers in the demodulator perfectly match those in the modulator.

Similarly, for the sake of brevity Figure 3.13 and similar diagrams are often depicted as that shown in Figure 3.14 in complex form. Here the demodulation

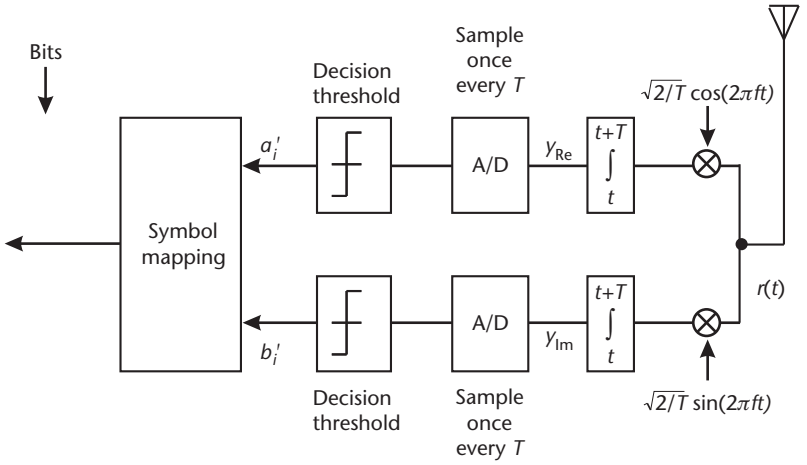


Figure 3.13 A demodulator using the matched filter approach. Note that a_i' and b_i' are the detected real and imaginary parts of the symbol, which may be different from the original a_i and b_i sent.

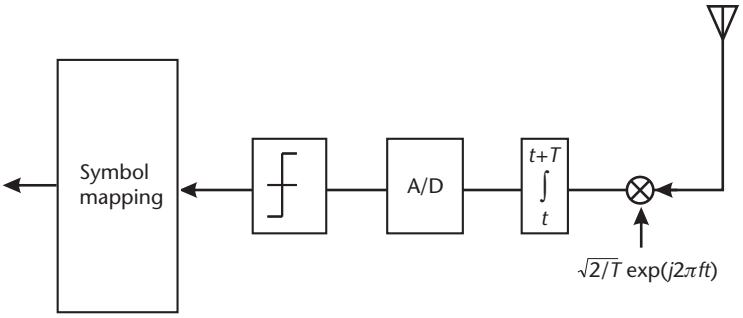


Figure 3.14 A complex equivalent of the demodulator using the matched filter approach.

process is shown simply as the complex multiplication by $\exp(j2\pi ft)$, which is the complex conjugate of the modulating carrier $\exp(-j2\pi ft)$.

3.5.2 Symbol Error

Let us consider the effect of noise for a moment. Figure 3.15 shows two transmissions, which both transmit the same symbol 00. In the first transmission, the symbol transmitted has the magnitude $\sqrt{E_1}$. In the second transmission, the symbol transmitted has the magnitude $\sqrt{E_2}$ such that $E_2 > E_1$.

Noise is multidimensional, but the matched filter extracts only the noise contribution along the real dimension (n_{Re}) and along the imaginary dimension (n_{Im}). In this example, the noise introduced by the channel has the same magnitude for both transmissions, and the magnitude of the noise is slightly larger than $\sqrt{E_1/2}$. As we can see, in the first transmission, the noise vector n_{Re} is large enough so that it could easily "push" the transmitted symbol into the upper left quadrant—the decision region for symbol 01. If so, the detector would interpret the received symbol to be 01, not 00, and a symbol error would occur. Similarly, the noise vector n_{Im} is also large enough so that it could easily "push" the transmitted symbol into the

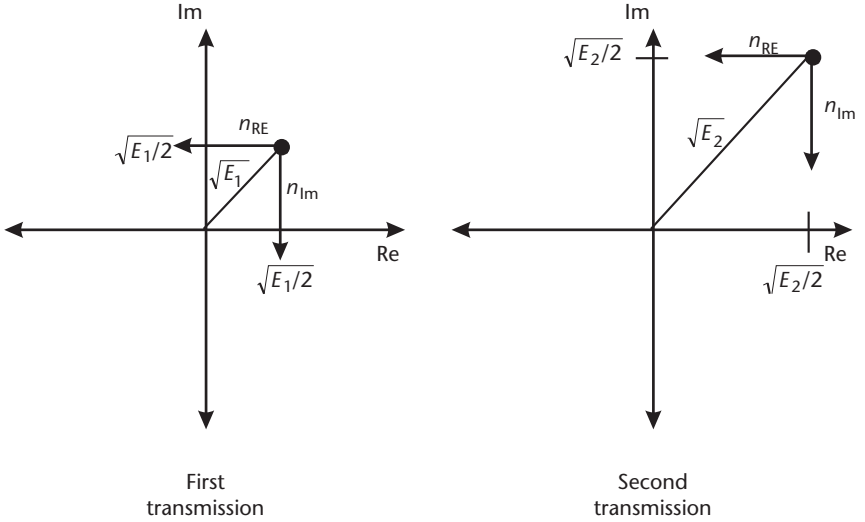


Figure 3.15 Two transmissions of the same QPSK symbol where $E_2 > E_1$.

lower right quadrant—the decision region for symbol 10. If that happens, the detector would interpret the received symbol to be 10, not 00. A symbol error would occur here, too.

In the second transmission, neither n_{Re} nor n_{Im} is large enough to push the received symbol into an adjacent decision region. In other words, both a_i and b_i stay below the thresholds, and the received symbol stays in the decision region for symbol 00. Thus, the detector interprets the received symbol to be 00, and there is no symbol error.

From Figure 3.15, readers can easily see that for a given magnitude of noise, the larger the symbol magnitude, the farther away the received symbol likely is from an adjacent decision region. This larger distance results in fewer symbol errors. Of course, the cost is that a larger symbol magnitude (hence signal power) is required.

If the SINR is high, then a received symbol would likely drift, at most, only into an adjacent decision region (due to noise). Because of Gray coding, an adjacent symbol is decoded as b bits that differ by 1 bit, and there is a single bit error as a result of a received symbol being misinterpreted as an adjacent symbol. Therefore, the relationship between the probability of symbol error P_e and the probability of bit error P_b can be approximated as [9]:

$$P_b \approx \frac{1}{b} P_e \tag{3.17}$$

if the SINR is high.

3.6 Adaptive Modulation and Coding

One important advantage of OFDMA is that each subcarrier can use the best possible combination of modulation and coding suitable to the fading experienced by that subcarrier. For example, if a subcarrier is experiencing small fade and high SINR, then that subcarrier can use a high order modulation (e.g., 64-QAM) and a high code rate (e.g., rate 3/4) to maximize bandwidth efficiency. If a subcarrier is experiencing large fade and low SINR, then that subcarrier can fall back to a low order modulation (e.g., QPSK) and a low code rate (e.g., rate 1/2) to maximize reliability but at the expense of bandwidth efficiency [10]. The process by which this dynamic change in modulation and coding is carried out is called adaptive modulation and coding (AMC).

The theoretical bandwidth efficiency e_B for a single carrier is

$$e_B = R \log_2 (M)(1 - P_P) \tag{3.18}$$

where R is the error-correcting code rate, M is the total number of symbols in the constellation (and modulation scheme), and P_P is the probability of packet error, so bandwidth efficiency is in bps/Hz. For a given symbol rate and bandwidth (hertz), a high bandwidth efficiency (bps/Hz) results in a high bit rate (bps). Conversely, a low bandwidth efficiency results in a low bit rate for a given symbol rate.

A higher order modulation such as $M = 64$ increases bandwidth efficiency. However, the SINR required by 64-QAM to maintain a respectable distance between symbols (hence an acceptable symbol error rate) is quite high. This is because 64-QAM has subtle amplitude and phase differences among its symbols, so 64-QAM needs a higher SINR to keep the symbols apart to minimize error. This is in contrast to QPSK whose symbols have more dramatic amplitude and phase differences, so QPSK does not require a high SINR to keep its symbols apart to minimize error. Therefore, high bandwidth efficiency, hence high bit rate, is only possible when the link SINR is high.

Figure 3.16 uses (3.18) to show that, through AMC, the bandwidth efficiency changes as a function of distance from the transmitter. When the receiver is close to the transmitter, the link SINR is high, so the transmitter can change to a high order modulation and high code rate to achieve high bandwidth efficiency. When the receiver is far from the transmitter, the link SINR is low, so the transmitter changes to a low order modulation and low code rate, resulting in a low bandwidth efficiency.

As the distance between transmitter and receiver increases and the SINR decreases, the transmitter can increase transmit power first, but stays with the current modulation and coding. If the SINR starts to drop, increasing transmit power can at first maintain the SINR at the required level for a given modulation and coding. However, if the SINR continues to drop, the transmitter may reach a maximum transmit power beyond which it can no longer increase power. At this time, the transmitter can switch to a lower order modulation and/or lower code rate. Power control for IEEE 802.16e is discussed in Chapter 8.

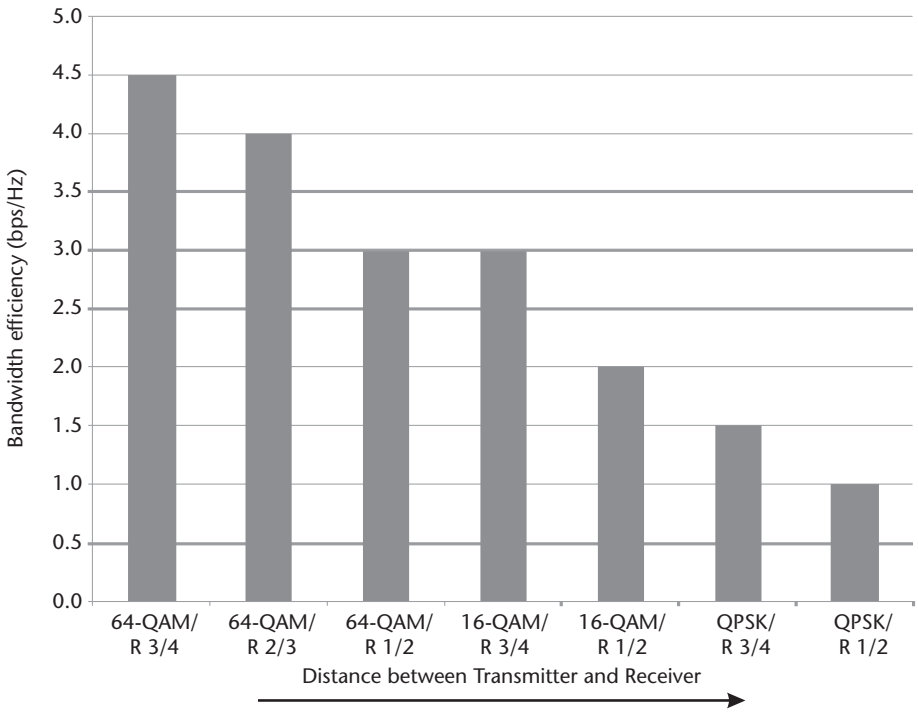


Figure 3.16 Bandwidth efficiency as a function of distance. P_P is assumed to be 0%.

In order for AMC to work, the transmitter has to know what the SINR is on the link so that it can change modulation and coding accordingly. This knowledge of the link SINR requires that the receiver sends some estimate of the link SINR back to the transmitter. With this knowledge of the SINR, the transmitter then determines the optimal modulation, coding, and transmit power to use. Of course, how often the transmitter needs this knowledge of the SINR depends on how quickly the channel changes. If the channel changes quickly as a function of time, then the transmitter would need the knowledge of the link SINR more often. Recall from Chapter 2 that coherence time T_c is the length of time over which the channel varies little:

$$T_c \approx \frac{c}{2fv} \tag{3.19}$$

where c is the speed of light, f is the frequency, and v is the velocity of the mobile. The smaller the T_c, the more often the channel changes.

3.7 Cyclic Redundancy Check (CRC)

Each packet generated by the MAC layer typically uses block coding to indicate the quality of the packet (containing a block of data bits). For example, the IEEE 802.16e system uses a *cyclic redundancy check* (CRC), which is one of the most common block codes. CRC is used to detect error in a packet but cannot correct error by itself. To correct error in a packet, the MAC layer has to request retransmission of the packet (in which CRC detected error).

For CRC, the information bits are treated as one long binary number. This number is divided by a unique *prime* number that is also binary, and the remainder is appended to the information bits as redundancy bits. When the packet is received, the receiver performs the same division using the same prime divisor and compares the calculated remainder with the remainder received in the packet [11]. If both remainders are identical, then the decision is that no error occurred. If they are different, then the decision is that some error occurred.

To demonstrate binary division, the (7, 4) Hamming code discussed previously can be generated using a prime divisor of (1, 0, 1, 1). The method can be more clearly seen if we represent binary bits (or a binary number) in a *polynomial* form. For example, the binary bits or number (1, 0, 1, 1) can be represented as a polynomial:

$$g(x) = x^3 + x + 1 \tag{3.20}$$

where each term in the polynomial corresponds to each 1-bit of the binary number. The polynomial $g(x)$ is a *prime polynomial*.

Suppose the message (1, 0, 1, 0) needs to be coded using the (7, 4) Hamming code. To do so, we first convert the message into its polynomial form, that is,

$$m(x) = x^3 + x \tag{3.21}$$

then we *shift* the message up by $(n-k)$ positions. This can be done very easily in the polynomial form by multiplying the message polynomial $m(x)$ by x^{n-k}. In this case $(n-k) = (7-4) = 3$, so we multiply $m(x)$ by x^3

$$x^3 m(x) = x^6 + x^4 \tag{3.22}$$

Note that this polynomial corresponds to (1, 0, 1, 0, 0, 0, 0).

The redundancy bits can be obtained by dividing $x^3 m(x)$ by $g(x)$, or

$$x^6 + x^4 = (x^3 + 1)(x^3 + x + 1) + (x + 1) \tag{3.23}$$

where $(x^6 + x^4)$ is $x^3 m(x)$, $(x^3 + 1)$ is the quotient, $(x^3 + x + 1)$ is the generator polynomial $g(x)$, and $(x + 1)$ is the remainder. The remainder polynomial $(x + 1)$ represents the redundancy bits to be appended to the message, that is, the redundancy bits are (0, 1, 1). As we can see in the (7, 4) Hamming code in Table 3.1, (0, 1, 1) are indeed the redundancy bits to be appended to the message (1, 0, 1, 0). See [12] for a good discussion on cyclic redundancy codes. References [3] and [6] give a good discussion on cyclic codes in general.

In an IEEE 802.16e system, CRC is used to check the bits in the header and payload portions of the MAC PDU (i.e., "packet"). Specifically, the OFDMA version uses CRC-32, which generates 32 redundancy bits. The generator polynomial used to generate the redundancy bits is [5]:

$$g(x) = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^8 + x^7 + x^5 + x^4 + x^2 + x + 1 \tag{3.24}$$

The way CRC is used to detect error is shown in Figure 3.17. This process takes place in the MAC layer. Using the procedure described earlier, the sender calculates the CRC (remainder) over the MAC header and MAC payload. If there is any encryption that is applied to the payload, then the CRC calculation is done over the header and encrypted payload.

The CRC is appended to the header and payload and all are sent over the medium. At the receiver, the receiver takes the received MAC PDU and passes it

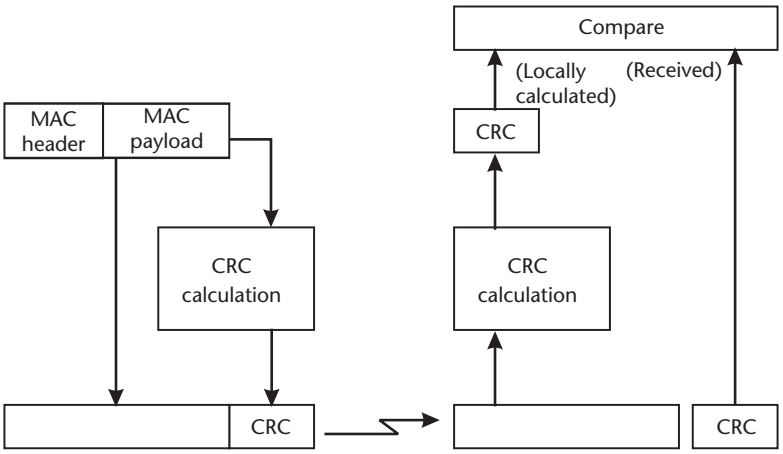


Figure 3.17 Packet error detection using CRC.

through the same CRC calculation used by the sender. The same CRC calculation produces a local CRC calculated at the receiver. Then the receiver compares the (locally) calculated CRC with the received CRC. If they are the same, then the decision is that the packet contains no error. If they are different, then the decision is that the packet contains error.

Because of the need for retransmission if a packet contains error, the concept of the probability of packet error P_P is more relevant at layer 2. If there are P bits in a packet, then the number of symbols in a packet is P/b. If one assumes that error is randomly distributed in a packet, then the probability of no error in a packet is:

$$1 - P_P = (1 - P_e)^{P/b} \tag{3.25}$$

Thus the probability of packet error is [9]:

$$P_P = 1 - (1 - P_e)^{P/b} \tag{3.26}$$

If there is error in the packet, then the receiver, or more specifically, layer 2 at the receiver would request retransmission of that packet.

3.8 Automatic Repeat Request (ARQ)

The process of requesting and receiving a retransmitted packet is called automatic repeat request (ARQ). There are two main variants of ARQ: stop-and-wait ARQ and sliding window ARQ.

3.8.1 Stop-and-Wait ARQ

In stop-and-wait ARQ, the sender sends a packet and then waits for an acknowledgment back from the receiver. The receiver may send either a positive acknowledgment (ACK) letting the sender know that the packet has been received correctly, or a negative acknowledgment (NAK) letting the sender know that the packet was received in error. Figure 3.18 illustrates the process of stop-and-wait ARQ. The sender sends a series of packets. In the figure, the first packet was received correctly, so the receiver sends an ACK to the sender. However, the second packet was received with error, so the receiver sends a NAK to the sender (and effectively requests a retransmission of the packet). The sender retransmits the packet, which was received correctly; the receiver sends an ACK to the sender. Then the sender continues with sending the next packet.

The sender has a timer and would start the timer after sending a packet. If a packet was lost and never reached the receiver, then the receiver can never acknowledge the packet. In this case, the sender waits and the timer would run out at the sender. After the timer runs out, the sender resends the same packet again.

If an acknowledgment (ACK or NAK) was lost and never reached to the sender, the sender cannot send the next packet because it does not know if the previous packet was correctly received. In this case, the sender still waits, and the timer would run out at the sender. After the timer runs out, the sender resends the same packet again.

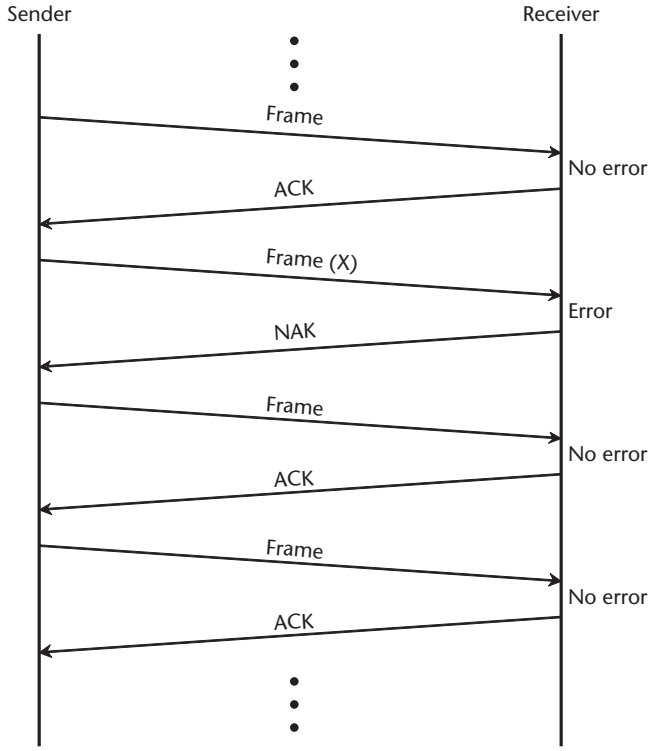


Figure 3.18 The stop-and-wait ARQ.

In short, with stop-and-wait ARQ, the sender cannot send the next packet without getting an acknowledgment (either ACK or NAK) from the receiver. In effect, the sender sends a packet, stops, and waits for an acknowledgment back from the receiver before sending another packet. If there is no acknowledgment after the timer runs out, then the sender resends the last packet again.

3.8.2 Sliding Window ARQ

In sliding window ARQ, the sender sends a packet but does not stop and wait for an acknowledgment back from the receiver. Instead, the sender sends the next packet right away. As such, each packet has to be numbered (with a *sequence number*) in order to keep track of the packets, and the acknowledgments have to be numbered as well in order to know to which packet an acknowledgment pertains.

In sliding window ARQ, the sender maintains a "window" W_{sender} in the unit of number of packets. This (sliding) window is effectively the number of packets that the sender is capable of buffering (because it may have to retransmit some of them later). The sender also keeps track of the maximum sequence number of a packet that may be sent $N_{max\ to\ send}$ and the sequence number of the last acknowledgment (ACK or NAK) that was received $N_{last\ ACK\text{-}NAK\ received}$. At all times, the following relationship must hold true at the sender:

$$W_{sender} \geq N_{max\,to\,send} - N_{last\,ACK-NAK\,received} \tag{3.27}$$

Figure 3.19 shows an example. In this case, W_{sender} is a system parameter that is set at 8. $N_{last\ ACK\text{-}NAK\ received} = 3$. This means that the sender can keep on sending packets with sequence numbers up to (8 + 3) or 11 (=$N_{max\ to\ send}$). The sender may not send a packet with sequence number of 12 because that packet would be outside the sliding window.

Similarly, the receiver maintains a window $W_{receiver}$ in the unit of number of packets. This (sliding) window is effectively the number of packets that the receiver is capable of buffering. The receiver keeps track of the sequence number of last packet that was received, $N_{last\ packet\ received}$, and the maximum sequence number of a packet that the receiver can accept, $N_{max\ to\ accept}$. At all times, the following relationship must hold true at the sender:

$$W_{receiver} \geq N_{max\ to\ accept} - N_{last\ packet\ received} \tag{3.28}$$

Figure 3.19 also shows an example at the receiver. In this case, $W_{receiver}$ is a system parameter that is set at 8. $N_{last\ packet\ received} = 9$. This means that the receiver can keep on accepting packets with sequence numbers up to (9 + 8) or 17. The receiver may not receive a packet with sequence number of 18 or greater because that packet would be outside the sliding window. Afterwards, the receiver may not receive a packet with a sequence number of 9 or less because that packet would also be outside the sliding window.

Given the constraints imposed by the sliding windows, the receiver sends acknowledgments for those packets received inside the (receiver) sliding window. Acknowledgments can be sent for specific packets, or an acknowledgment can be sent for the highest sequence number of a series of packets that has been received correctly.

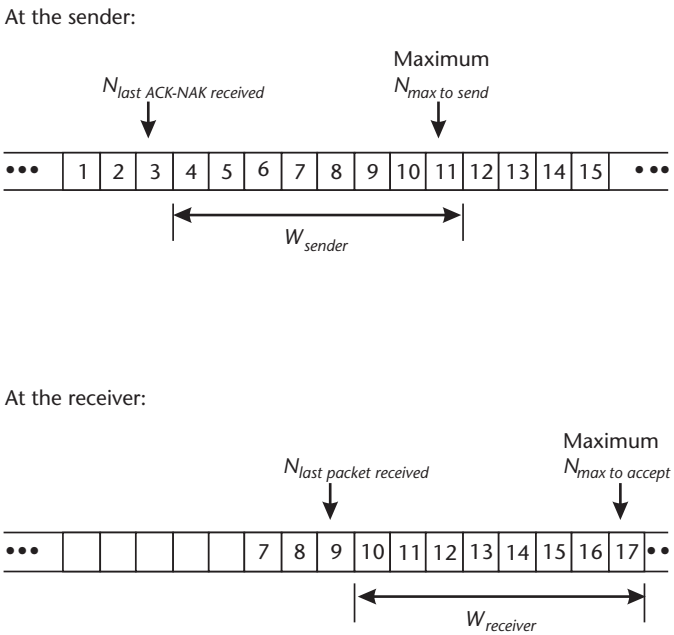


Figure 3.19 The sliding windows at the sender and the receiver.

In the sliding window ARQ, the sender also has a timer and would start the timer after sending a packet. If the timer runs out for that packet before any acknowledgment is received by the sender, the sender resends the same packet again.

Typically, the sender window W_{sender} is the same as the receiver window $W_{receiver}$. This setting means that the receiver is capable of buffering the amount of packets sent by the sender. Note that the stop-and-wait ARQ is really a sliding window ARQ that has a window of 1. See [13] for a good discussion of a sliding window.

3.9 Hybrid ARQ

Unlike ARQ, which is performed by the MAC layer alone, hybrid ARQ (HARQ) is performed by both the MAC layer and the physical layer. This *crosslayer* processing has become popular in recent years and represents a gradual shift toward more cooperation between protocol layers (as opposed to isolated processing within each layer) to improve performance. In traditional ARQ, if the MAC layer detects an error in a packet, it would discard the erroneous packet and request retransmission. However, potentially useful information contained in the erroneous packet is thrown away. In HARQ, if the MAC layer detects an error in a packet (i.e., CRC fails), it would still send a NAK to request retransmission, but when the newly retransmitted packet arrives, the physical layer would combine this retransmitted packet with the original packet to yield more information for FEC decoding. If the combined packet passes CRC, then the receiver sends an ACK and the HARQ process ends. If the combined packet still does not pass the CRC, then the receiver sends a NAK to request another retransmitted packet to be used by HARQ.

Figure 3.20 shows that the combining of packets takes place prior to FEC decoding, and FEC decoding operates on the bits in the combined packet. In fact, several subsequently retransmitted packets may be jointly processed with the original packet for decoding, but the system typically specifies a maximum number of retransmitted packets.

There are two types of HARQ: chase combining and incremental redundancy. In *chase combining*, the bits in the subsequently retransmitted packet(s) are FEC-coded the same way as those in the original packet. For example, if the bits in the

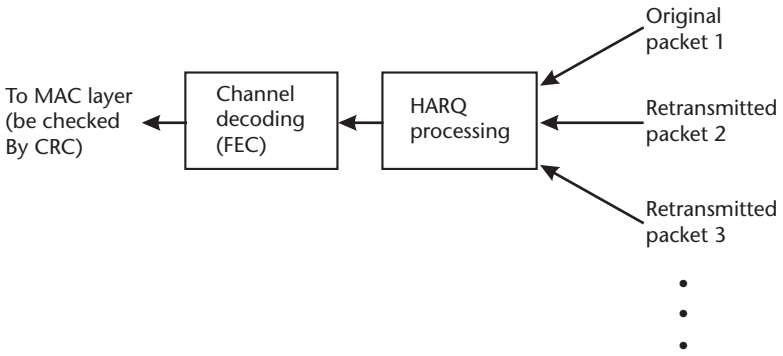


Figure 3.20 HARQ process. If an initially received packet fails CRC, then the receiver requests retransmission. The transmitter retransmits the packet. All packets are combined prior to FEC decoding.

original packet are coded using a rate 2/3 convolutional code, then the bits in any retransmission of the same packet are also coded using the same rate 2/3 convolutional code. The puncturing patterns used to generate the two rate 2/3 codes are the same. At the receiver, the original packet and the retransmitted packet(s) are simply combined in a bit-wise fashion to yield a combined packet, and the bits in the combined packet are decoded by FEC decoding.

In *incremental redundancy*, the bits in the subsequently retransmitted packet (or packets) are FEC-coded differently from those in the first packet. The idea is that the redundancy bits previously punctured out are "incrementally" sent with each retransmitted packet. Figure 3.21 illustrates the process. At the transmitter, the bits in the original packet undergo a specific puncturing pattern to have its bits selectively punctured to achieve the rate 3/4. In IEEE 802.16e, the puncturing pattern is characterized by the *subpacket ID* (SPID), and there are four puncturing patterns that can be applied in HARQ using incremental redundancy. For example, the bits in the original packet undergo a puncturing pattern (SPID = 0). If this packet fails CRC at the receiver, the bits in the retransmitted packet would undergo a different puncturing pattern (e.g., SPID = 2) to again achieve the rate 3/4. But the puncturing pattern is different this time so that bits previously punctured are now sent. At the receiver, the receiver then jointly combines the original packet and the retransmitted packet.

As one can see, HARQ is a form of time diversity that leverages information sent in a previous packet. In HARQ, because the sender has to wait for the receiver to finish processing the original packet plus the retransmitted packet(s) and to send an ACK or a NAK, the ARQ exchange necessarily has to be stop-and-wait ARQ.

Because HARQ can be set per mobile or per connection in each mobile, a system designer should not use HARQ for those mobiles that are experiencing good SINR. This is because there is not that much to be gained from HARQ at high SINR. This makes sense because at high SINR, a packet would most likely pass the

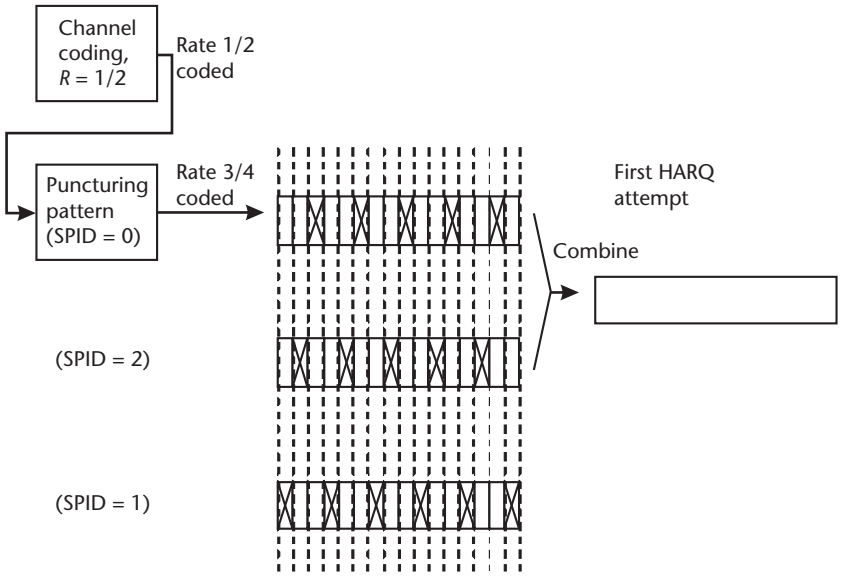


Figure 3.21 In incremental redundancy, bits previously punctured are sent in a retransmitted packet.

CRC upon the first transmission [14]. It has been shown that HARQ is suitable when used in conjunction with aggressive modulation (i.e., high M) and coding (i.e., high R) [15].

References

[1] Yang, S. C., *CDMA RF System Engineering*, Norwood, MA: Artech House, 1998.

[2] Sklar, B., "A Structured Overview of Digital Communications," *IEEE Communication Magazine*, August 1993.

[3] Blahut, R. E., *Theory and Practice of Error Control Codes*, Reading, MA: Addison-Wesley, 1984.

[4] Sklar, B., *Digital Communications: Fundamentals and Applications*, Englewood Cliffs, NJ: Prentice-Hall, 2001.

[5] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[6] Lin, S., and D. J. Costello, Jr., *Error Control Coding*, Englewood Cliffs, NJ: Prentice-Hall, 2004.

[7] TIA/EIA IS-95A, "Mobile Station-Base Station Compatibility Standard for Wideband Spread Spectrum Cellular Systems," Telecommunications Industry Association, May 1995.

[8] TIA/EIA/IS-2000.2-A, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, Telecommunications Industry Association, March 2000.

[9] Lee, E. A., and D. G. Messerschmitt, *Digital Communication*, Boston, MA: Kluwer Academic Publishers, 1988.

[10] Goldsmith, A. J., and S. -G. Chua, "Variable-Rate Variable-Power MQAM for Fading Channels," *IEEE Trans. on Communications*, Vol. 45, No. 10, 1997, pp. 1218–1230.

[11] Stallings, W., *Business Data Communications*, Upper Saddle River, NJ: Pearson Prentice-Hall, 2009.

[12] Stallings, W., *Data and Computer Communications*, Upper Saddle River, NJ: Prentice-Hall, 2006.

[13] Peterson, L. L., and B. S. Davie, *Computer Networks: A Systems Approach*, San Francisco, CA: Morgan Kaufmann, 2007.

[14] Andrews, J. G., A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*, Englewood Cliffs, NJ: Prentice-Hall, 2007.

[15] Ansari, A. Q., A. Rajput, and M. Hashmani, "WiMAX Network Optimization—Analyzing Effects of Adaptive Modulation and Coding Schemes Used in Conjunction with ARQ and HARQ," *7th Annual Communication Networks and Services Research Conference*, May 11–13, 2009, pp. 6–13.

Selected Bibliography

Haykin, S., *Communication Systems*, New York: John Wiley & Sons, 2009.

Jayant, N. S., and P. Noll, *Digital Coding of Waveforms,* Englewood Cliffs, NJ: Prentice-Hall, 1984.

Papoulis, A., *Probability, Random Variables and Stochastic Processes*, New York: McGraw-Hill, 2002.

Proakis, J. G., and M. Salehi, *Digital Communications*, New York: McGraw-Hill, 2008.

Simon, M. K., S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Englewood Cliffs, NJ: Prentice-Hall, 1995.
Viterbi, A. J., and J. K. Omura, *Principles of Digital Communication and Coding*, Mineola, NY: Dover Publications, 2009.
Wozencraft, J. M., and I. M. Jacobs, *Principles of Communication Engineering*, Prospect Heights, IL: Waveland Press, 1990.

CHAPTER 4

Fundamentals of OFDM and OFDMA: Transceiver Structure

4.1 Basic Transmitter Functions

Both Chapter 1 and Chapter 2 illustrate the theoretical advantages of OFDMA for a broadband wireless system in the terrestrial environment, but there are some implementation details that need to be addressed in order to make OFDM and OFDMA work. Care has to be taken in both the frequency domain and the time domain. Figure 4.1 depicts the simplified OFDM transmitter [1] shown in Chapter 1. In short, the baseband high-rate stream of data symbols at a rate of R_s goes into the serial-to-parallel converter, which assigns successive data symbols to K separate low-rate substreams. As a result, each low-rate substream has a rate of R_s/K. At the output of the serial-to-parallel converter, there is a set of K data symbols in parallel at any given time, so the serial-to-parallel converter arranges the baseband, serial data symbol stream into groups of K parallel data symbols.

The IDFT function transforms the set of K parallel data symbols from the frequency domain into the time domain. In OFDM, the system pretends that the data symbols originally exist in the frequency domain. That is why later at the receiver, the data symbols are recovered at the peaks of the (overlapping) sync functions in the frequency domain. In any case, the set of K transformed symbols in parallel then pass through the parallel-to-serial converter, which puts the K transformed symbols in series. A set of K transformed symbols in series is called an OFDM symbol, and the OFDM symbols at the output of the parallel-to-serial converter are running at a rate of R_s/K (OFDM symbols per second or blocks per second). Then the OFDM symbols are upconverted to produce the transmitted signal.

4.2 Time Domain: Guard Time

The simplified transmitter just presented has a problem. Figure 4.1 dictates that each OFDM symbol comes right after the previous OFDM symbol, without any guard time in between. Figure 4.2 shows this serial stream of OFDM symbols (each OFDM symbol contains K transformed symbols) with no guard time. Because of

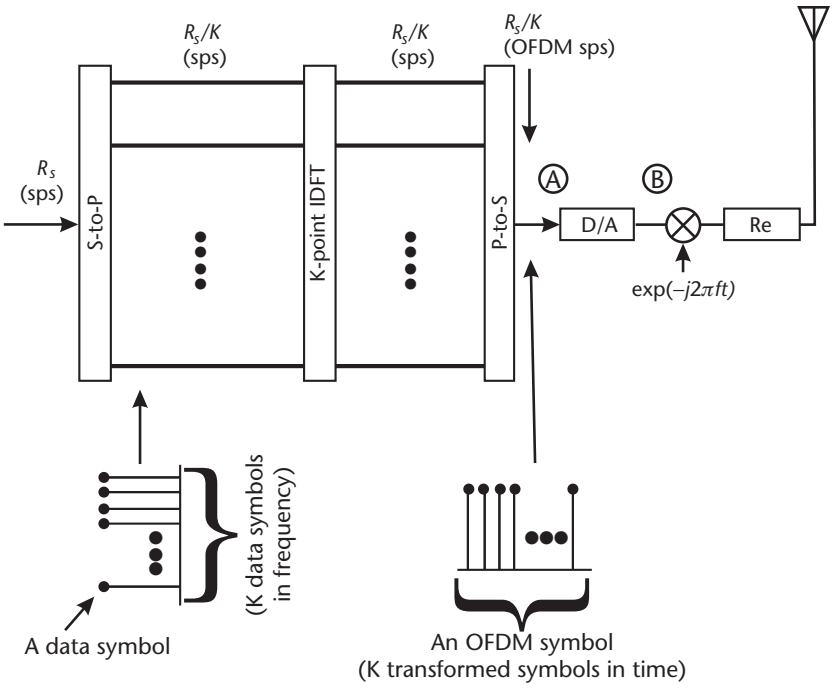


Figure 4.1 A simplified OFDM transmitter.

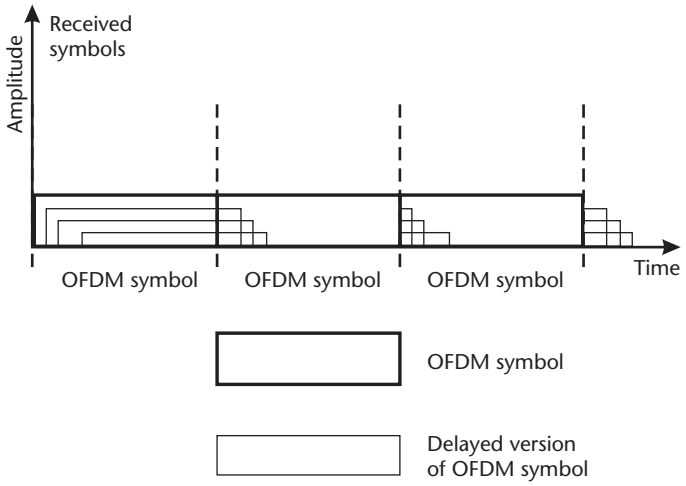


Figure 4.2 OFDM symbols without guard time and with IBI.

the multipath, delayed versions of an OFDM symbol can fall on the next OFDM symbol. As a result, there is inter-OFDM symbol interference (i.e., interblock interference, or IBI) between adjacent OFDM symbols.

An advantage of OFDM is that the data symbols in a single OFDM symbol do not interfere with one another inside an OFDM symbol. To state it in another way, the *K* data symbols do not affect one another within the symbol time of an OFDM symbol as far as data recovery at the receiver is concerned. This is because data are recovered at the peaks of the overlapping sync functions (in the *frequency* domain).

However, successive OFDM symbols (in the *time* domain) can still interfere with one another if they are not sufficiently separated in time.

To reduce such interference between adjacent OFDM symbols, one needs to add an extra guard time between adjacent OFDM symbols. In practice, extra symbols are inserted at the beginning of each OFDM symbol to add the guard time. Figure 4.3 shows the implementation. Here, g extra symbols are added right before the parallel-to-serial converter, so that the parallel-to-serial converter produces a total of $(K + g)$ symbols for each OFDM symbol.

Figure 4.4 shows that when adjacent OFDM symbols are separated by g symbols, the interference between OFDM symbols can be avoided. The guard symbols are also called the *cyclic prefix*. In practice, the cyclic prefix is generated by simply copying the last g transformed symbols in an OFDM symbol and repeating them at the front of the OFDM symbol. For example, if $K = 8$, $g = 2$, and the eight transformed symbols at the output of an 8-point IFFT are {A B C D E F G H}, then the cyclic prefix is {G H} and the OFDM symbol with the cyclic prefix appended is {G H A B C D E F G H}. If the guard time afforded by the cyclic prefix is larger than the delay spread, then interference between adjacent OFDM symbols can be eliminated.

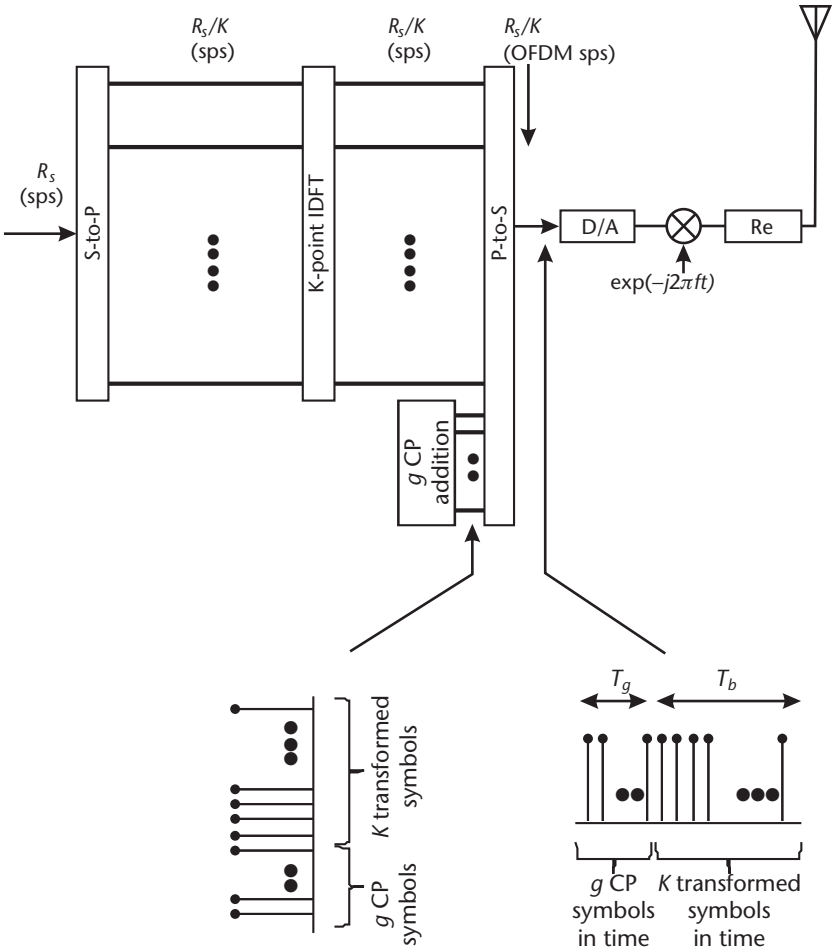


Figure 4.3 An OFDM transmitter. CP designates the cycle prefix.

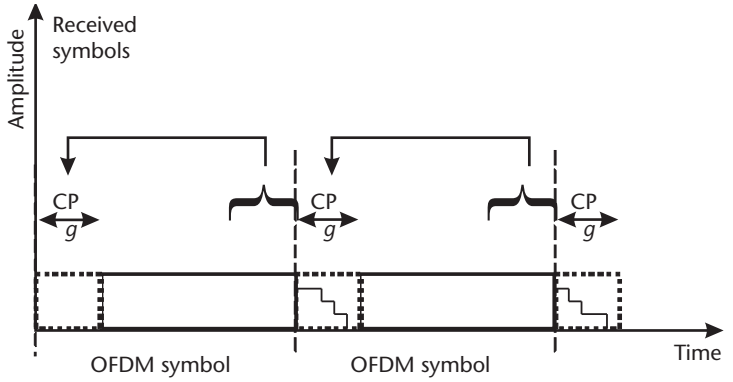


Figure 4.4 OFDM symbols with guard time. CP designates the cycle prefix.

Another advantage of adding a cyclic prefix is that it turns the channel operation from a *linear* convolution to a *circular* convolution, which can be easily implemented using DFT. See Section 4.5 for more details.

4.3 Frequency Domain: Synchronization

In OFDM, subcarriers overlap in the frequency domain, but data symbols can still be recovered at the receiver because they are sampled at the peaks of the sync functions (see Figure 4.5). Chapter 1 lists the advantages of such an arrangement of subcarriers in the frequency domain, including:

- Enhanced ability to combat ISI;
- Adaptive modulation and coding for each subcarrier;
- Simple equalization;
- Low-complexity modulation;
- Better spectral efficiency.

However, the disadvantage of this arrangement of subcarriers is that it is very sensitive to frequency offset. As can be seen in Figure 4.5, the peak of a subcarrier has to occur precisely at the zero-crossings of other subcarriers. Any offset would introduce interference from one subcarrier to where the peak of another subcarrier is and to where the data symbol is recovered.

One cause of frequency offset between the transmitter and the receiver is relative motion between them. Such a motion introduces a Doppler shift (see Chapter 2). Another cause of frequency offset is the mismatch of the transmitter and the receiver circuits. Some frequency offset will always be present. IEEE 802.16e attains frequency synchronization by using symbols that are known a priori. For example, on the downlink, the preamble containing known symbols is used to obtain frequency and timing synchronization; on the uplink, the ranging subchannels transmitting known symbols are used to obtain synchronization.

In addition to the preamble (downlink) and ranging (uplink), the OFDM system itself can be configured to minimize the effect of frequency offset. For a given

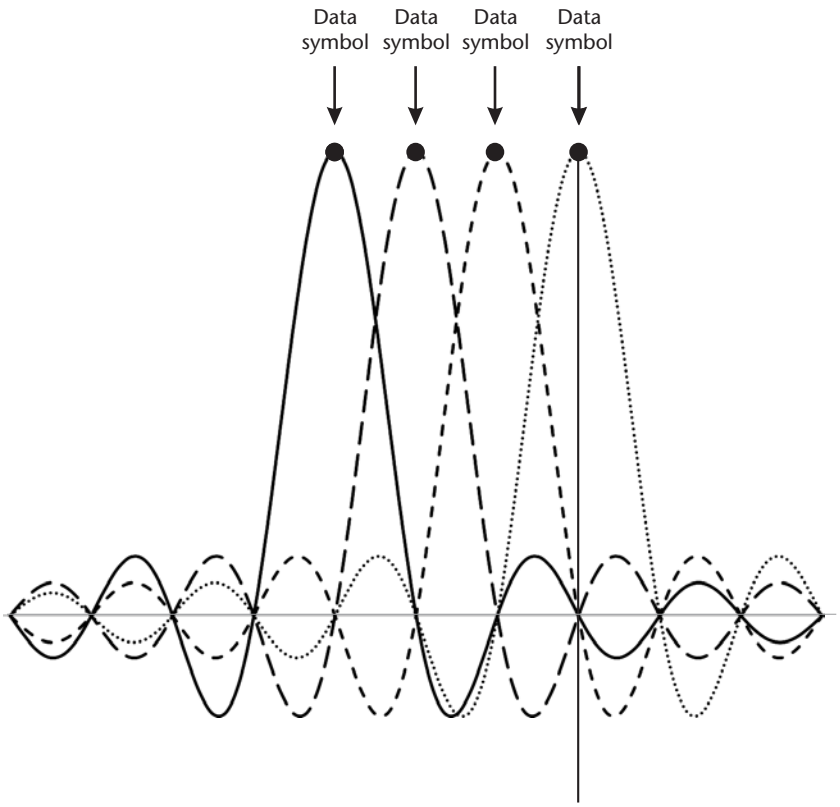


Figure 4.5 A set of OFDM subcarriers. The spectrum is shown over the duration of one OFDM symbol.

frequency offset (in hertz), a wider subcarrier bandwidth would help lessen the effect of frequency offset. This is because, given a fixed frequency offset (in hertz), the percent frequency offset (in %) decreases if a subcarrier becomes wider. One way to increase the subcarrier bandwidth is to decrease the number of subcarriers in a given band. However, this option may not be available if a network access provider has already chosen a technology to implement.

4.4 Basic Receiver Functions

A simple OFDM receiver is shown in Figure 4.6. After downconversion and a low-pass filter (LPF), the signal is now at the baseband but is still continuous in time. The analog-to-digital converter converts baseband continuous-time signals into baseband discrete-time symbols.

The serial-to-parallel converter assembles the incoming symbols into groups of OFDM symbols, each OFDM symbol consisting of K symbols and g cyclic prefix symbols. After throwing away the g cyclic prefix symbols, the remaining K symbols go into the DFT function, which transforms the K symbols in the time domain to K received data symbols in the frequency domain. The equalizer in each path corrects the data symbol carried by the corresponding subcarrier and removes the effects of the channel, and the detector in each path decides what data symbol was actually

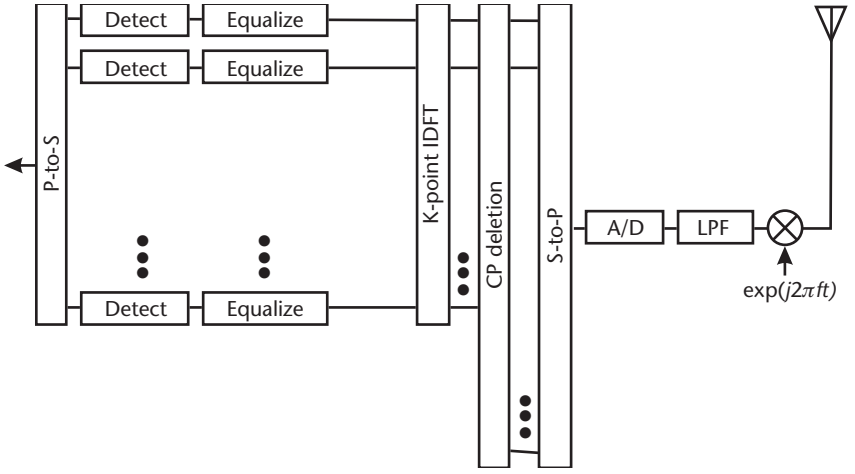


Figure 4.6 An OFDM receiver.

carried by the corresponding subcarrier. Afterwards, the parallel-to-serial converter rearranges the K parallel substreams of recovered data symbols into a single, high-rate stream of data symbols.

4.5 Equalization

The equalizers in Figure 4.6 are unique in OFDM and bear some explanation. We know that in the analog world (continuous in time and continuous in frequency) a linear communication system can be modeled by the diagram shown in Figure 4.7.

In the time domain, the system includes: input signal $x(t)$, impulse response of the channel $h(t)$, and output signal $y(t)$. It is well known that the output signal $y(t)$ is the convolution of the input signal $x(t)$ with the impulse response of the channel $h(t)$, that is,

$$y(t) = x(t) * h(t) \tag{4.1}$$

Given that convolution in time is multiplication in frequency, we have in the frequency domain:

$$Y(f) = X(f)H(f) \tag{4.2}$$

where $X(f)$ and $Y(f)$ are the Fourier transforms of $x(t)$ and $y(t)$, respectively; $H(f)$ is the Fourier transform of $h(t)$ and is also known as the transfer function.

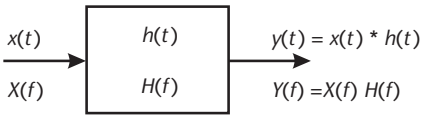


Figure 4.7 Channel operation turning input signal $x(t)$ into output signal $y(t)$.

Thus, at the receiver, a simple linear equalization used to recover the input (transmitted) signal in the frequency domain is dividing the output (received) signal by the transfer function, that is,

$$X(f) = \frac{Y(f)}{H(f)} \tag{4.3}$$

In the digital domain (discrete in time and discrete in frequency), the applicable Fourier transform to use is the discrete Fourier transform (DFT).[1] Specifically, the DFT converting from the discrete-time, time-domain signal x_n to the discrete, frequency-domain signal X_k is:

$$\mathrm{DFT}\{x_n\} = X_k = \sum_{n=0}^{K-1} x_n \left(e^{-j2\pi \frac{kn}{K}} \right) \tag{4.4}$$

The IDFT that converts from the discrete, frequency-domain signal X_k to the discrete-time, time-domain signal x_n is:

$$\mathrm{IDFT}\{X_k\} = x_n = \frac{1}{K} \sum_{k=0}^{K-1} X_k \left(e^{j2\pi \frac{kn}{K}} \right) \tag{4.5}$$

One advantage of using the DFT and the IDFT is that they can be efficiently calculated. In fact, the fast Fourier transform (FFT) and the inverse fast Fourier transform (IFFT) [2] are efficient implementations of the DFT and the IDFT and have enabled many new applications in digital signal processing. Another advantage of the DFT is that it is the only class of Fourier transform[2] that can be finitely parameterized [3].

Similarly, the convolution-multiplication property of the DFT states that if

$$y_n = x_n \circledast h_n \tag{4.6}$$

then

$$Y_k = X_k H_k \tag{4.7}$$

where $\circledast$ denotes circulation convolution. In other words, *circular convolution* of two signals in time is equivalent to multiplication of DFTs of two signals in frequency. y_n is the *circular convolution* of x_n and h_n and is operationally

1. Readers may recall another transform called discrete-time Fourier transform (DTFT), but DTFT is applied to signals that are continuous in frequency and discrete in time.

2. The four classes of Fourier transform are Fourier series (FS), Fourier transform (FT), discrete-time Fourier transform (DTFT), and discrete Fourier transform (DFT).

$$y_n = \sum_{m=0}^{K-1} x_{n-m \bmod K} h_m \tag{4.8}$$

Circular convolution is used because the convolution-multiplication property of the DFT requires that x_n is periodic with the period K.[3] In practice, x_n is made to look like a periodic sequence by adding the cyclic prefix. Recall that the cyclic prefix is generated by copying the last g transformed symbols in an OFDM symbol and repeating them at the front of the OFDM symbol. Doing so makes the $(K + g)$ data symbols look periodic, at least for the duration over which circular convolution is performed.

In DFT, being able to perform circular convolution is what makes the relationship $Y_k = X_k H_k$ true. Once this relationship is true (by adding the cyclic prefix), the effect of the channel is simply to multiply each original data symbol X_k by a complex number H_k [4], where H_k is the channel response at subcarrier k. Therefore, at the OFDM receiver, a simple linear equalization can be used to recover the input (transmitted) X_k in the frequency domain by just dividing the output (received) Y_k by the channel response H_k, that is,

$$X_k = \frac{Y_k}{H_k} \tag{4.9}$$

In actuality, the channel also introduces noise n_k; thus, (4.7) is rewritten as

$$Y_k = X_k H_k + n_k \tag{4.10}$$

and the equalized data symbol is:

$$\frac{Y_k}{H_k} = \frac{X_k H_k}{H_k} + \frac{n_k}{H_k} = X_k + \frac{n_k}{H_k} \tag{4.11}$$

In general, H_k (i.e., the channel) has to be known or at least estimated before the transmitted signal X_k can be recovered. In OFDM, each subcarrier k experiences its own channel response H_k, and the channel response may be different at different frequencies. This means that each subcarrier k requires its own estimated channel response H_k. That is why the OFDM receiver diagram shown previously has K separate equalizers, one for each subcarrier.

In OFDM, a number of the subcarriers are used as pilot subcarriers. Pilot subcarriers carry known signals, and the receiver can estimate the response of the channel based on what are actually received on the pilot subcarriers. Figure 4.8 shows an arrangement of pilot subcarriers and data subcarriers. Figure 4.8, as an example, shows that there are six data subcarriers between two pilot subcarriers. Typically, the system transmits pilot subcarriers at a higher power to ensure that channel estimates are reliable. Because the channel response may be different at

3. Strictly speaking, h_n should be periodic as well when applying the DFT, but to compute circular convolution, only x_n needs to be periodic. For more details, consult [3].

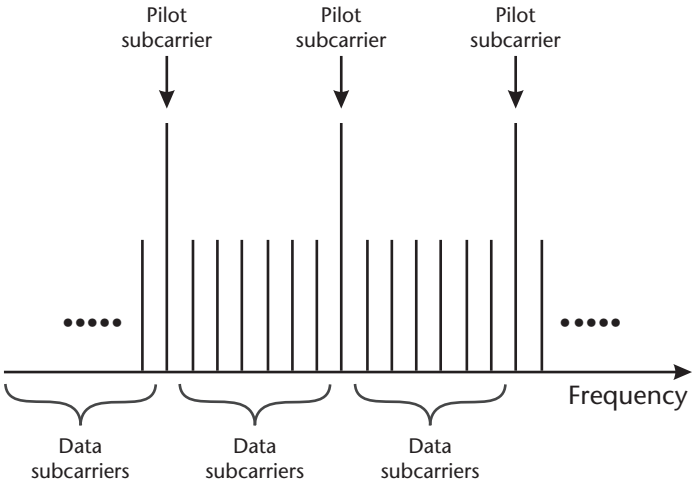


Figure 4.8 An example of arrangement of data and pilot subcarriers.

different frequencies, the actual response for a data subcarrier has to be interpolated based on measurements of the two nearest pilot subcarriers. In general, the more pilot subcarriers are provisioned, the more accurate the channel estimates are. However, the obvious tradeoff is that the more pilot subcarriers are provisioned, the fewer subcarriers are available to carry data.

4.6 OFDM Symbol

We are now ready to examine in more detail the subcarriers that make up the OFDM signal. Thus far, we have stated that the OFDM spectrum (over the duration of an OFDM symbol) consists of a group of overlapping subcarriers. In exploring the OFDM signal, we will work backwards through the transmitter chain and see how this spectrum of overlapping subcarriers is produced. Recall that a multicarrier signal can be produced conventionally by a series of complex multipliers shown in Figure 4.9.

This series of complex multipliers generates a multicarrier signal $x(t)$ that has the spectrum shown in Figure 4.10. The delta functions in frequency correspond to the complex sinusoids in time.

The ensemble of complex sinusoids shown in the figure can be characterized by a series of complex data symbols X_k carried by a series of complex subcarriers $\exp(j2\pi f_k t)$. In the complex baseband equivalent form, it is

$$x(t) = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{j2\pi f_k t} \tag{4.12}$$

where f_k is the center frequency of the kth subcarrier and K is the number of subcarriers.

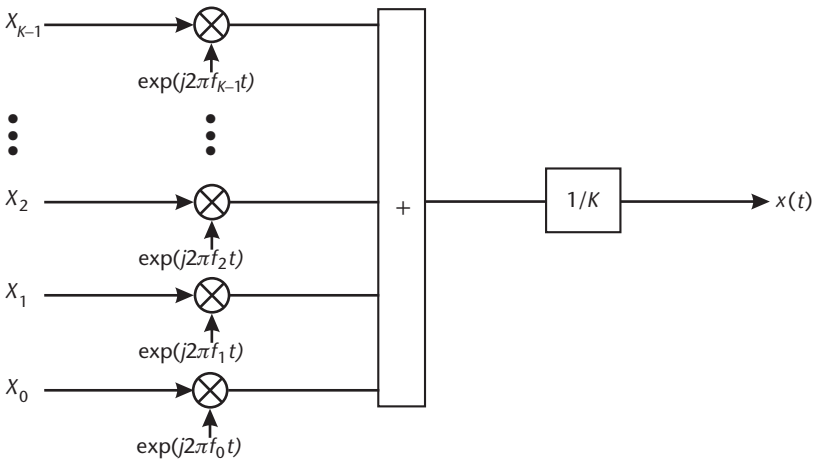


Figure 4.9 Generating a multicarrier signal using complex multipliers.

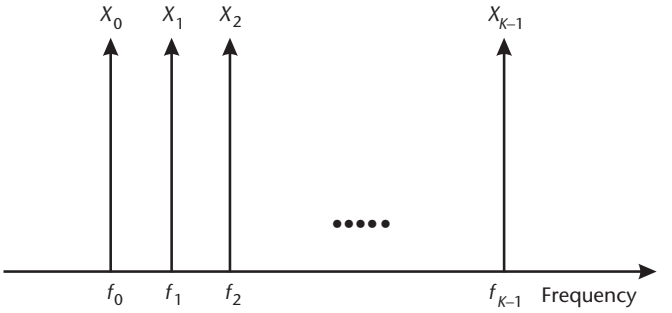


Figure 4.10 The spectrum of the multicarrier signal for all time.

The spectrum of the multicarrier signal consists of delta functions scaled by complex data symbols at f_k. The spectral lines constitute the multicarrier signal because the complex sinusoids $\exp(j2\pi f_k t)$ exist for all time.

Now, let us truncate the multicarrier signal in time so that it exists only for a limited duration T_{os}, that is,

$$x(t) = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{j2\pi f_k t} \quad 0 < t < T_{os} \tag{4.13}$$

(As seen later in this section, T_{os} is really the duration of an OFDM symbol.)

Limiting any signal to a range in time is equivalent to multiplying it by a rectangular function in time, and multiplication by a *rectangular* function in time is equivalent to convolution with a *sinc* function in frequency. If the rectangular function in time lasts T_{os} seconds, then its corresponding sinc function in frequency is $2/T_{os}$ wide (between the first two zeros).

Therefore, truncating the multicarrier signal in time results in a magnitude spectrum that is the convolution of a series of delta functions with a sinc function. Figure 4.11 shows that convolution of a series of delta functions with a sinc

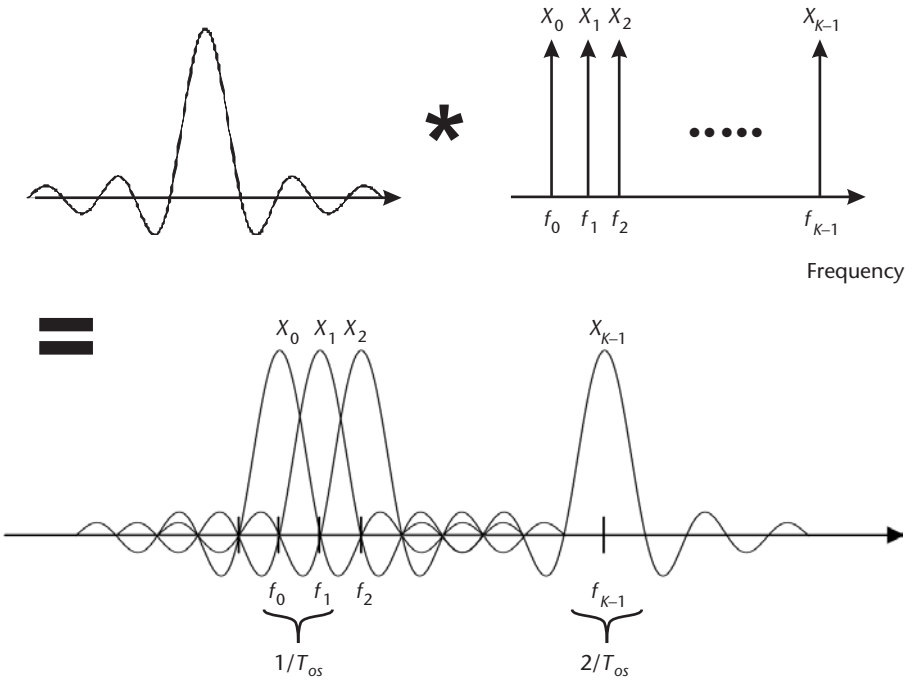


Figure 4.11 The magnitude spectrum of the multicarrier signal over the duration of T_{os}.

function results in copies of the sinc function duplicated at where the delta functions used to be.

If the sinc functions overlap, then according to Figure 4.11 the difference between the centers of two adjacent sinc functions is $1/T_{os}$, so $f_k = k/T_{os}$. Substituting k/T_{os} for f_k in the (4.13) yields:

$$x(t) = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{j2\pi\frac{k}{T_{os}}t} \quad 0 < t < T_{os} \tag{4.14}$$

This equation is the continuous-time (analog) version of the multicarrier signal. In other words, it is the signal found at position "B" immediately after the digital-to-analog converter in Figure 4.1.

To derive the discrete-time (digital) form of the multicarrier signal, one proceeds to sample $x(t)$ in time. Remember that $x(t)$ exists only between $t = 0$ and $t = T_{os}$. In the duration of T_{os} seconds, K equally spaced samples are taken in time, so the nth sample takes place at $t = (T_{os}/K)n$. Replacing t with $(T_{os}/K)n$ produces:

$$x\left(\frac{T_{os}}{K}n\right) = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{j2\pi\frac{k}{T_{os}}\left(\frac{T_{os}}{K}\right)n} = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{j2\pi\frac{kn}{K}} \quad 0 < t < T_{os} \tag{4.15}$$

which can be written as

$$x(n) = x_n = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{j2\pi\frac{kn}{K}} \quad 0 < t < T_{os} \tag{4.16}$$

because the argument of $x(\cdot)$ in discrete time is the sample number n itself.

Due to sampling, (4.16) is the discrete-time (digital) form of the multicarrier signal. In other words, it is the signal found at position "A" immediately before the digital-to-analog converter in Figure 4.1. More importantly, one can easily recognize now that (4.16) for x_n is simply the IDFT of X_k shown previously in (4.5). What this means is that x_n in time (within an OFDM symbol) can be easily generated by a K-point IDFT function, that is,

$$\text{IDFT}\{X_k\} = x_n = \frac{1}{K}\sum_{k=0}^{K-1} X_k \left(e^{j2\pi\frac{kn}{K}}\right) \quad 0 < t < T_{os} \tag{4.17}$$

In summary, (4.16) is the discrete-time (digital) form of the OFDM signal, over the duration T_{os}, produced by the K-point IDFT. Equation (4.14) is the continuous-time (analog) version of the OFDM signal, also over the duration T_{os}, after the digital-to-analog conversion. T_{os}, of course, is also known as the duration of the OFDM symbol.

There are three important points to remember regarding the generation of the OFDM signal:

- In the magnitude spectrum of the OFDM signal, the sync functions are present because the multicarrier signal is truncated in time [5].
- In the magnitude spectrum of the OFDM signal, adjacent sync functions overlap and are separated by $1/T_{os}$ peak-to-peak because the multicarrier signal is truncated in time and limited to a duration of T_{os}.
- Having adjacent sync functions separate by $1/T_{os}$ peak-to-peak also enables the discrete-time version of the OFDM signal to match the IDFT of X_k. See the derivation of (4.14) and (4.16). This match allows the generation of the OFDM signal using IDFT rather than using many complex multipliers.

Figure 4.12 illustrates an OFDM symbol in the time domain. The OFDM symbol lasts from $t = 0$ to $t = T_{os}$. In particular, Figure 4.12(a) shows that this OFDM symbol is made up of four data symbols (four subcarriers), and each subcarrier at a specific frequency is represented by a (truncated) cosinusoid with that frequency. The four subcarriers all have the same magnitude (e.g., 1, 1, 1, 1); thus, the four subcarriers all carry identical data symbols (e.g., 1, 1, 1, 1).

Two important observations can be made regarding this figure. First, in an OFDM symbol, each data symbol lasts the entire T_{os}. Second, these four subcarriers have frequencies $1/T_{os}$, $2/T_{os}$, $3/T_{os}$, and $4/T_{os}$ (at baseband); thus:

- The subcarrier with frequency $1/T_{os}$ completes one cycle in T_{os}.
- The subcarrier with frequency $2/T_{os}$ completes two cycles in T_{os}.
- The subcarrier with frequency $3/T_{os}$ completes three cycles in T_{os}.

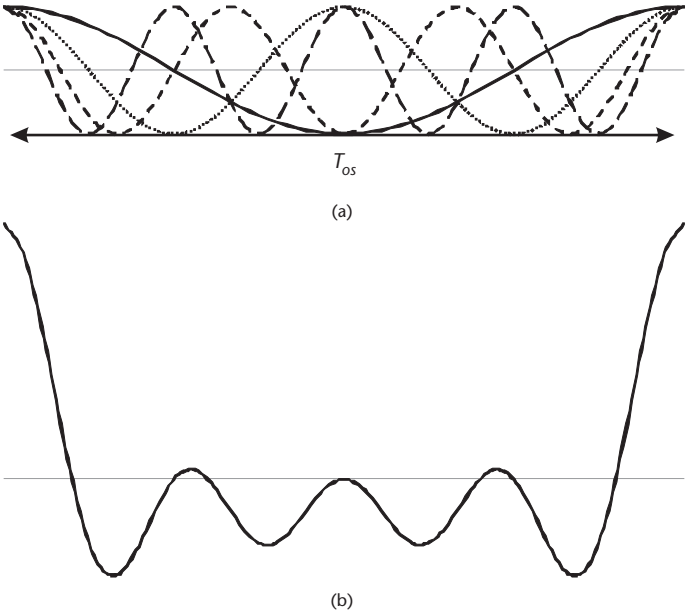


Figure 4.12 (a) An illustration of an OFDM symbol in the time domain over the duration of T_{os}. The OFDM symbol consists of four data symbols. (b) The actual superposition of four data symbols in time.

- The subcarrier with frequency $4/T_{os}$ completes four cycles in T_{os}.

In other words, a subcarrier always completes an integer number of cycles from $t = 0$ to $t = T_{os}$. Figure 4.12(b) depicts the actual superposition of four data symbols in time over the duration of T_{os}.

It is easy to recognize orthogonality among subcarriers in the frequency domain. In Figure 4.11, one can see that a data symbol X_k is recovered at the peak of the sync function, and the sync functions are arranged in frequency so that the peak of one sync function is at the zeros of all other sync functions. Because a data symbol X_k is recovered at the peak of the sync function, other sync functions do not interfere with X_k [6].

While it is straightforward to see orthogonality among subcarriers in the frequency domain, can one quantitatively show that the subcarriers (sync functions shifted by $1/T_{os}$) are orthogonal to each other and do not interfere with each other while in their analog form? To put it another way, can one be sure that the data symbols X_k carried by the subcarriers do not interfere with each other? The answer is yes, and such a proof can be more clearly shown in the time domain. To demonstrate the orthogonality among subcarriers in the time domain, we multiply the analog signal $x(t)$ by the complex conjugate of another subcarrier and integrate over the duration of an OFDM symbol ($0 < t < T_{os}$):

$$\int_0^{T_{os}} x(t) e^{-j2\pi \frac{l}{T_{os}}} dt \quad 0 < t < T_{os} \tag{4.18}$$

Note that the complex conjugate of this other subcarrier has an arbitrary frequency l/T_{os}. This integral is evaluated as follows:

$$\begin{aligned}
&\int_0^{T_{os}} x(t)e^{-j2\pi\frac{l}{T_{os}}t}\,dt = \frac{1}{K}\int_0^{T_{os}} e^{-2\pi\frac{l}{T_{os}}t}\sum_{k=0}^{K-1} X_k e^{j2\pi\frac{k}{T_{os}}t}\,dt \\
&= \frac{1}{K}\sum_{k=0}^{K-1} X_k \int_0^{T_{os}} e^{-j2\pi\frac{l}{T_{os}}t} e^{j2\pi\frac{k}{T_{os}}t}\,dt \\
&= \frac{1}{K}\sum_{k=0}^{K-1} X_k \int_0^{T_{os}} e^{j\frac{2\pi}{T_{os}}(k-l)t}\,dt \\
&= \frac{T_{os}}{K}X_k \quad \text{if } l = k \\
&\qquad 0 \quad \text{if } l \neq k
\end{aligned} \tag{4.19}$$

Thus, we see that:

- If the complex conjugate of a subcarrier has the same center frequency l/T_{os} as the center frequency k/T_{os} of a subcarrier carrying X_k, then the data symbol X_k is recovered.
- Data symbols carried by other subcarriers l ($\neq k$) do not interfere with X_k. In other words, subcarrier k is orthogonal to any other subcarrier l ($\neq k$).

The reason the above expression is 0 if $l \neq k$ is that a subcarrier (in an OFDM signal) always completes an integer number of cycles from $t = 0$ to $t = T_{os}$ as shown in Figure 4.12(a), and the integration of a sinusoid over an integer number of cycles is always 0.

In OFDM systems, the data symbols to be sent are X_k. Recall that OFDM pretends that the data symbols originally exist in the frequency domain. In the receiver, after the receiver receives x_n (plus noise and distortion) in time, it passes x_n (within an OFDM symbol) through the K-point DFT function to recover the original data symbols X_k (plus noise and distortion), that is,

$$\text{DFT}\{x_n\} = X_k = \sum_{n=0}^{K-1} x_n \left(e^{-j2\pi\frac{kn}{K}}\right) \quad 0 < t < T_{os} \tag{4.20}$$

4.7 OFDMA Transmitter

This section examines an actual OFDMA transmitter (and receiver) specified in the IEEE 802.16e [7] standard as a case. Figure 4.13(a) shows the basic structure of the transmitter. Note that it is similar to the transmitter shown in Figure 4.3 with some additions; the additions are shown boldfaced in Figure 4.13(a). The stream of information bits from the MAC layer are first fed into the data randomizer. The

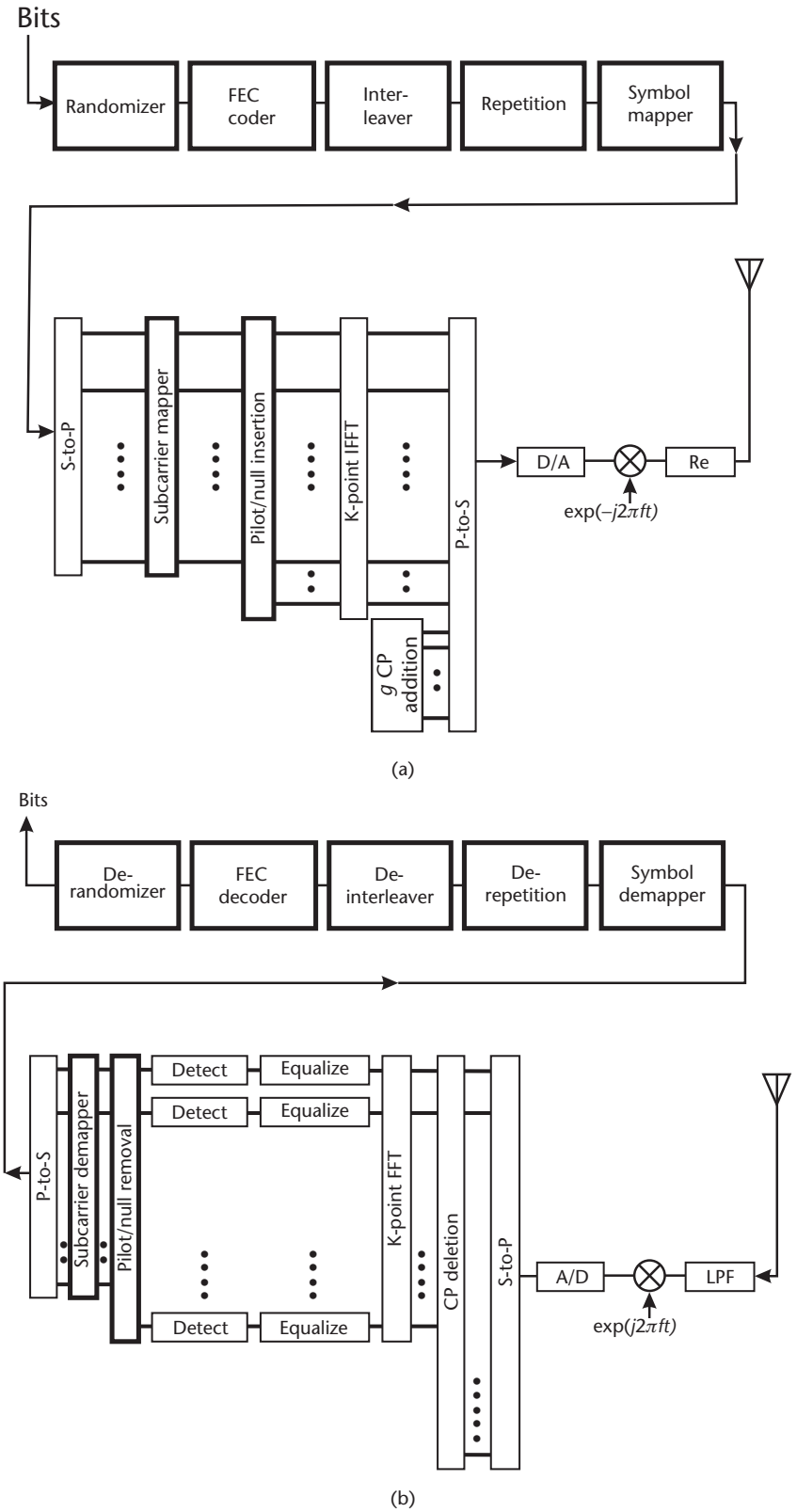


Figure 4.13 (a) Basic structure of the OFDMA transmitter in IEEE 802.16e. (b) Basic structure of the OFDMA receiver complementing the transmitter in IEEE 802.16e.

data *randomizer* XORs the data bits with bits produced by a shift register. The randomizer has three purposes:

- It scrambles the bits so that a casual eavesdropping receiver cannot easily intercept the data bits.
- It redistributes the bits to avoid long runs of 1s or 0s. A long run of 1s or 0s can cause a subcarrier to become unmodulated.
- It redistributes the bits to avoid long runs of 1s or 0s. A long run of 1s or 0s can cause the received bit stream (at the receiver) to lose synchronization. Bit-level synchronization requires a sufficient number of bit transitions (1-to-0 and 0-to-1) in a given time.

The randomizer only operates on the information bits and is present in both the uplink and the downlink.

After data randomization, the scrambled bits go into the forward error correction (FEC) function, which uses an error-correcting code to add redundancy bits for error correction. In addition to convolutional codes described in Chapter 3, the IEEE 802.16e standard specifies three additional FEC codes: block turbo code, convolutional turbo code, and low density parity check code. The convolutional code is mandatory in IEEE 802.16e.

After FEC, the coded bits are operated by the *interleaver*. The purpose of the interleaver is to ensure that the coded bits become sufficiently separated in frequency space and constellation space. In fact, the interleaver operates in two phases. The first phase is for frequency space. In the first phase, consecutive, coded bits are reordered to make sure that these bits are later mapped (by the subcarrier mapper) to nonadjacent subcarriers for frequency diversity. The second phase is for constellation space. In the second phase, consecutive, coded bits are reordered to make sure that these bits are later mapped fairly (by the symbol mapper) to more and less significant bits of the constellation.

After interleaving, the bits are organized into slots. At this point, the system can use the *repetition* function to further increase the reliability of the transmitted bits. The bits may be repeated by a repetition rate of 2, 4, or 6. Repetition provides a quick way for system designers to trade capacity for coverage. If a system designer would like to increase coverage, he or she can do so by increasing the repetition rate by a desired amount. This is because repetition decreases the required SNR at the receiver. However, the tradeoff is reduced capacity because available slots are occupied by repetitive bits [8].

The *symbol mapper* maps the (interleaved and/or repetitive) bits to data symbols based on the constellation used at the time (i.e., QPSK, 16-QAM, or 64-QAM). Note that each data symbol (later carried by a corresponding subcarrier) can be a QPSK, 16-QAM, or 64-QAM data symbol depending on the channel condition experienced by that subcarrier. If that subcarrier is experiencing high SINR, then it may carry a 64QAM data symbol to maximize bit rate. If that subcarrier is experiencing low SINR, then it may just carry a QPSK data symbol to ensure reliability.

The *serial-to-parallel converter* converts the serial stream of data symbols into parallel streams. Then the data symbols go into the *subcarrier mapper*, which assigns the individual data symbols to the individual subcarriers (i.e., assigning

a subcarrier index to each data symbol). The subcarrier mapper is necessary in OFDMA because different data symbols may have come from different users, and assigning data symbols to different subcarriers allows multiple users to access the air interface simultaneously.

The *pilot/null insertion* function inserts pilot symbols and null symbols. The pilot subcarriers are for channel estimation, and the null subcarriers include the guard subcarriers and the DC subcarrier. The null guard subcarriers and the null DC subcarrier have no power. This way, the (zero-power) guard subcarriers help contain the signal spectrum at the band edges, and the (zero-power) DC subcarrier introduces no DC component to the OFDM signal.

The input to the *K-point IFFT* consists a total of K parallel symbols, including data symbols, pilot symbols, and null symbols. The K-point IFFT transforms data, pilot, and null symbols from the frequency domain to the time domain. The K transformed symbols, along with g cyclic prefix symbols, go into the *parallel-to-serial converter*, which produces a serial output of transformed symbols in the time domain. At the output of the parallel-to-serial converter, the block of K transformed symbols constitutes an OFDM symbol, and the g CP symbols constitute the cyclic prefix.

The *digital-to-analog converter* changes the discrete-time symbols to analog signals, which are upconverted and transmitted over the air.

4.8 OFDMA Receiver

Figure 4.13(b) shows the basic structure of the receiver. At the receiver, the reverse of the process of Figure 4.13(a) takes place. The 802.16e standard specifies, in detail, what the transmitter does and contains. However, as has become the convention, the standards do not explicitly specify the architecture of the receiver. This is done to leave some details of the end-to-end implementation to the vendor, and many vendors differentiate themselves by how they implement their receivers. Figure 4.13(b) depicts one possible implementation of an OFDMA receiver arrived at by modifying the receiver shown in Figure 4.6.

After downconversion and analog-to-digital conversion, the received discrete-time symbols at baseband go into the serial-to-parallel converter, which converts the serial stream of received symbols into parallel streams. The g cyclic prefix symbols are removed, and the remaining K symbols go into the K-point FFT function. The K-point FFT function transforms the K symbols in the time domain to the K data (and pilot and null) symbols in the frequency domain.

The equalizer (in each path) takes the effects of the channel out of each received data symbol and corrects the received data symbol. Then the detector (in each path) estimates what the original data symbol is (see also Figure 4.6). The pilots are read and used for channel estimation, and then both the pilot symbols and the null symbols are removed. What remain are the recovered data symbols.

The subcarrier demapper rearranges the recovered data symbols in parallel back in the order by which users were originally assigned (to the individual subcarriers). In effect, the subcarrier demapper assigns the individual data symbols (recovered in the frequency domain) back to the individual users.

The parallel-to-serial converter rearranges the parallel substreams of recovered data symbols into a single, high-rate stream of data symbols. This high-rate stream of data symbols then go into the symbol demapper, which matches each data symbol in the stream to the bit pattern that data symbol represents. The resulting high-rate stream of bits then go through derepetition, deinterleaver, FEC decoder, and derandomizer, which constitute the inverse of the first four functions in the transmitter.

4.9 OFDMA

OFDMA is a method that assigns different users to groups of orthogonal subcarriers so they can access the air interface at the same time. The subcarrier mapper shown in Figure 4.13(a) is key in implementing OFDMA because it is the subcarrier mapper that assigns users to subcarriers. More accurately, it is the subcarrier mapper that assigns users' data symbols to subcarriers. Figure 4.14 and Figure 4.15 show a train of OFDM symbols along both time and frequency dimensions, and these figures compare and contrast OFDM and OFDMA in a situation where three users (A, B, and C) would like to access the air interface. There are a total of eight subcarriers, and the figures show how these users are assigned to subcarriers as a function of time.

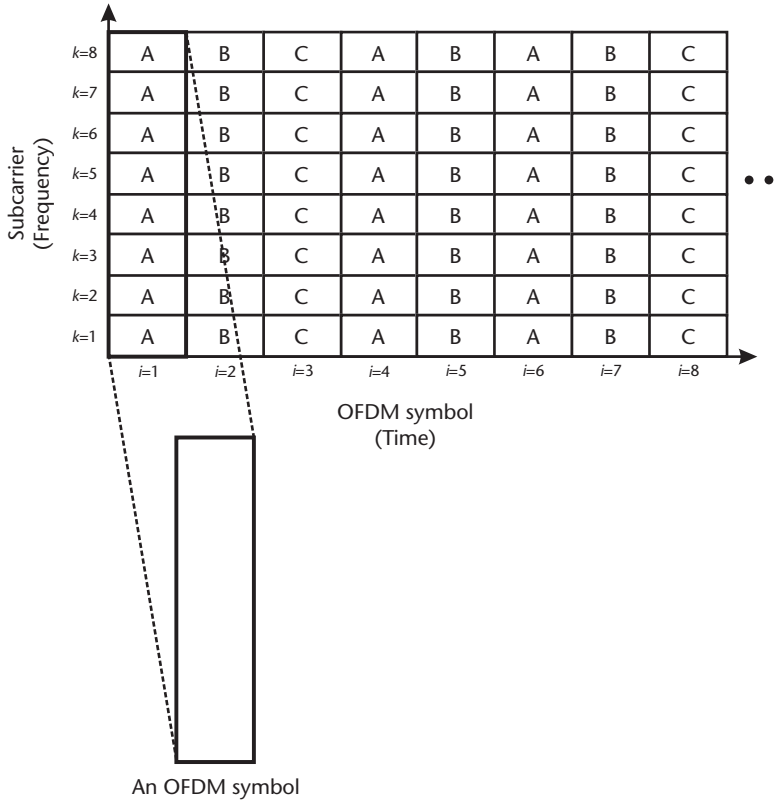


Figure 4.14 Multiple users using OFDM.

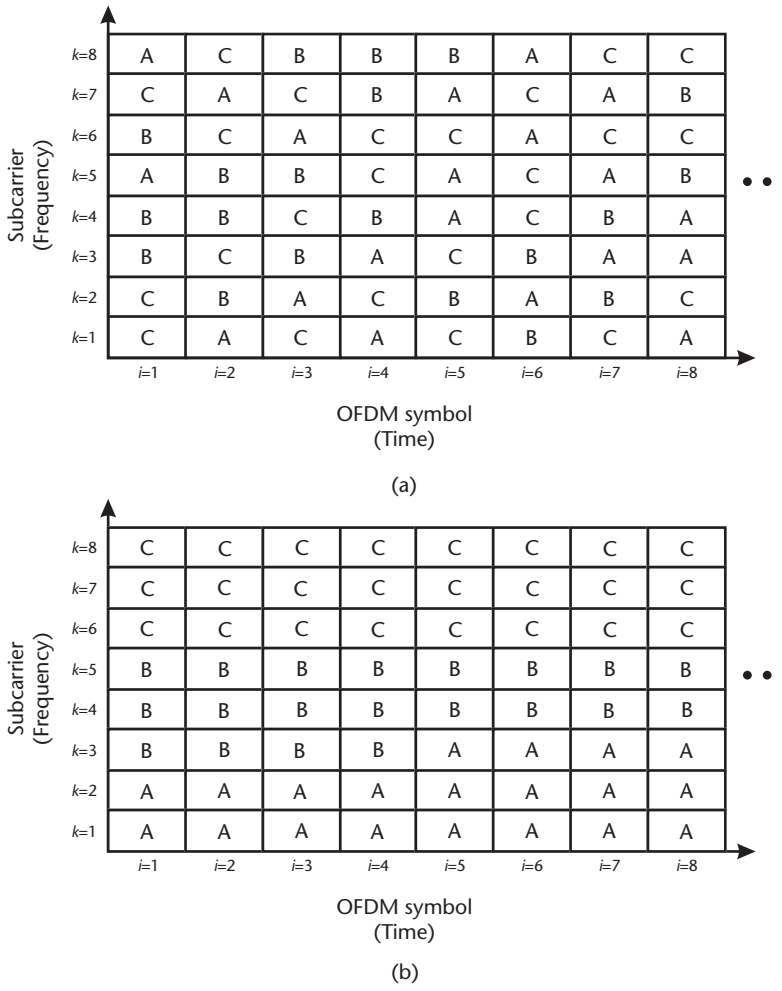


Figure 4.15 Multiple users using OFDMA: (a) distributed subcarriers and (b) contiguous subcarriers.

Figure 4.14 shows the situation in OFDM. In OFDM, the subcarrier mapper assigns all subcarriers to a single user at a time, so only one user can access the air interface at a time. In this case, each user is assigned an OFDM symbol in time, so users take turns to access the air interface. In effect, Figure 4.14 shows an OFDM/TDMA arrangement. Note that in this figure, user C is allocated less bandwidth than either user A or user B.

Figure 4.15 shows the situation in OFDMA. In OFDMA, the subcarrier mapper assigns different users to different subcarriers at a time. In general, there are two ways of assigning users to subcarriers: distributed subcarriers and contiguous subcarriers. A logical set of subcarriers is sometimes called a *subchannel*. (Each user may be assigned one or more subchannels.)

Figure 4.15(a) illustrates the arrangement of distributed subcarriers, where users are assigned pseudorandomly to subcarriers. For example, in the first OFDM symbol, user A is assigned two subcarriers ($k = 5, 8$), whereas user B is assigned three subcarriers ($k = 3, 4, 6$). Alternatively, in the first OFDM symbol, user A's two data symbols are carried by subcarrier $k = 5$ and subcarrier $k = 8$, whereas user B's

three data symbols are carried by subcarrier $k = 3$, subcarrier $k = 4$, and subcarrier $k = 6$. Distributing subcarriers pseudorandomly affords *frequency diversity* (to a single user).

Figure 4.15(b) illustrates the arrangement of contiguous subcarriers. In contiguous subcarriers, subcarriers are assigned to users in continuous groups. For example, in the first OFDM symbol, user A is assigned two subcarriers ($k = 1, 2$), user B is assigned three subcarriers ($k = 3, 4, 5$), and user C is assigned three subcarriers ($k = 6, 7, 8$). Contiguous subcarriers can take advantage of *multiuser diversity*.

IEEE 802.16e has certain ways of arranging users' data symbols in time and frequency, called permutation modes. The next chapter describes those ways in more detail.

4.9.1 Frequency Diversity

Frequency diversity is achieved by forming a subchannel through distributed subcarriers. For distributed subcarriers, the subcarrier mapper pseudorandomly distributes a user's subcarriers across the band. As far as a single user is concerned, such a distribution of subcarriers offers frequency diversity. This is so because if a user's subcarriers are distributed pseudorandomly, some of its subcarriers likely would not experience fades while some of its other subcarriers would.

Frequency diversity afforded by distributed subcarriers is well suited for *mobile* users because as a mobile user changes its location, the user experiences different multipath fadings at different locations. Figure 4.16 illustrates. For a user traveling from location X to location Y, that user experiences two different channel

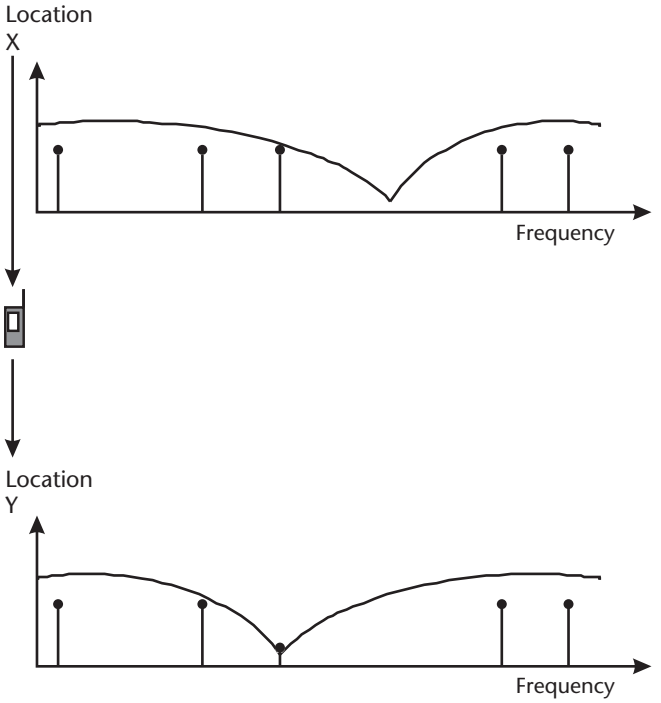


Figure 4.16 A mobile user travels from location X to location Y. At location X, none of the user's subcarriers is degraded by the channel response. At location Y, a subcarrier is experiencing deep fade.

responses at the two locations. If a user's subcarriers are distributed across the band, some of its subcarriers may avoid fading.

4.9.2 Multiuser Diversity

Multiuser diversity occurs when different users at different locations experience different channel responses. This form of diversity can be achieved by forming a subchannel through contiguous subcarriers. For contiguous subcarriers, a group of adjacent subcarriers are assigned to a single user. This scheme cannot take advantage of frequency diversity because all of the user's subcarriers are in the same vicinity of the spectrum. If a deep fade falls on top of those subcarriers belonging to a user, then that user will experience degraded channel. However, contiguous subcarriers can offer multiuser diversity.

Multiuser diversity afforded by contiguous subcarriers is well suited for *fixed* users because a fixed user's location does not change, so the channel response experienced by the user is relatively constant. This way, the system can assign a user a set of contiguous subcarriers based on the user's channel response. Figure 4.17 shows the case for two users. User A is fixed at location X, and user B is fixed at location Y. The two users experience two different channel responses because they are at two different locations. Therefore, the system can assign user A a set of contiguous subcarriers where user A is experiencing a good channel response, and it

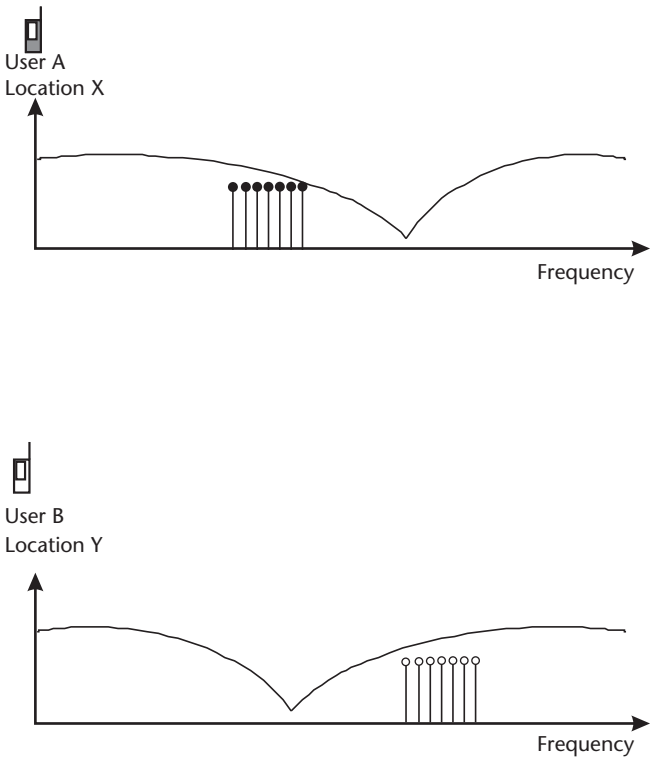


Figure 4.17 User A is at location X and user B is at location Y. At location X, user A's subcarriers are experiencing a good channel response (away from the null). At location Y, user B's subcarriers are also experiencing a good channel response (away from the null).

can assign user B a different set of contiguous subcarriers where user B has a good channel response.

Typically, the base station measures the channel response across the band by using pilot subcarriers that locate throughout the band. In OFDMA systems, it is possible to assign subcarriers based on their SINRs. For contiguous carriers, the base station can assign to a user a set of contiguous subcarriers that experience high SINR.

For a fixed user, contiguous subcarriers are especially helpful on the uplink in conserving the user's battery power. By transmitting on those subcarriers that experience strong SINR, the user device does not have to expend much power to attain a desired uplink bit rate.

4.9.3 Concluding Remarks

In dynamically assigning subcarriers to users, the subcarrier mapper performs an important *scheduler* function, which "schedules" different subcarriers to carry different users' data symbols. For example, in the first four OFDM symbols shown in Figure 4.15(b), user A is allocated less bandwidth than user B. However, in the second four OFDM symbols shown in the same figure, user A is allocated more bandwidth than user B. This change in granted bandwidth as a function of time is probably due to different users' bandwidth requests at different times subject to any quality-of-service (QoS) constraints. In fact, the scheduler makes optimizations decisions on the assignments of subcarriers and allocation of bandwidth resources based on multiple factors, including a user's current request for bandwidth, other users' pending (and competing) requests for bandwidth, users' QoS requirements, and channel quality experienced by each user.

Because OFDMA can dynamically allocate resources both in frequency (subcarriers) and in time (OFDM symbols), it is expected that broadband mobile systems will mostly use the more flexible OFDMA in the physical layer. In fact, the Mobile WiMAX System Profile specifies only the OFDMA implementation at the physical layer because OFDMA's scalable architecture is more suitable for mobile usage [9]. As such, this book focuses on OFDMA in subsequent chapters. After the reader becomes familiar with OFDMA, the OFDM implementation should be easily grasped.

4.10 Peak-to-Average Power Ratio

Figure 4.12(a) showed an example of an OFDM symbol made up of four subcarriers, and Figure 4.12(b) showed the actual superposition of these four subcarriers in one OFDM symbol. As seen in Figure 4.12(b), the peaks of the OFDM symbol in time are quite high. In this case, the peaks occur when all four subcarrier come in phase and add up together at the beginning and at the end of the OFDM symbol period. For an OFDM symbol, the continuous-time representation of an OFDM symbol is made up of its constituent subcarriers, each with its own frequency. As the subcarriers are superimposed and combined in time, high peaks can manifest themselves. The high peaks (in magnitude) can occur when the subcarriers come in phase and when the data symbols carried by the subcarriers are mostly positive

(or mostly negative). For the example shown in Figure 4.12(b), the two high peaks occur because the four data symbols carried by the four subcarriers are all identical (i.e., 1, 1, 1, 1) in the OFDM symbol.

This high peak-to-average power ratio (PAPR) is a typical problem with a signal made up of subcarriers of different frequencies (i.e., a multicarrier signal). If there are multiple subcarriers at multiple frequencies, they invariably come in phase (to form peaks and valleys) and out of phase (to form values close to zero) for different combinations of carried data symbols. Formally, the PAPR is defined as

$$\text{PAPR} = \frac{\max\left\{\left|x(t)\right|^2\right\}}{\text{avg}\left\{\left|x(t)\right|^2\right\}} \tag{4.21}$$

For example, the PAPR of the OFDM symbol shown in Figure 4.12(b) is 6.7, or 8.3 dB.

High PAPR is actually a nontrivial issue for OFDM-based systems because it decreases the efficiency of the power amplifier, and low efficiency of the power amplifier is a problem, especially in small mobile devices on the uplink. Techniques exist to deal with the PAPR issue in OFDM, including both signal scrambling schemes (e.g., block coding) and signal distortion schemes (e.g., clipping) [10]. An overview of the different PAPR-reduction techniques for OFDM can be found in [10, 11]. Incidentally, LTE has adopted single-carrier frequency-division multiple access (SC-FDMA) on the uplink because SC-FDMA has lower PAPR than multicarrier transmissions [12].

4.11 Conclusions

IEEE 802.16e specifies two implementations at the physical layer: WirelessMAN-OFDM and WirelessMAN-OFDMA. WirelessMAN-OFDM uses 256-point IFFT/FFT, whereas WirelessMAN-OFDMA can use up to a 2,048-point IFFT/FFT. Specifically, WirelessMAN-OFDMA's IFFT/FFT size can vary from 128 to 2,048 (i.e., 128, 512, 1,024, and 2,048). Because of the presence of the subcarrier mapper, the basic transmitter shown in the last section applies to both the OFDM option and the OFDMA option. For the OFDM option, the subcarrier mapper can basically be a straight-through device.

A broadband mobile channel over a metropolitan area or wide area introduces both delay spread and Doppler spread. Delay spread comes about because the system is operating in a terrestrial environment, and the longer transmission paths in metro and wide areas (as compared to local areas) mean that delay spread is non-negligible. Doppler spread comes from the fact that, in a mobile channel, there is relative motion between the transmitter and the receiver. These impairments cause the channel to be both frequency selective and time varying.

OFDM deals with the effects of a frequency selective channel by dividing a high-speed symbol stream into many low-speed symbol streams and narrow subcarriers. As a result, each subcarrier experiences little ISI because the data symbol time of a subcarrier is now much greater than the channel delay spread. Equivalently, the data symbol rate (and hence bandwidth) of a subcarrier is now much

less than the channel coherence bandwidth W_c. However, by now readers probably detected a tradeoff. If the subcarrier bandwidth decreases, then the effect of the Doppler shift f_D as compared to the bandwidth becomes more prominent, and a non-negligible f_D causes subcarriers to interfere with one another. Needless to say, increasing the number of subcarriers in a fixed band necessarily decreases the subcarrier bandwidth, and narrower subcarrier bandwidth is more advantageous in environments with large delay spreads.

IEEE 802.16e fixes the subcarrier spacing at 10.9375 kHz (in frequency), which does not change regardless of the bandwidth of the RF channel. This way, if the bandwidth of the RF channel increases or decreases (e.g., due to a country's specific spectrum regulation), then the number of subcarriers can increase or decrease proportionally. Fixing the subcarrier spacing fixes the atomic unit of physical-layer resource in frequency. This way, scaling bandwidth has a minimal impact to higher layers [13]. In addition, considerable cost savings result because one does not have to design a brand-new physical layer for every possible RF channel bandwidth. This scheme is also called scalable OFDMA (SOFDMA) [14].

Having gone through the fundamental workings of OFDM and OFDMA, we are now ready to go into other details of the IEEE 802.16e implementation of OFDMA. The next chapter investigates how its physical layer is structured in time and frequency. Unless otherwise noted, subsequent chapters will focus on the more advanced attributes of the IEEE 802.16e standard.

References

[1] Haykin, S., and M. Moher, *Modern Wireless Communications*, Upper Saddle River, NJ: Pearson Prentice Hall, 2005.

[2] Cooley, J. W., and N. N. Turkey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Math. Comput.*, Vol. 19, No. 2, April 1965, pp. 297–301.

[3] Roberts, R. A., and C. T. Mullis, *Digital Signal Processing*, Reading, MA: Addison Wesley, 1987.

[4] Litwin, L., "OFDM: An Introduction to Multicarrier Modulation," *IEEE Potential*, April/May 2000, pp. 36–38.

[5] Weinstein, S. B., and P. M. Ebert, "Data Transmission by Frequency-Division Multiplexing Using the Discrete Fourier Transform," *IEEE Trans. on Communication Technology*, Vol. 19, No. 5, 1971, pp. 628–634.

[6] Van Nee, R., and R. Prasad, *OFDM for Wireless Multimedia Communications*, Norwood, MA: Artech House, 2000.

[7] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[8] Upase, B., M. Hunukumbure, and S. Vadgama, "Radio Network Dimensioning and Planning for WiMAX Networks," *Fujitsu Scientific and Technical Journal*, Vol. 43, No. 4, 2007, pp. 435–450.

[9] Nakamura, M., T. Chujo, and T. Saito, "Standardization Activities for Mobile WiMAX," *Fujitsu Scientific and Technical Journal*, Vol. 44, No. 3, 2008, pp. 285–291.

[10] Jiang, T., and Y. Wu, "An Overview: Peak-to-Average Power Ratio Reduction Techniques for OFDM Signals," *IEEE Trans. on Broadcasting*, Vol. 54, No. 2, 2008, pp. 257–268.

[11] Han, S. H., and J. H. Lee, "An Overview of Peak-to-Average Power Ratio Reduction Techniques for Multicarrier Transmission," *IEEE Personal Communications*, Vol. 12, No. 2, 2005, pp. 56–65.

[12] Astely, D., et al., "LTE: The Evolution of Mobile Broadband," *IEEE Communications*, Vol. 47, No. 4, 2009, pp. 44–51.

[13] Yin, H., and S. Alamouti, "OFDMA: A Broadband Wireless Access Technology," *IEEE Sarnoff Symposium on Advances in Wired and Wireless Communication*, Piscataway, NJ, March 2006, pp. 1–4.

[14] Jain, R., C. So-In, and A. -K. Al Tamimi, "System-Level Modeling of IEEE 802.16e Mobile WiMAX Networks: Key Issues," *IEEE Wireless Communications*, Vol. 15, No. 5, 2008, pp. 73–79.

Selected Bibliography

Ergen, M., *Mobile Broadband Including WiMAX and LTE*, New York: Springer, 2009.

CHAPTER 5

Physical Layer: Time and Frequency

5.1 Introduction

In a layered protocol architecture, the physical layer is responsible for the actual transmission of bits across the medium. In doing so, the physical layer at the transmitter accepts *bits* from the MAC layer above and converts these bits into physical *waveforms* (i.e., symbols). The physical layer then injects the waveforms into the medium. At the receiver, the physical layer intercepts the waveforms, reinterprets them into bits, then delivers the bits to the MAC layer above. See Figure 5.1.

This chapter examines the physical layer in more detail. Whereas Chapters 3 and 4 describe how raw data symbols are transmitted and received over the air interface and the principles of OFDM and OFDMA, this chapter looks at how an actual system (i.e., IEEE 802.16e) organizes the data symbols logically to carry user and control information.

The standard supports both time division duplex (TDD) and frequency division duplex (FDD). In *single-band* operation, TDD uses the same RF band for both downlink and uplink transmissions. So in TDD, the base station and the mobile take turns in time to transmit in the same RF band (see Figure 5.2). Using TDD has three advantages:

- The system can leverage channel reciprocity because both the downlink and the uplink transmit in the same frequency band.
- Resources on the downlink and the uplink can be proportioned dynamically in time, thus realizing asymmetric bit rates on the downlink and the uplink; asymmetric bit rates are typical of a Web traffic profile.
- The complexity and cost of the mobile can be lower. Because the mobile does not have to transmit and receive at the same time in TDD, there is no duplexer and no duplexer loss in the mobile [1].

A disadvantage of TDD is that temporal resources in the same band have to be shared between the downlink and the uplink, so timing synchronization is critical.

In *paired-band* operation, FDD can be used where one band is dedicated to the downlink and another band is dedicated to the uplink (see Figure 5.3). An advantage of FDD is that the downlink can transmit continuously in its own band and

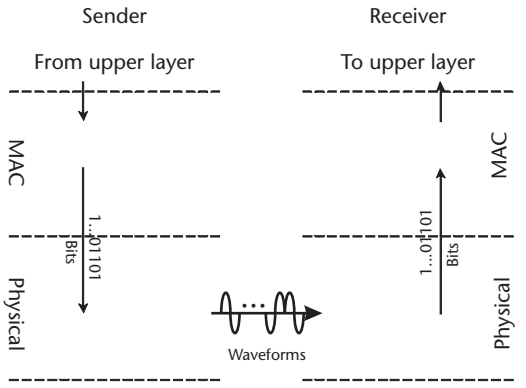


Figure 5.1 The relationship between the physical layer and the MAC (sub)layer.

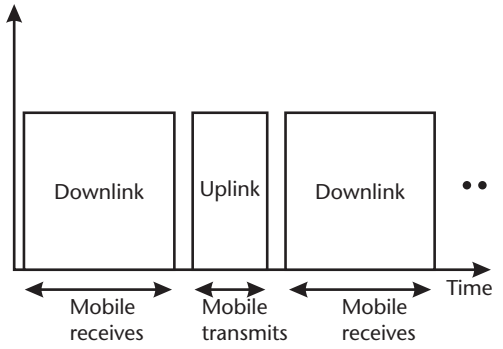


Figure 5.2 Time division duplex (TDD).

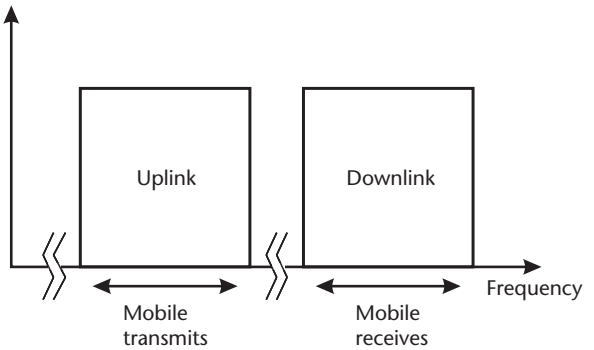


Figure 5.3 Frequency division duplex (FDD).

the uplink can transmit continuously in its own band. The disadvantage, of course, is the inability to leverage channel reciprocity. In addition, IEEE 802.16e supports half duplex FDD (H-FDD). In H-FDD, one band is for the downlink transmission and another band is for the uplink transmission, but at any given time, the mobile can only transmit or receive, not both (hence half duplex). H-FDD can be used in countries that mandate paired-band operation. Because an H-FDD mobile does not transmit and receive simultaneously, its cost can be lower.

Both TDD and FDD can be used with OFDMA. It is expected that network access providers would predominantly use the OFDMA option for flexible bandwidth management and performance optimization. Also, TDD is expected to be popular because of the importance of channel reciprocity and the ability to divide allocations in time between the downlink and the uplink. Thus, this chapter focuses on TDD and the OFDMA option of IEEE 802.16e and is based on the system specifications stated in the standard [2, 3].

Figure 5.4 illustrates that logical units of data are organized in a hierarchical manner at the physical layer. At the lowest level, *subcarriers* (that carry data symbols) are transmitted and received over the air. Chapter 4 discusses OFDMA subcarriers. At the next level up, subcarriers are organized into *subchannels* by using one of the subcarrier permutation modes.

At the next higher level, subchannels (in frequency) and OFDM symbols (in time) are organized into *slots*. Slots are important because a slot is the smallest unit of physical-layer resource allotted to a user. The ability to allocate one or more slots to a particular user is important because a user can receive some minimum allocation of bandwidth resources, and the base station can then change the allocation based on the bandwidth requirements of a particular user.

The following are some relevant parameters:

- N_{FFT} is the total number of subcarriers (= K in previous chapters).
- N_{used} is the number of used subcarriers, including data subcarriers, pilot subcarriers, and the DC subcarrier. Thus, ($N_{FFT} - N_{used}$) is the number of guard subcarriers.
- $N_{subcarriers}$ is the number of subcarriers per subchannel.
- $N_{subchannels}$ is the number of subchannels.

In particular, *data subcarriers* carry discrete data symbols for data transmission. *Pilot subcarriers* carry continuous signals for channel estimation. The *DC subcarrier* is a null subcarrier that has no power and is at the middle of the channel band. *Guard subcarriers* are also null subcarriers that have no power but are at the edges of the channel band. This way, guard subcarriers help contain the signal spectrum at band edges.

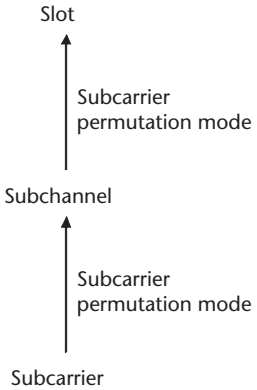


Figure 5.4 Hierarchy of data organization in IEEE 802.16e.

Because IEEE 802.16e uses scalable OFDMA (SOFDMA), the standard fixes the subcarrier separation at 10.9375 kHz. This way, the total number of subcarriers is a function of the channel bandwidth. For example, if the channel bandwidth is 5 MHz, then N_{FFT} = 512; if the channel bandwidth is 10 MHz, then N_{FFT} = 1,024.

Starting in the next section, we examine how subcarriers are organized to form subchannels using one of the subcarrier permutation modes.

5.2 Distributed Subcarrier Permutation: Forming Subchannels on Downlink

On the downlink, there are two required subcarrier permutation modes: full usage of subchannels (FUSC) and partial usage of subchannels (PUSC). These two modes were first specified in the original IEEE 802.16-2004 standard. In addition, the IEEE 802.16e standard specified two optional subcarrier permutation modes that are important to system design and optimization: tile usage of subchannels 1 (TUSC1) and tile usage of subchannels 2 (TUSC2). All these modes use what is called *distributed subcarrier permutation* because these modes form a subchannel by using nonadjacent subcarriers that are pseudorandomly scattered throughout the frequency band. As such, frequency diversity can be achieved. These modes of distributed subcarrier permutation primarily differ in how data subcarriers are assigned. Ultimately, we want to assign each subcarrier k to a subchannel s.

5.2.1 Full Usage of Subchannels (FUSC)

If the downlink uses FUSC, the system first assigns the pilot subcarriers, and then it assigns the remaining data subcarriers to different subchannels. The procedure of allocating data subcarriers to subchannels is as follows:

- Assign the pilot subcarriers. The number of pilot subcarriers to be assigned depends on N_{FFT}. They are:
  - If N_{FFT} = 128, the number of pilot subcarriers is 10.
  - If N_{FFT} = 512, the number of pilot subcarriers is 42.
  - If N_{FFT} = 1,024, the number of pilot subcarriers is 82.
  - If N_{FFT} = 2,048, the number of pilot subcarriers is 166.
- Assign the remaining data subcarriers to subchannels according to a permutation formula.

At the end, each subchannel would have 48 data subcarriers. Because of the permutation, data subcarriers in each subchannel are distributed across the band.

For pilot subcarriers, there are actually two types of pilot subcarriers in the downlink FUSC: constant-set pilot subcarriers and variable-set pilot subcarriers. The positions of constant-set pilot subcarriers do not change over successive OFDM symbols, whereas the positions of variable-set pilot subcarriers do change over suc-

cessive OFDM symbols. The reason for having variable-set pilot subcarriers is that they can estimate the channel at a variety of frequencies over time.

In FUSC, one slot is defined as one subchannel by one OFDM symbol. So the minimum allocation is 48 data subcarriers (in one OFDM symbol).

5.2.2 Partial Usage of Subchannels (PUSC)

If the downlink uses PUSC, the system initially assigns subcarriers to *clusters*. Then in each cluster, the system assigns the pilot subcarriers and the data subcarriers. The procedure of assigning data subcarriers to subchannels is as follows:

- Divide the subcarriers into clusters. Each cluster contains 14 subcarriers.
- Allocate the clusters to six groups in a permuted fashion. This way, physically adjacent clusters are more likely to be assigned to different groups.
- Assign the pilot subcarriers in each cluster of each group. Each cluster has two pilot subcarriers (and hence 12 data subcarriers).
- Assign the remaining data subcarriers (in all clusters of each group) to subchannels according to the same permutation formula used in downlink FUSC.

At the end, each subchannel would have four pilot subcarriers and 24 data subcarriers. For example, for $N_{FFT} = 1{,}024$ in 10 MHz of RF channel, there are 1,024 subcarriers. Out of these subcarriers, 840 subcarriers are for the pilot subcarriers and data subcarriers, and 184 subcarriers are for the DC subcarrier (at the middle of the RF channel) and the guard subcarriers (near the edges of the RF channel). The system divides the 840 subcarriers into 60 clusters, each consisting of 14 subcarriers. Each cluster contains two pilot subcarriers and 12 data subcarriers. Then using predefined algorithms, the system allocates the 60 clusters to six groups so that physically adjacent clusters are assigned to different groups (to achieve frequency diversity).

Afterwards, the system assigns pilot subcarriers in each cluster to predefined locations. (The locations of pilot subcarriers are different depending on if they are sent in an even-numbered or odd-numbered OFDM symbol in a permutation zone.) Then, in an OFDM symbol, the system proceeds to form subchannels for each group. It does so by choosing 24 data subcarriers from the clusters assigned to a group; it uses a permutation formula to choose the 24 data subcarriers so that they are picked from different frequency locations. As an illustration, group 0 has 12 clusters. In one OFDM symbol, the system forms a subchannel for group 0 by choosing 24 data subcarriers from the 12 clusters in group 0.

The 12 clusters in group 0 contain 144 data subcarriers because each cluster has 12 data subcarriers. Thus the system can form six subchannels for group 0.

Note that in even-numbered and odd-numbered OFDM symbols, the system forms different versions of the six subchannels to accommodate the different locations of the pilot subcarriers (in even-numbered and odd-numbered OFDM symbols). In addition, the assignment of data subcarriers to subchannels is different across neighboring cells.

The reason why PUSC is called *partial* usage of subchannels is that, in the PUSC permutation mode, a sector of a base station can only use some (but not all) of the available subchannels. In other words, a sector of a base station can use one group of subcarriers. This is done through *segmentation*, which is similar to sectorization in cellular systems. A segment is a subdivision of available subchannels for deploying a single instance of MAC [1]. In effect, the clusters are assigned to six groups, and a sector of a base station can use one group of subcarriers. (See Section 5.9.2 for more details on segmentation.)

In downlink PUSC, one slot is defined as one subchannel by two OFDM symbols. So the minimum allocation is again 48 data subcarriers (in two OFDM symbols).

5.2.3 Tile Usage of Subchannels 1 (TUSC1)

This subcarrier permutation mode was first specified by the IEEE 802.16e standard as an enhancement. The physical structure of TUSC1 on the downlink matches that of PUSC on the uplink. PUSC on the uplink is described in Section 5.3.1.

In TDD, this correspondence between the downlink (e.g., TUSC1) and the uplink (e.g., PUSC) turns out to be important when the system uses multiple-antenna techniques with closed-loop feedback. The reason is that if the structures match between the downlink and the uplink, then the system can exploit channel reciprocity between the downlink and the uplink.

In TUSC1, one slot is defined as one subchannel by three OFDM symbols. As seen in Section 5.3.1, the minimum allocation is 48 data subcarriers (over three OFDM symbols).

5.2.4 Tile Usage of Subchannels 2 (TUSC2)

This subcarrier permutation mode was also first specified by the IEEE 802.16e standard as an enhancement. The physical structure of TUSC2 on the downlink matches that of optional PUSC on the uplink. Optional PUSC on the uplink is described in Section 5.3.2. Similarly in TDD, using TUSC2 on the downlink and optional PUSC on the uplink allows the system to exploit channel reciprocity.

In TUSC2, one slot is defined as one subchannel by three OFDM symbols. As seen later in Section 5.3.2, the minimum allocation is also 48 data subcarriers (over three OFDM symbols).

5.3 Distributed Subcarrier Permutation: Forming Subchannels on Uplink

On the uplink, the required subcarrier permutation mode is partial usage of subchannels (PUSC). There is also an optional mode called optional PUSC. Both modes were first specified in the IEEE 802.16-2004 standard. These modes also use *distributed subcarrier permutation* because they form a subchannel by using nonadjacent subcarriers (which are pseudorandomly distributed throughout the frequency band). Again, the goal is to assign each subcarrier k to a subchannel s.

5.3.1 Partial Usage of Subchannels (PUSC)

Using PUSC on the uplink, the system initially assigns subcarriers to *tiles*. Then in each tile, the system assigns the pilot subcarriers and the data subcarriers. The procedure of assigning data subcarriers to subchannels is as follows:

- Divide the subcarriers over three OFDM symbols into tiles. Each tile contains four subcarriers over three OFDM symbols, or 12 subcarriers.
- Assign the pilot subcarriers in each tile. Each tile has four pilot subcarriers (and hence eight data subcarriers).
- Assign tiles to subchannels according to a permutation formula. This way, the data subcarriers in the tiles are assigned to subchannels as well.

Each subchannel is formed using six tiles. Thus, each subchannel has (6×4) or 24 pilot subcarriers and (6×8) or 48 data subcarriers. The subcarriers in a tile are adjacent, but the tiles in a subchannel are not physically adjacent. In addition, the assignment of data subcarriers to subchannels is different across neighboring cells.

Each subchannel has 24 pilot subcarriers and 48 data subcarriers. The high number of pilot subcarriers as compared to the number of data subcarriers affords excellent channel estimation on the uplink. Thus PUSC can be used in those areas with large multipath delay spread (e.g., urban environments). Recall from Chapter 2 that coherence bandwidth W_c is the range of frequency over which the transfer function $H(f)$ of the channel varies little, and W_c is inversely proportional to delay spread τ_{MAX}. As delay spread goes up, coherence bandwidth goes down. A small coherence bandwidth means that the transfer function of the channel has many variations as a function of frequency, and the channel requires more pilots to estimate. However, the better channel estimation comes at a cost of reduced user bit rate due to the high number of pilot subcarriers.

In uplink PUSC, one slot is defined as one subchannel by three OFDM symbols. So the minimum allocation is 48 data subcarriers (over three OFDM symbols).

5.3.2 Optional Partial Usage of Subchannels (Optional PUSC)

The optional PUSC on the uplink is similar to PUSC described in the previous section. The major differences are that in the optional PUSC,

- Each tile contains three subcarriers over three OFDM symbols, or nine subcarriers.
- Each tile has one pilot subcarrier (and hence eight data subcarriers).

Here, each subchannel is also formed using six tiles. This way, each subchannel has (6×1) or six pilot subcarriers and (6×8) or 48 data subcarriers. The low number of pilot subcarriers as compared to the number of data subcarriers allows a higher user bit rate, but it does not offer as good of a channel estimation. Thus optional PUSC can be used in those areas with small multipath delay spread.

In optional PUSC, one slot is defined as one subchannel by three OFDM symbols. Again, the minimum allocation is 48 data subcarriers (over three OFDM symbols).

5.4 Adjacent Subcarrier Permutation: Downlink and Uplink

An optional method of forming subchannels is called *adjacent subcarrier permutation*. The IEEE 802.16-2004 standard first specified this method for the downlink only. Then IEEE 802.16e specified the same method for the uplink. Adjacent subcarrier permutation essentially forms a subchannel using subcarriers that are adjacent to each other in frequency. In other words, each subchannel contains subcarriers that are contiguous in frequency. Adjacent subcarrier permutation is also known as adaptive modulation and coding (AMC) mode. In AMC mode, the system can quickly assign a specific modulation and FEC coding by using fast feedback channels [4]. Downlink AMC has the same structure as the uplink AMC mode [5].

The procedure of allocating data subcarriers to subchannels is as follows:

- Divide the subcarriers in one OFDM symbol into bins. Each bin contains nine subcarriers.
- Assign one pilot subcarrier in each bin. Each bin has one pilot subcarrier (and hence eight data subcarriers).
- Assign bins to subchannels according to several fixed schemes or *types*. They are:
  - *First type:* Assign six consecutive bins to a subchannel.
  - *Second type:* Assign two bins to a subchannel.
  - *Third type:* Assign three bins to a subchannel.
  - *Fourth type:* Assign one bin to a subchannel.
- Then the slot is defined as follows:
  - *First type:* One slot is defined as one subchannel by one OFDM symbol.
  - *Second type:* One slot is defined as one subchannel by three OFDM symbols.
  - *Third type:* One slot is defined as one subchannel by two OFDM symbols.
  - *Fourth type:* One slot is defined as one subchannel by six OFDM symbols.

As readers can see, each slot is always formed using six bins. Thus, each slot always has (6×1) or six pilot subcarriers and (6×8) or 48 data subcarriers. Incidentally, the position of the pilot subcarrier in each bin does not change, and this fixed position supports the operation of the advanced antenna systems (AAS) [6].

5.5 Summary of Subcarrier Permutation Modes

Figure 5.5 depicts a summary of different subcarrier permutation modes. Distributed subcarriers are well suited for frequency diversity, while adjacent subcarriers are more appropriate for multiuser diversity. In addition, downlink TUSC1 structurally matches uplink PUSC, and downlink TUSC2 structurally matches uplink optional PUSC. This way, in TUSC1 and TUSC2 the system can allocate (to a user) a downlink burst and an uplink burst that occupy the same subcarriers. In TDD, this arrangement is beneficial to closed-loop multiple-antenna techniques, where the transmitter (i.e., base station) requires channel feedback from the receiver (i.e.,

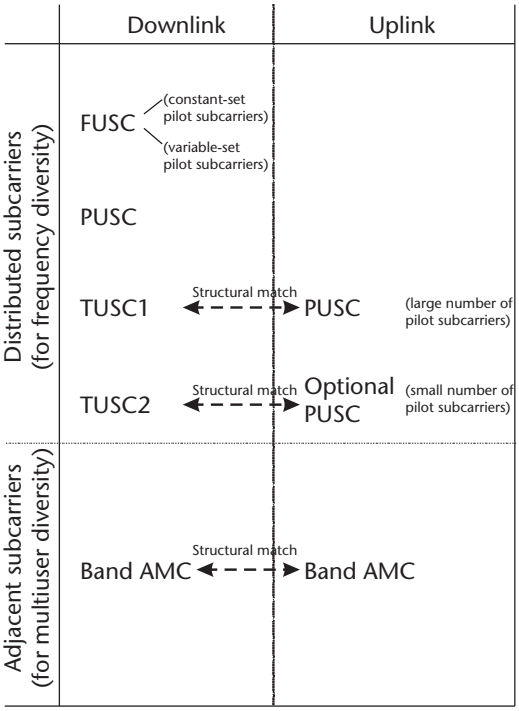


Figure 5.5 Summary of subcarrier permutation modes in IEEE 802.16e.

mobile). If a (user's) downlink burst and the uplink burst occupy the same subcarriers (and frequencies), then the base station can exploit channel reciprocity and infer the channel state based on the uplink. This arrangement minimizes the amount of channel feedback necessary from the mobile.

Interestingly, the minimum allocation in all the modes is always 48 data subcarriers (carrying 48 data symbols).

5.6 Bursts and Permutation Zones

After subchannels and slots are defined, the next higher units of organization are bursts and permutation zones. A *burst* is an allocation of a group of contiguous slots. Figure 5.6 shows an example (on the downlink). Occupying this two-dimensional space of subchannels (i.e., frequency) by OFDM symbols (i.e., time), a burst consists of a group of neighboring slots. As Figure 5.6 shows, a slot is the smallest unit of physical-layer resource that can be allocated to a user.

On the downlink, a burst is a rectangular allocation of logically contiguous subchannels and contiguous OFDM symbols. The base station transmits a burst using a particular modulation and coding scheme specified by the downlink map (DL-MAP) message [1]. A burst is also known as a data region [6], which can be allocated to a single user (i.e., unicast), to selected users (i.e., multicast), or to all users (i.e., broadcast) [4].

On the uplink, a single burst must span the entire permutation zone. Typically, only one permutation zone exists on the uplink. This is because uplink traffic is

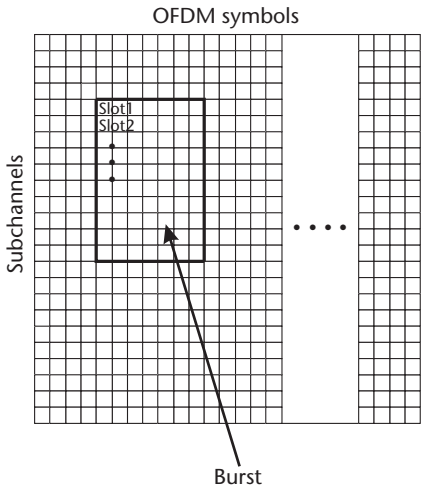


Figure 5.6 A burst in the two-dimensional layout of OFDM symbols by subchannels.

typically lighter than downlink traffic. In contrast, on the downlink there is typically more than one permutation zone in the downlink subframe.

A *permutation zone* is a section of OFDM symbols in time that uses the same subcarrier permutation mode. Figure 5.7 shows an example. In this two-dimensional space of subchannels (i.e., frequency) by OFDM symbols (i.e., time), a permutation zone is simply a section of OFDM symbols during which a particular subcarrier permutation mode (e.g., PUSC) applies. It is important to note that more than one permutation zone may exists in a downlink subframe (or in an uplink subframe). On the downlink, the base station dictates the transition to a different permutation zone through the DL-MAP message. On the uplink, the base station communicates the permutation zone to use through the uplink map (UL-

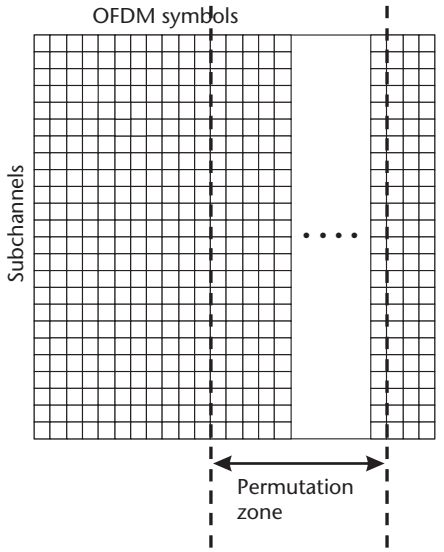


Figure 5.7 Permutation zone in the two-dimensional layout of OFDM symbols (in time) by subchannels (in frequency).

MAP)message. In addition, the IEEE 802.16e standard also supports a special zone specifically used for AAS.

5.7 Subframes and Frames

Bursts are used to constitute a downlink subframe and an uplink subframe, which together make up a *frame*. Figure 5.8 shows an example of a frame, which consists of a downlink subframe and an uplink subframe. In TDD, the uplink subframe follows the downlink subframe. At the end of the downlink subframe, the base station stops transmitting, waits for a period of time equal to the transmit/receive transition gap (TTG), then starts receiving the uplink subframe. The TTG allows the base station's hardware and software to switch from transmitting to receiving, as well as allows the mobile to switch from receiving to transmitting. At the end of the uplink subframe, the base station stops receiving, waits for a period of time equal to the receive/transmit transition gap (RTG), then starts transmitting the next frame. The RTG allows the base station to switch from receiving to transmitting; it also allows the mobile to switch from transmitting to receiving.

In general, the advantage of a longer frame is higher transmission efficiency because the frame can transport more user data for a given amount of frame overhead. The disadvantage of a longer frame is a longer delay because users have to wait for the entire (longer) frame to be processed before getting their data. The 5-ms frame shown in Figure 5.8 results in a balance between low delay (and jitter) and reasonable transmission efficiency [7]. For 5-MHz and 10-MHz channels, there are a total of 47 OFDM symbols available for downlink and uplink subframes (excluding TTG and RTG).

Figure 5.9 shows an example of a downlink subframe. During the period of the downlink subframe, the base station transmits a preamble, a frame control header (FCH), a downlink map (DL-MAP) message, and an uplink map (UL-MAP) mes-

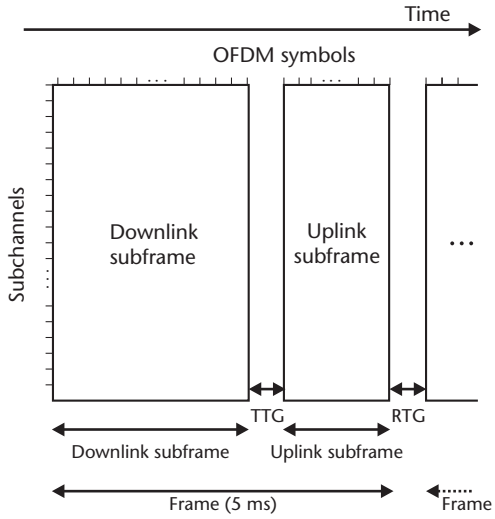


Figure 5.8 Illustration of a frame in OFDMA/TDD. (*After:* [3].)

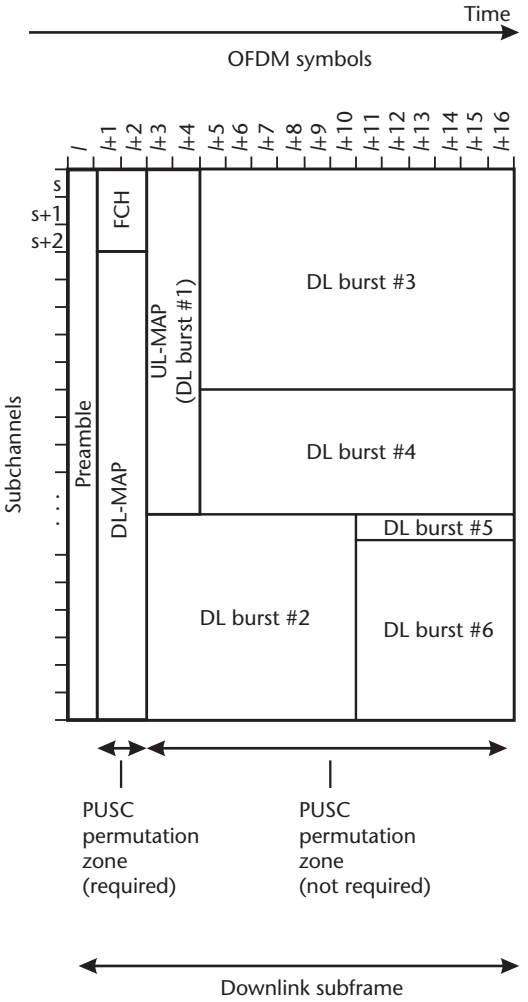


Figure 5.9 Illustration of a downlink subframe in OFDMA/TDD. (*After:* [3].)

sage. Then the base station transmits different bursts to different mobiles in its coverage area.

Figure 5.10 shows an example of an uplink subframe. After TTG, the base station starts to receive the uplink subframe, which may contain bursts from more than one mobile in the coverage area of the base station. Note here that some subchannels are used for ranging; these subchannels are dynamically assigned, and their location is shown in the UL-MAP message. The mobile uses the ranging channel to synchronize with the system, as well as to make contention-based bandwidth requests. (See Chapter 8 for more details on ranging.) In addition, the uplink channel quality indicator channel (CQICH) and the uplink ACK channel may be allocated in the uplink subframe. The uplink CQICH is for the mobile to quickly send information on the channel back to the base station, and this information on the channel is quantized to 64 levels using six bits. The uplink ACK channel is for the mobile to send downlink HARQ ACKs back to the base station [8].

Figure 5.9 and Figure 5.10 also show the different permutation zones that may exist in a frame. The first permutation zone in every frame must be PUSC, which is

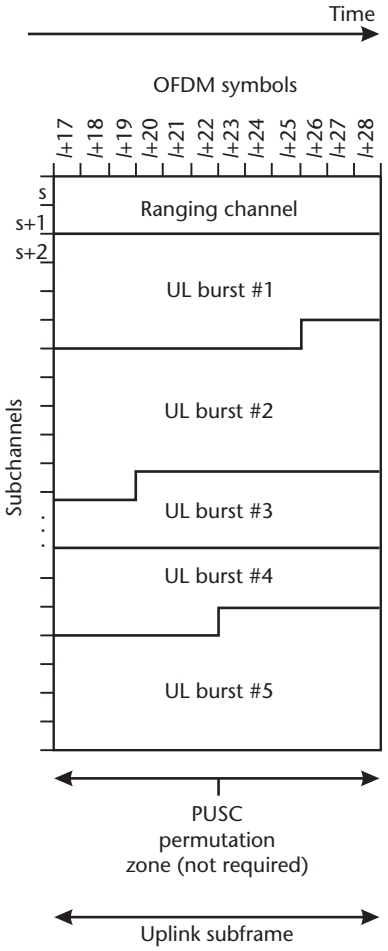


Figure 5.10 Illustration of an uplink subframe in OFDMA/TDD. (*After:* [3].)

used to transmit the DL-MAP message and the FCH. In Figure 5.9, after the first PUSC permutation zone, another PUSC permutation zone follows in the downlink subframe. In Figure 5.10, the uplink subframe contains a single PUSC permutation zone. The length of the frame is variable depending on how many bursts are (and how much user data is) sent. In fact, because only the preamble and the first PUSC permutation zone are required in each frame, it is possible for a frame to have only the preamble and the first PUSC permutation zone (containing the DL-MAP and the FCH), with no subsequent bursts on the downlink and the uplink.

In this illustration of the downlink subframe, both the FCH and the DL-MAP message last two OFDM symbols; this is because in downlink PUSC, one slot is defined as one subchannel by two OFDM symbols, and both FCH and DL-MAP are in the (required) PUSC permutation zone. Recall that a slot is the smallest unit of physical-layer resource allotted. Similarly for the different bursts in the next PUSC permutation zone (in the same downlink subframe), a burst lasts a multiple of two OFDM symbols because, again, one slot is defined as one subchannel by two OFDM symbols in downlink PUSC.

In the uplink subframe, on the other hand, a burst lasts a multiple of three OFDM symbols; this is because in uplink PUSC, one slot is defined as one subchannel by three OFDM symbols.

In an OFDMA system, allocations of bursts to users can be different from one frame to the next frame. This way, the system can quickly respond to users' requests for bandwidth and dynamically allocate resources to users from frame to frame. The ratio of downlink subframe duration to uplink subframe duration is typically 3:1 for Web traffic because the Web surfing profile is highly asymmetric. The same ratio becomes 1:1 for voice traffic because voice traffic is symmetric. The default ratio becomes 2:1 for a mix of Web traffic and voice traffic [7].

5.7.1 Preamble

The first OFDM symbol of a frame is the preamble, which is known a priori to the mobile. In other words, its data symbols in the OFDM symbol are known to the mobile. Transmitted on the downlink, the preamble is used for initial timing synchronization, initial frequency estimation, and initial channel estimation. Using the preamble, the mobile can measure the carrier-to-interference and noise ratio (CINR) and report it back to the base station via the MOB_SCN-REP (scanning result report) message or the MOB_MSHO-REQ (mobile station handover request) message. Figure 5.11 shows how subcarriers in the preamble are organized. All subcarriers in the preamble are assigned to three groups (or "carrier sets" as they are called in the standard). Those subcarriers marked "0" belong to group 0, those subcarriers marked "1" belong to group 1, and those subcarriers marked "2" belong to group 2. As shown in the figure, the subcarriers in the same group are simply spaced three subcarriers apart.

The reason why the subcarriers in the preamble are assigned to three groups is because they can then be allocated to three segments (sectors) of a base station. In particular, segment 0 uses subcarriers in group 0 of the preamble, segment 1 uses subcarriers in group 1 of the preamble, and segment 2 uses subcarriers in group 2 of the preamble. To distinguish the different segments in a geographic area, each segment in a cluster of base stations is modulated by a different pseudonoise (PN) code. For maximum link performance, the subcarriers in the preamble are transmitted using BPSK at an elevated transmit power.

5.7.2 Frame Control Header (FCH)

The FCH contains a data structure called DL_Frame_Prefix. DL_Frame_Prefix has 24 bits of information,[1] including:

- The "used subchannel bitmap," which shows what subchannels are used (in PUSC) by the segment (sector) transmitting the FCH.
- The forward error correction (FEC) code used to transmit the subsequent DL-MAP message (i.e., convolutional code and block turbo code). The DL-MAP is always coded at the rate of 1/2.

1. DL_Frame_Prefix has 12 bits for $N_{FFT} = 128$.

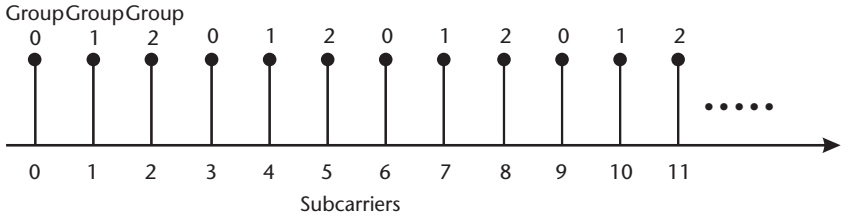


Figure 5.11 The organization of subcarriers in the preamble.

- The length of the DL-MAP message. This way, the mobile has sufficient information to read the DL-MAP message that follows the FCH.

The FCH transmission begins in the first subchannels of the second OFDM symbol of the frame (see Figure 5.9). For robust link performance, the FCH is transmitted using QPSK at a rate of 1/2. To ensure that the mobile receives the FCH, the base station transmits the FCH with a repetition rate of four.[2]

The standard refers to DL_Frame_Prefix as a *data structure* rather than a message. This is because a *message* needs to come from the MAC layer. DL_Frame_Prefix is technically not a message because it originates from the physical layer, not the MAC layer. Thus, the first message in the frame is really the DL-MAP message that follows the FCH.

5.7.3 Downlink MAP (DL-MAP) and Uplink MAP (UL-MAP)

After the FCH, the base station transmits downlink map (DL-MAP) followed by uplink map (UL-MAP), which are MAC messages. Specifically, DL-MAP contains information on subchannels and OFDM symbols that are assigned to each mobile in the downlink subframe, and UL-MAP has information on subchannels and OFDM symbols that are assigned to each mobile in the uplink subframe. Every mobile needs to read both DL-MAP and UL-MAP to find out what its allocations are in the frame. DL-MAP and UL-MAP are called *map* messages because they define maps of different bursts that are allocated to different mobiles (in the subchannel × OFDM symbol space). DL-MAP and UL-MAP are critical control messages used for allocating resources to the mobiles. The map messages effectively point out the place (in time and frequency) where the system has allocated resources to a mobile.

5.8 TDD and FDD

This chapter focuses much of its discussions on TDD. But it is important to note that IEEE 802.16e also supports FDD (if, for example, paired bands are available). The subframe structure in FDD is not that much different from that in TDD. In TDD, a base station transmits a downlink subframe and receives the uplink subframe at different times, as shown in Figure 5.8. In FDD, a base station simply transmits a downlink subframe in one frequency band and receives the uplink

2. Repetition is not applied for $N_{FFT} = 128$.

subframe in another frequency band. Similar operations take place at the mobile as well. Obviously, both TTG and RTG are not used in FDD.

5.9 System Design Issues

In this section, we address some of the system design issues arising out of the topics discussed in this chapter. The IEEE 802.16e standard has built-in options that allow system designers much flexibility in adopting their systems to specific environments.

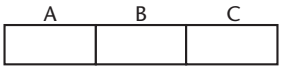
5.9.1 Frequency Diversity and Multiuser Diversity

In forming subchannels, distributed subcarrier permutation modes use permutation schemes to pseudorandomly disperse subcarriers to form a subchannel, thus achieving *frequency diversity*. Distributed subcarrier permutation can be used in those environments where there are moving mobiles. While mobiles are moving, their transfer functions would change as a function of time as well. In these situations, frequency diversity can be employed to ensure a higher probability of survival of a subchannel, and it can be assumed that using distributed subcarrier permutation, subchannels have a similar quality. Another advantage of distributed subcarriers is intercell (or intersector) interference averaging. If adjacent cells (or sectors) use the same physical RF channel and subcarriers are scattered pseudorandomly across the channel, then the chance that identical subcarriers are used is lowered [6].

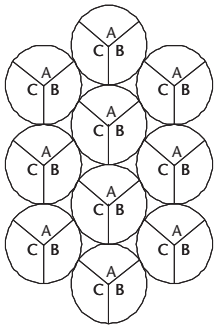
On the other hand, adjacent subcarrier permutation modes construct subchannels using subcarriers that are contiguous in frequency. Needless to say, adjacent subcarrier permutation modes cannot deliver much frequency diversity because subcarriers (used to form a subchannel) are together in frequency. However, adjacent subcarrier permutation can leverage *multiuser diversity*. The idea is that if users are distributed in locations around a base station, then there is already diversity in these mobiles' transfer functions. Thus, a user can be allocated a subchannel at a part of the frequency band that has a high SINR. In an environment with high multiuser diversity, each user should experience high SINR at a different part of the frequency band; subcarriers with high SINRs can use higher-order modulation, which increases bandwidth efficiency (bps/Hz) [9]. By assigning each user a subchannel where the subchannel has high SINR, the base station can maximize its aggregate throughput. Typically, adjacent subcarriers are more suited for fixed or nomadic users [5].

5.9.2 Segmentation

In IEEE 802.16e, segmentation is akin to sectorization in cellular systems. To see the use of segmentation, consider a network access provider who has won through auction a slice of spectrum that can accommodate three physical RF channels (see Figure 5.12). This access provider intends to deploy an IEEE 802-16-based system. Using TDD, an access provider can assign these three RF channels A, B, and C to three sectors of a base station in a system with (intracell) frequency reuse factor $K = 3$, as shown in Figure 5.12. For those access providers that have sufficient spectrum, frequency reuse is expected to be a predominant approach of deploying



(a)

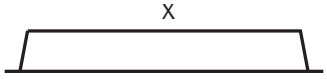


(b)

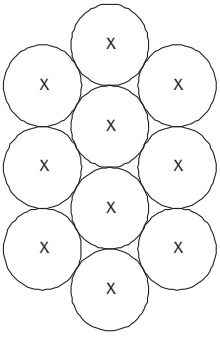
Figure 5.12 A system using (intracell) frequency reuse factor $K = 3$. (a) Three physical RF channels in frequency. (b) Three physical RF channels arranged in space.

an IEEE 802.16-based system because of the need to minimize adjacent cell/sector interference.

On the other hand, a network access provider who does not have sufficient spectrum may have a problem with managing adjacent cell/sector interference. Consider an access provider who has only enough spectrum for a single physical RF channel (see Figure 5.13). Using TDD, this RF channel X can only be deployed using frequency reuse factor $N = 1$, as shown in Figure 5.13. But deploying a system with frequency reuse factor $N = 1$ can introduce severe adjacent cell interference.



(a)



(b)

Figure 5.13 A system using frequency reuse factor $N = 1$. (a) Only one physical RF channel in frequency. (b) The one physical RF channel is used everywhere.

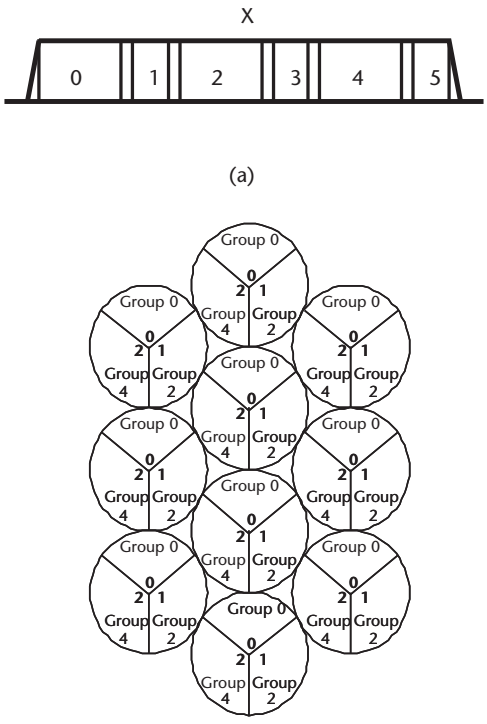


Figure 5.14 A system using segmentation: (a) Subchannel groups for segmentation. Note that subchannels in each group are not physically adjacent to each other due to distributed subcarrier permutation. (b) Subchannel groups arranged in space.

The problem of interference would be especially severe in those places that border between cells/sectors.

Segmentation, in effect, affords the second access provider (who does not have sufficient spectrum) a way to implement frequency reuse. Figure 5.14 shows how this can be done. Figure 5.14(a) shows that within a single RF channel, the standard allocates six groups of subchannels (groups 0, 1, 2, 3, 4, and 5).[3] This way, the system designer can assign these groups of subchannels to different segments (sectors). The standard stipulates that, if segmentation is used, segment 0 should use subchannels in group 0, segment 1 should use subchannels in group 2, and segment 2 should use subchannels in group 4 [see Figure 5.14(b)]. For N_{FFT} = 1,024 and N_{FFT} = 2,048, groups 1, 3, and 5 are available and can be assigned on-demand among the three segments. For example, groups 1, 3, and 5 may all be assigned to one segment, or groups 1, 3, and 5 may be assigned to segments 0, 1, and 2, respectively. Note that segmentation is supported on both the downlink and the uplink.

For reference, Table 5.1 shows, for different N_{FFT}, the subchannels used in different subchannel groups.

For example, for N_{FFT} = 1,024 in 10 MHz of the RF channel in downlink PUSC, groups 0, 2, and 4 have six subchannels (or 12 clusters of subcarriers) per group, while groups 1, 3, and 5 have four subchannels (or eight clusters of subcarriers) per group. In downlink PUSC (see Section 5.2.2), each subchannel has four

3. For N_{FFT} = 128 and N_{FFT} = 512, no subchannel is allocated to group 1, group 3, and group 5.

Table 5.1 Subchannels Used in Subchannel Groups

Subchannel Group	N_{FFT}			
	128	512	1,024	2,048
0	0	0–4	0–5	0–11
1	—	—	6–9	12–19
2	1	5–9	10–15	20–31
3	—	—	16–19	32–39
4	2	10–14	20–25	40–51
5	—	—	26–29	52–59

pilot subcarriers and 24 data subcarriers. Thus, in groups 0, 2, and 4 there are 24 pilot subcarriers and 144 data subcarriers per group, while in groups 1, 3, and 5 there are 16 pilot subcarriers and 96 data subcarriers per group.

5.10 Adaptive Burst Profiles

5.10.1 Burst Profiles

In IEEE 802.16e, a burst profile is a specific combination of modulation and FEC (including FEC rate) assigned to a user. To implement adaptive modulation and coding (see Chapter 3), the base station can dynamically change a user's burst profile in response to changing channel conditions experienced by that user. If the channel condition is good, then the link can use a more bandwidth-efficient burst profile (e.g., 64-QAM, rate 3/4 convolutional code) to maximize bit rate. If the channel condition is poor, then the link can throttle back to a more robust burst profile (e.g., QPSK, rate 1/2 convolutional code) to maximize link reliability. In general, a system that uses adaptive modulation and coding performs better than a system that does not use adaptive modulation and coding [10, 11].

The burst profile applies to a burst and can change from burst to burst for each user, thus enabling the system to trade off bit rate with reliability in real time. Table 5.2 lists some examples of burst profiles. During the registration process, the mobile tells the base station the set of burst profiles that the mobile supports.

An index called downlink interval usage code (DIUC) identifies the specific burst profile on the downlink in IEEE 802.16; another index uplink interval usage code (UIUC) identifies the burst profile used on the uplink. The base station transmits the downlink and uplink burst profiles to the mobile by using MAC management messages such as the downlink channel descriptor (DCD) message and the uplink channel descriptor (UCD) message.

In determining the burst profile to use, the base station can use three parameters [12]:

- *Entry threshold:* This is the SINR value above which the link may transition to a more bandwidth-efficient burst profile.

Table 5.2 Sample Burst Profiles

	QPSK	*16-QAM*	*64-QAM*
Convolutional code, rate = 1/2	√	√	√
Convolutional code, rate = 2/3	—	—	√
Convolutional code, rate = 3/4	√	√	√
Convolutional turbo code, rate = 1/2	√	√	√
Convolutional turbo code, rate = 2/3	—	—	√
Convolutional turbo code, rate = 3/4	√	√	√
Block turbo code, rate = 1/2	√	—	—
Block turbo code, rate = 3/4	√	—	—
Low density parity code, rate = 1/2	√	√	√
Low density parity code, rate = 2/3	√	√	√
Low density parity code, rate = 3/4	√	√	√

- *Exit threshold:* This is the SINR value below which the link transitions to a more robust burst profile.
- *Existing SINR:* This is the current SINR on the link.

Table 5.3 lists the minimum required SNR values for burst profiles that consist of QPSK, 16-QAM, and 64-QAM and convolutional turbo codes at different rates. These SNR values are for a probability of a bit error of 10^{-6} in the additive white Gaussion noise (AWGN) channel.

5.10.2 Channel Quality Feedback

Obviously, the base station knows the SINR of the uplink because it can directly measure it. On the downlink, the base station gets the SINR of the downlink through channel feedback sent by the mobile. The IEEE 802.16e standard supports two types of channel feedback: received signal strength indicator (RSSI) and carrier-to-interference plus noise ratio (CINR).

Table 5.3 Minimum Required SNR (Convolutional Turbo Code for $P_b = 10^{-6}$ in AWGN Channel) [2]

	SNR value
QPSK, convolutional turbo code, rate = 1/2	2.9 dB
QPSK, convolutional turbo code, rate = 3/4	6.3 dB
16-QAM, convolutional turbo code, rate = 1/2	8.6 dB
16-QAM, convolutional turbo code, rate = 3/4	12.7 dB
64-QAM, convolutional turbo code, rate = 1/2	13.8 dB
64-QAM, convolutional turbo code, rate = 2/3	16.9 dB
64-QAM, convolutional turbo code, rate = 3/4	18 dB

For RSSI, the mobile measures its total receive signal strength and reports it back to the base station through the channel measurement report response (REP-RSP) message. The advantage of the RSSI measurement is that it can be taken relatively quickly and does not require receiver demodulation. The disadvantage is that the measurement has everything in it, including signal, noise, and interference.

For CINR, the mobile demodulates the base station's transmission, separates out the signal from the noise and the interference, and reports the CINR back to the base station through the REP-RSP message or over a fast channel quality indicator channel (CQICH) (in the uplink subframe). The CINR measurement is a more accurate reflection of the channel condition, but its measurement requires receiver demodulation.

References

[1] Balachandran, K., et al., "Design and Analysis of an IEEE 802.16e-Based OFDMA Communication System," *Bell Labs Technical Journal*, Vol. 11, No. 4, 2007, pp. 53–73.

[2] IEEE Standard 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York: IEEE, October 1, 2004.

[3] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[4] Yaghoobi, H., "Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN," *Intel Technology Journal*, Vol. 8, No. 3, 2004, pp. 201–212.

[5] Upase, B., M. Hunukumbure, and S. Vadgama, "Radio Network Dimensioning and Planning for WiMAX Networks," *Fujitsu Scientific and Technical Journal*, Vol. 43, No. 4, 2007, pp. 435–450.

[6] Nuaymi, L., *WiMAX: Technology for Broadband Wireless Access*, New York: John Wiley & Sons, 2007.

[7] Jain, R., C. So-In, and A. -K. Al Tamimi, "System-Level Modeling of IEEE 802.16e Mobile WiMAX Networks: Key Issues," *IEEE Wireless Communications*, Vol. 15, No. 5, 2008, pp. 73–79.

[8] Etemad, K., "Overview of Mobile WiMAX Technology and Evolution," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 31–40.

[9] Wong, C. Y., et al., "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation," *IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 10, 1999, pp. 1747–1757.

[10] Sternad, M., et al., "Towards Systems Beyond 3G Based on Adaptive OFDMA Transmission," *Proceedings of the IEEE*, Vol. 95, No. 12, 2007, pp. 2432–2455.

[11] Svensson, A., "An Introduction to Adaptive QAM Modulation Schemes for Known and Predicted Channels," *Proceedings of the IEEE*, Vol. 95, No. 12, 2007, pp. 2322–2336.

[12] Belghith, A., and L. Nuaymi, "WiMAX Capacity Estimations and Simulation Results," *Proc. IEEE Vehicular Technology Conference*, May 2008, pp. 1741–1745.

CHAPTER 6

Physical Layer: Spatial Techniques

6.1 Introduction

While time and frequency techniques employed by OFDMA at the physical layer are well suited for high-speed, wireless data transmission, various spatial techniques are also available to further increase performance. Advanced broadband wireless systems not only exploit resources in time (i.e., OFDM symbols) and in frequency (i.e., subcarriers), but also take advantage of resources in space (i.e., antennas). Increasingly, system design efforts incorporate a view of time, frequency, and space together [1]. In these three domains (time, frequency, and space), two broad classes of methods exist: multiplexing and diversity. Figure 6.1 shows a framework depicting the different techniques.

In terms of *diversity,* distributed subcarriers and interleaving can take advantage of frequency diversity, and of course, HARQ and FEC are intrinsically forms of time diversity [2]. In terms of *multiplexing,* OFDM multiplexes data in frequency primarily to combat ISI. OFDMA, by assigning different users to different subcarriers, can leverage an additional degree of freedom to scale bandwidths for different users. In this chapter, we also examine diversity and multiplexing in the spatial domain.

With spatial techniques, the system uses multiple transmit and receive antennas, thereby artificially creating additional wireless paths between pairs of transmit and receive antennas; then it exploits these additional paths to achieve higher performance (e.g., bit rate and/or reliability). This chapter provides an introduction to some popular spatial techniques.

In general, there are three types of gains[1] that can be achieved with multiple antennas (M_t transmit antennas and M_r receive antennas) [3]:

- *Spatial diversity gain:* Spatial diversity improves link reliability by transmitting via multiple means in space and by appropriately combining the received signals. Spatial diversity gain depends on the paths being uncorrelated. It is well known that the maximum diversity gain is M_tM_r (i.e., diversity order)

1. Another type of gain from multiple antennas is array gain, which depends on the amount of signal power collected by multiple antennas, not on the paths being uncorrelated. It is attained when the receiver coherently combines the signals and increases the received SNR.

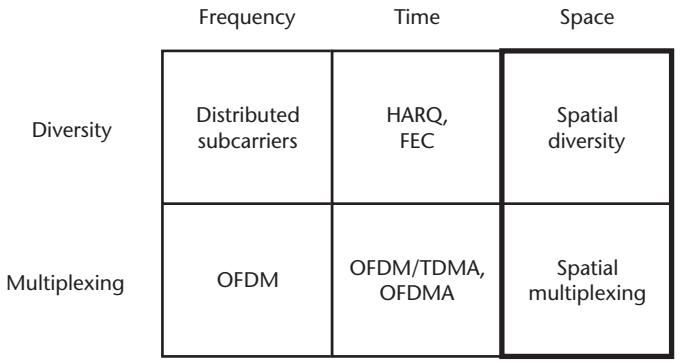


Figure 6.1 Different techniques in time, frequency, and space.

because $M_t M_r$ is the maximum number of uncorrelated paths between the transmitter and the receiver.[2]

- *Spatial multiplexing gain:* Spatial multiplexing increases the bit rate by transmitting via additional paths created by multiple antennas [4]. It is well known that spatial multiplexing can increase the bit rate by $\min(M_t, M_r)$ because $\min(M_t, M_r)$ is the number of unique received paths (over which unique data streams may travel) between the transmitter and the receiver.[3]
- *Cochannel interference reduction:* Multiple antennas can be used to discriminate between the desired signal and the cochannel interfering signals, hence reducing cochannel interference.

Along these three types of gains are three general classes of spatial techniques: spatial diversity (to improve reliability), spatial multiplexing (to increase bit rate), and beamforming (to reduce cochannel interference), each discussed in the following sections. Figure 6.2 depicts these techniques and their benefits. Spatial diversity results in lower error rates, spatial multiplexing achieves higher bit rates, and beamforming increases SINR, which can be used for lower error rates and/or higher bit rates.

In addition, a spatial technique can be termed open-loop or closed-loop. *Open-loop* refers to those techniques in which the transmitter has no knowledge of the channel transfer function, whereas *closed-loop* refers to those techniques in which the transmitter makes use of knowledge of the channel transfer function.

6.2 Spatial Diversity: Receive Diversity

Receive diversity has been used on cellular systems since they began appearing in early 1980s. Although receive diversity in the form of two receive antennas (at the base station) has been mostly implemented on the uplink, it has nevertheless

2. Assuming that the paths between multiple transmit antennas and multiple receive antennas are independent and identically distributed (i.i.d.) Rayleigh faded.

3. If the receiver has perfect knowledge of the channels.

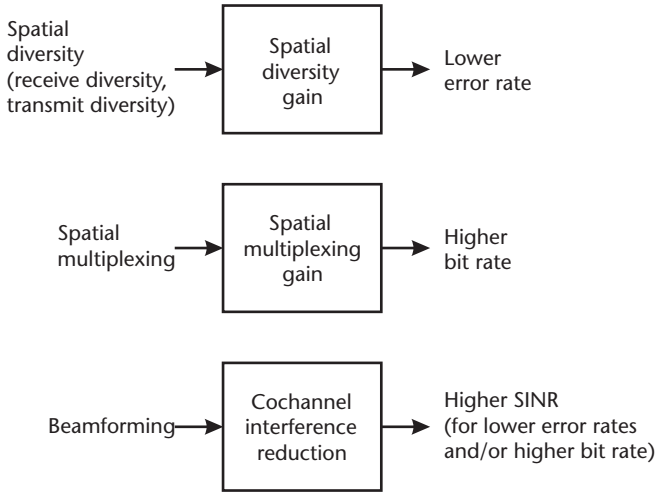


Figure 6.2 Different spatial techniques. (*After:* [5].)

enhanced the link that is often the weakest due to limited transmit power available at the mobile.

Figure 6.3 depicts a system that uses receive diversity. Here there is one transmit antenna and M_r receive antennas. It is well known that if the receive antennas are placed sufficiently apart, the M_r paths between the transmitter and the receiver would be approximately uncorrelated. These uncorrelated paths are important because when one path experiences a deep fade, another path would likely not experience a fade. Given this desired outcome (owing to uncorrelated paths), the question then becomes how the antenna postprocessor combines the M_r signals from the M_r paths for the receiver. For systems that use receive diversity, two combining schemes are popular: receive antenna selection and maximal ratio combining.

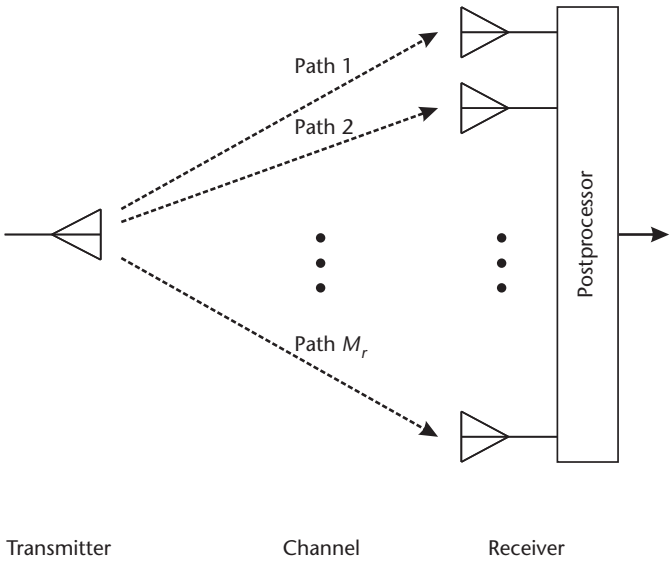


Figure 6.3 Receive diversity. Note that in an actual system there may be multiple transmit antennas as well.

6.2.1 Receive Diversity: Antenna Selection

In receive antenna selection, the receive antenna system measures the signal powers of the M_r paths and simply selects the path with the best signal power (see Figure 6.4). Although this scheme is simple in terms of implementation, it does not perform as well as maximal ratio combining because antenna selection throws away useful, uncorrelated (albeit weaker) signals gathered by other antennas. Receive antenna selection primarily offers (spatial) diversity gain because the signal from the best path is chosen [6].

6.2.2 Receive Diversity: Maximal Ratio Combining

In maximal ratio combining, the receive antenna system makes use of the signals from all paths. Specifically, it scales the signal of each path by a coefficient f_i and adds up all (scaled) signals to produce the final signal for the receiver (see Figure 6.5).

In general, the received symbols **y** can be written in the matrix format, that is,

$$\mathbf{y} = \mathbf{F}(\mathbf{Hx} + \mathbf{n}) \tag{6.1}$$

where **y** is the received symbol vector ($M \times 1$), **x** is the transmitted symbol vector ($M \times 1$), **H** is the channel matrix ($M_r \times M_t$), and **F** is the combining matrix ($M \times M_r$). **n** is the noise vector ($M \times 1$). If the system transmits one symbol at a time ($M = 1$), (6.1) can be rewritten as:

$$y = \mathbf{F}(\mathbf{H}x + \mathbf{n}) \tag{6.2}$$

For a system that has three receive antennas ($M_r = 3$) and one transmit antenna ($M_t = 1$), (6.2) can be rewritten as

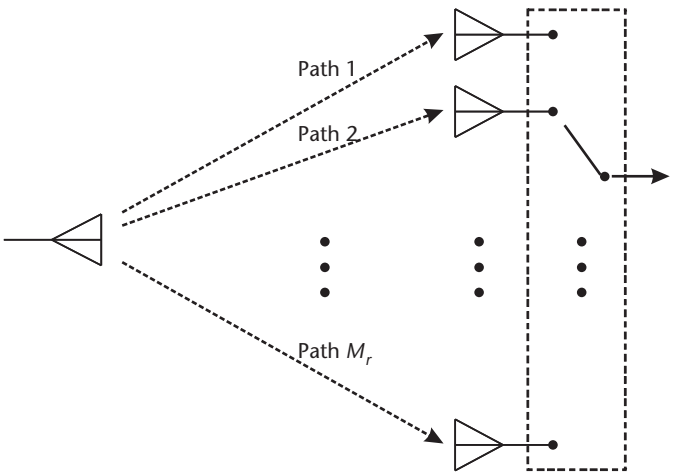


Figure 6.4 Antenna selection.

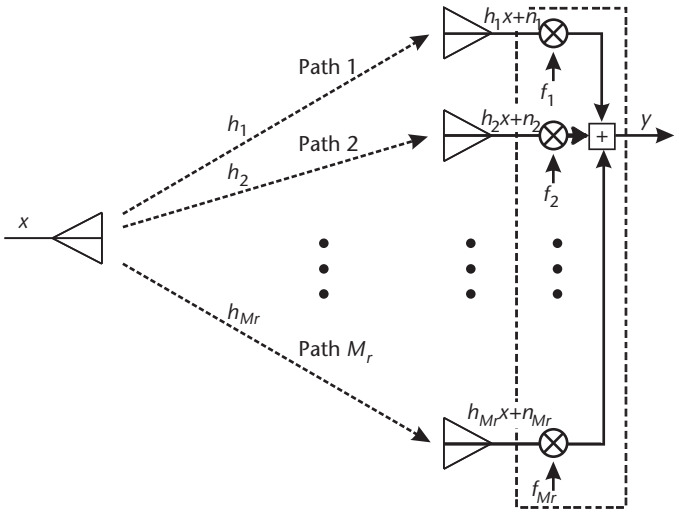


Figure 6.5 Maximal ratio combining.

$$\begin{aligned} y &= \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} x + \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \\ &= \left(f_1 h_1 + f_2 h_2 + f_3 h_3\right) x + \left(f_1 n_1 + f_2 n_2 + f_3 n_3\right) \end{aligned} \tag{6.3}$$

The coefficients (i.e., f_1, f_2, and f_3) can be derived in such a way that the resultant SNR of the combined signal y is maximized. Thus, the maximal ratio combining offers both array gain and diversity gain [6].

6.3 Spatial Diversity: Transmit Diversity

Contrary to the long history of receive diversity, transmit diversity has only recently become popular in actual deployments. A transmit diversity system has multiple antennas at the transmitter. Transmit diversity is typically implemented on the downlink for two reasons. First, although the downlink is often thought of as the stronger of the two links (because the base station has more transmit power available), experience with 3G wireless systems has shown that a system is often downlink-limited. This is because in 3G many mobile devices run bandwidth-intensive applications such as video streaming that is asymmetrically biased toward the downlink. If enough mobiles in a cell run such applications, then the base station serving the cell can quickly exhaust its transmit power resources. This type of resource limitation is termed power-limited. Second, because of the small size and limited processing power of the mobile device, mounting even two receive antennas on a mobile device may be difficult. For example, at 700 MHz, half a wavelength is:

$$\frac{1}{2}\lambda = \frac{1}{2}\frac{c}{f} = \frac{1}{2}\frac{3\times10^8\,\text{m}}{700\times10^6\,\text{Hz}} = 0.21\text{m} \approx 8.4\text{ in} \tag{6.4}$$

While it is possible to mount two antennas 8.4 inches apart on a notebook computer, it would be difficult to mount the same two antennas on a smartphone because of its small size. Therefore, implementing transmit diversity (at the base station) lessens the need for multiple receive antennas at the mobile device, decreases receiver complexity, and shifts the cost and complexity to the base station.

Figure 6.6 shows a transmit diversity scheme. There are M_t transmit antennas at the transmitter and one receive antenna at the receiver. As a result, there are M_t paths between the transmitter and the receiver, and each path has a transfer function h_i. Analogous to the *postprocessor* at the receiver for receive diversity, there is now a *preprocessor* at the transmitter for transmit diversity. The preprocessor trains the transmit signal and encodes it for transmission out of multiple transmit antennas.

6.3.1 Transmit Diversity: Open-Loop 2 × 1

Figure 6.7 shows a basic open-loop transmit diversity system that has two transmit antennas and one receive antenna. This is referred to as a 2 × 1 system. In this system, the preprocessor prearranges consecutive symbols (from the transmitter) S_1 and S_2 and sends them out of two transmit antennas in the following manner: S_1 and $-S_2^*$ from antenna 1 and S_2 and S_1^* from antenna 2. The key idea is that the transmitter transmits a symbol in one symbol period through one antenna and retransmits the same symbol in the next symbol period through another antenna (see Figure 6.7).

At the receiving end, the received symbol in the first symbol period is

$$R_1 = S_1 h_1 + S_2 h_2 + N_1 \tag{6.5}$$

and the received symbol in the second symbol period is

$$R_2 = -S_2^* h_1 + S_1^* h_2 + N_2 \tag{6.6}$$

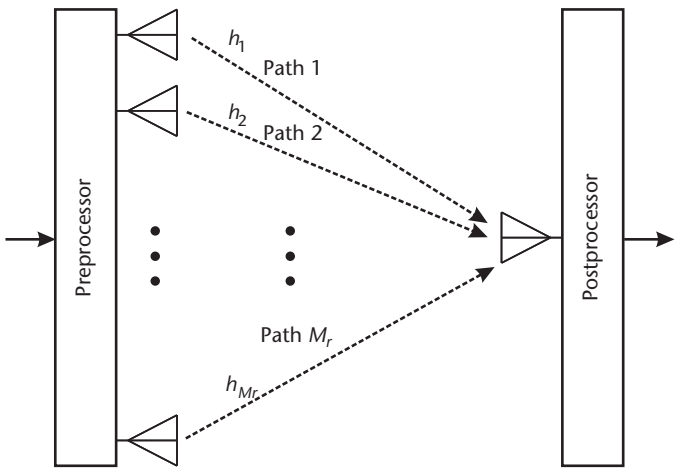


Figure 6.6 Transmit diversity. Note that in an actual system, there are typically multiple receive antennas as well.

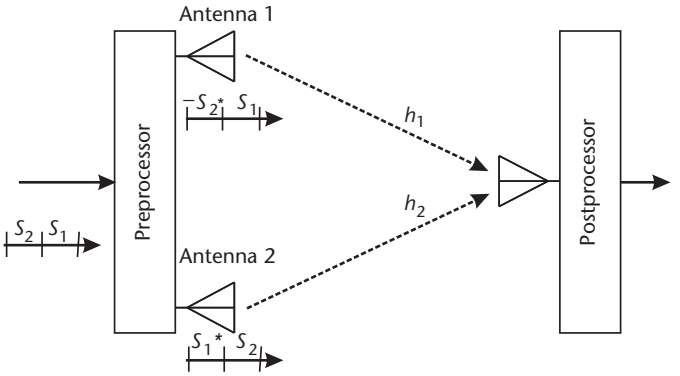


Figure 6.7 Open-loop 2 × 1 transmit diversity.

Given these received symbols, the postprocessor can estimate and recover the transmitted symbols $\tilde{S}_1$ and $\tilde{S}_2$ by using the following equations:

$$\tilde{S}_1 = R_1 h_1^* + R_2^* h_2 \tag{6.7}$$

$$\tilde{S}_2 = R_1 h_2^* - R_2^* h_1 \tag{6.8}$$

To see how the postprocessor can recover the transmitted symbols $\tilde{S}_1$ and $\tilde{S}_2$, substitute (6.5) and (6.6) into (6.7). This results in:

$$\begin{aligned}\tilde{S}_1 &= h_1 h_1^* S_1 + h_1^* h_2 S_2 + h_1^* N_1 - h_1^* h_2 S_2 + h_2 h_2^* S_1 + h_2 N_2^* \\ &= \left(|h_1|^2 + |h_2|^2\right) S_1 + h_1^* N_1 + h_2 N_2^*\end{aligned} \tag{6.9}$$

Note that the recovered symbol $\tilde{S}_1$ depends only on the transmitted symbol S_1 (and noise) but not S_2. Similarly, substituting (6.5) and (6.6) into (6.8) yields:

$$\begin{aligned}\tilde{S}_2 &= h_1 h_2^* S_1 + h_2^* h_2 S_2 + h_2^* N_1 + h_1^* h_1 S_2 - h_1 h_2^* S_1 - h_1 N_2^* \\ &= \left(|h_1|^2 + |h_2|^2\right) S_2 + h_2^* N_1 - h_1 N_2 *\end{aligned} \tag{6.10}$$

which only depends on the transmitted symbol S_2 (and noise) but not S_1.

Two observations can be made:

- By using (6.7) and (6.8), the postprocessor can eliminate the "crossterms" (i.e., $h_1^* h_2 S_2$ and $h_1 h_2^* S_2$).
- With the knowledge of the paths h_1 and h_2, the postprocessor in the receiver can recover the transmitted symbols. The receiver can estimate these paths by receiving the transmitted pilot symbols or some other means.

This transmit diversity scheme exploits two separate (and uncorrelated) paths h_1 and h_2 created by the two transmit antennas. This can be readily seen in (6.9)—if h_1 is in deep fade and near zero, h_2 is most likely still a viable channel. Hence, S_1

can still be recovered. The same is true also for S_2 in (6.10). Thus, the diversity gain achieved is due to transmitting the same symbol in two consecutive symbol periods.

Note that the received symbols can be written in the matrix format, that is,

$$\mathbf{R} = \mathbf{HS} + \mathbf{N} \tag{6.11}$$

where **R** is the received symbol matrix, **S** is the transmitted symbol matrix, **H** is the channel matrix, and **N** is the noise vector. Given this matrix representation, (6.5) and (6.6) can be rewritten as

$$\begin{bmatrix} R_1 & R_2 \end{bmatrix} = \begin{bmatrix} h_1 & h_2 \end{bmatrix} \begin{bmatrix} S_1 & -S_2^* \\ S_2 & S_1^* \end{bmatrix} + \begin{bmatrix} N_1 & N_2 \end{bmatrix} \tag{6.12}$$

Note that the transmitted symbol matrix **S** is used by the preprocessor to transmit the symbols. The channel matrix **H** is $M_r \times M_t$; since there is one receive antenna and two transmit antennas, the channel matrix **H** is 1×2.

This transmit diversity scheme was first proposed by Alamouti [7]. It is called open-loop because the *transmitter* does not need to know the channels before transmitting the symbols (although the *receiver* still needs to know the channels before recovering the symbols). The scheme is a special case of the orthogonal space-time block code (STBC). It is orthogonal because the cross terms cancel out in the recovery of the two symbols; it is called space-time because the transmitted symbol matrix **S** prearranges the symbols S_1 and S_2 in both space (across two antennas 1 and 2) and time (in consecutive symbol periods 1 and 2). This STBC is called Rate 1 because the transmitted symbol rate is the same as the original symbol rate. Another class of space-time coding is the space-time trellis code (STTC) [8].

6.3.2 Transmit Diversity: Open-Loop 2 × 2

A more elaborate transmit diversity scheme is one that has two transmit antennas and two receive antennas (i.e., a 2×2 system). Figure 6.8 shows that, in the open-loop 2×2 system, the preprocessor arranges consecutive symbols (from the transmitter) S_1 and S_2 and sends them out of two transmit antennas in the same

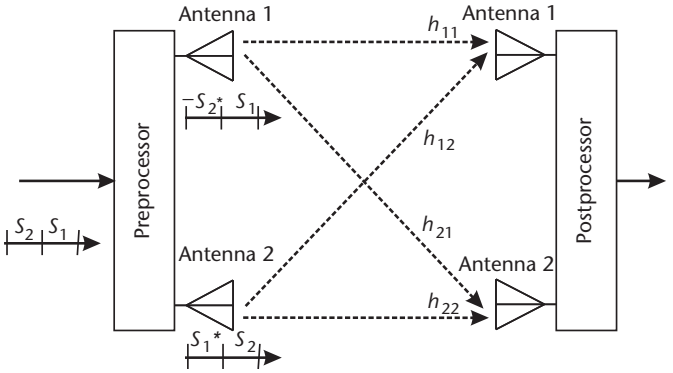


Figure 6.8 Open-loop 2×2 transmit diversity.

manner as that in the 1 × 2 system: S_1 and $-S_2^*$ from antenna 1 and S_2 and S_1^* from antenna 2.

However, at the receiving end, there are now four received symbols (because there are now two receive antennas). Specifically, the received symbol in the first symbol period, at receive antenna 1 is

$$R_{1,1} = S_1 h_{11} + S_2 h_{12} + N_{1,1} \tag{6.13}$$

and the received symbol in the second symbol period at receive antenna 1 is

$$R_{2,1} = -S_2^* h_{11} + S_1^* h_{12} + N_{2,1} \tag{6.14}$$

The received symbol in the first symbol period at receive antenna 2 is

$$R_{1,2} = S_1 h_{21} + S_2^* h_{22} + N_{1,2} \tag{6.15}$$

and the received symbol in the second symbol period at receive antenna 2 is

$$R_{2,2} = -S_2^* h_{21} + S_1^* h_{22} + N_{2,2} \tag{6.16}$$

Note that $R_{i,j}$ denotes received symbol in the ith symbol period at the jth receive antenna. Given these received symbols, the postprocessor can estimate and recover the transmitted symbols $\tilde{S}_1$ and $\tilde{S}_2$ by using the following equations:

$$\tilde{S}_1 = R_{1,1} h_{11}^* + R_{2,1}^* h_{12} + R_{1,2} h_{21}^* + R_{2,2}^* h_{22} \tag{6.17}$$

$$\tilde{S}_2 = R_{1,1} h_{12}^* - R_{2,1}^* h_{11} + R_{1,2} h_{22}^* - R_{2,2}^* h_{21} \tag{6.18}$$

To see how the postprocessor can recover the transmitted symbols $\tilde{S}_1$ and $\tilde{S}_2$, substitute (6.13) to (6.16) into (6.17). This results in:

$$\begin{aligned}
\tilde{S}_1 &= (h_{11}S_1 + h_{12}S_2 + N_{1,1})h_{11}^* + (-h_{11}^* S_2 + h_{12}^* S_1 + N_{2,1}^*)h_{12} + (h_{21}S_1 + h_{22}S_2 + N_{1,2})h_{21}^* \\
&\quad + (-h_{21}^* S_2 + h_{22}^* S_1 + N_{2,2}^*)h_{22} \\
&= S_1 |h_{11}|^2 + S_2 h_{12} h_{11}^* + N_{1,1} h_{11}^* - S_2 h_{11}^* h_{12} + S_1 |h_{12}|^2 + N_{2,1}^* h_{12} + S_1 |h_{21}|^2 \\
&\quad + S_2 h_{22} h_{21}^* + N_{1,2} h_{21}^* - S_2 h_{21}^* h_{22} + S_1 |h_{22}|^2 + N_{2,2}^* h_{22} \\
&= S_1 |h_{11}|^2 + S_1 |h_{12}|^2 + S_1 |h_{21}|^2 + S_1 |h_{22}|^2 + N_{1,1} h_{11}^* + N_{2,1}^* h_{12} + N_{1,2} h_{21}^* + N_{2,2}^* h_{22} \\
&= S_1 (|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2) + (N_{1,1} h_{11}^* + N_{2,1}^* h_{12} + N_{1,2} h_{21}^* + N_{2,2}^* h_{22})
\end{aligned} \tag{6.19}$$

Substituting equations (6.13) to (6.16) into (6.18) yields:

$$\begin{aligned}\tilde{S}_2 &= (h_{11}S_1 + h_{12}S_2 + N_{1,1})h_{12}^* - (-h_{11}^* S_2 + h_{12}^* S_1 + N_{2,1}^*)h_{11} + (h_{21}S_1 + h_{22}S_2 + N_{1,2})h_{22}^* \\ &\quad - (-h_{21}^* S_2 + h_{22}^* S_1 + N_{2,2}^*)h_{21} \\ &= S_1 h_{11}h_{12}^* + S_2|h_{12}|^2 + N_{1,1}h_{12}^* + S_2|h_{11}|^2 - S_1 h_{12}^* h_{11} - N_{2,1}^* h_{11} \\ &\quad + S_1 h_{21}h_{22}^* + S_2|h_{22}|^2 + N_{1,2}h_{22}^* + S_2|h_{21}|^2 - S_1 h_{22}^* h_{21} - N_{2,2}^* h_{21} \\ &= S_2|h_{12}|^2 + S_2|h_{11}|^2 + S_2|h_{22}|^2 + S_2|h_{21}|^2 + N_{1,1}h_{12}^* - N_{2,1}^* h_{11} + N_{1,2}h_{22}^* - N_{2,2}^* h_{21} \\ &= S_2(|h_{12}|^2 + |h_{11}|^2 + |h_{22}|^2 + |h_{21}|^2) + (N_{1,1}h_{12}^* - N_{2,1}^* h_{11} + N_{1,2}h_{22}^* - N_{2,2}^* h_{21})\end{aligned} \tag{6.20}$$

Again, one can see that:

- By using (6.17) and (6.18), the postprocessor can eliminate the cross terms: $S_2 h_{12}h_{11}^*$ and $S_2 h_{22}h_{21}^*$ in (6.19) and $S_1 h_{11}h_{12}^*$ and $S_1 h_{21}h_{22}^*$ in (6.20).
- With the knowledge of the paths h_{11}, h_{21}, h_{12}, and h_{22}, the postprocessor at the receiver can recover the transmitted symbols.

In the general case, the received symbols can also be written in the matrix format, that is,

$$\mathbf{R} = \mathbf{HS} + \mathbf{IN} \tag{6.21}$$

where **R** is the received symbol matrix, **S** is the transmitted symbol matrix, **H** is the channel matrix, and **N** is the noise matrix. **I** is the identity matrix. Given this matrix representation, (6.13) through (6.16) can be rewritten as

$$\begin{bmatrix} R_{1,1} & R_{2,1} \\ R_{1,2} & R_{2,2} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} S_1 & -S_2^* \\ S_2 & S_1^* \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} N_{1,1} & N_{2,1} \\ N_{1,2} & N_{2,2} \end{bmatrix} \tag{6.22}$$

Since there are now two receive antennas and two transmit antennas, the channel matrix **H** is a 2×2 matrix. Because the transmitter does not need to know the channels before transmitting the symbols, it is an open-loop orthogonal STBC system.

6.3.3 Transmit Diversity: Closed-Loop Antenna Selection

The antenna selection in closed-loop transmit diversity is akin to the antenna selection in receive diversity. In transmit antenna selection, the receive antenna system measures the signal powers of the M_t paths, determines the best path(s), and sends back to the transmitter information on such best path(s) [9]. A separate feedback channel is typically provisioned to carry such channel feedback information back to the transmitter. Figure 6.9 illustrates a scheme with just one receive antenna.

The transmit antenna system then only uses those antenna(s) that correspond to the best path(s). Examining Figure 6.9, one can easily see that this implementation is physically a mirror image of antenna selection in receive diversity (i.e., Figure 6.4). In fact, transmit antenna selection achieves an SNR performance similar to that of the receive antenna selection.

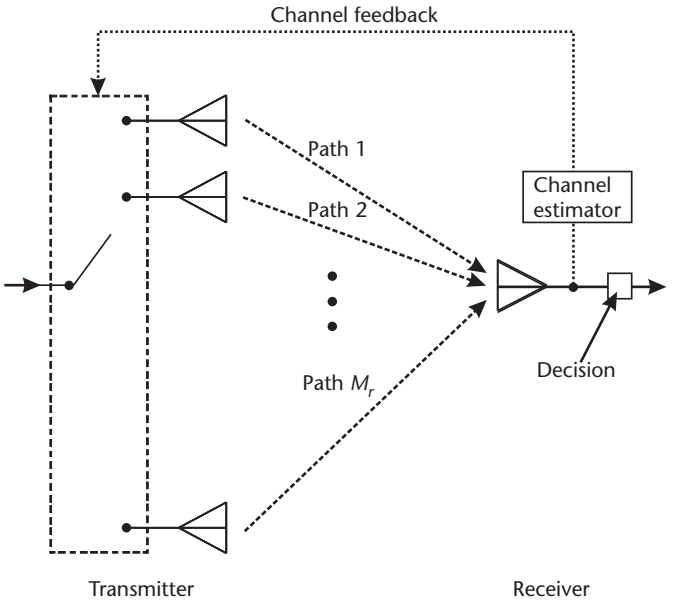


Figure 6.9 Closed-loop transmit diversity: antenna selection.

Example 6.1

A mobile wireless system uses transmit antenna selection shown in Figure 6.9 with three transmit antennas. It seeks to serve vehicles moving at speeds of up to 90 km/hour (or equivalently 25 m/s). What is the minimum rate of feedback if the system operates at 2.5 GHz? What is the minimum rate of feedback at 700 MHz?

At 2.5 GHz, the Doppler spread $2f_D$ is

$$2f_{D,2.5\text{ GHz}} = 2v\frac{f}{c} = 2(25\text{ m/s})\left(\frac{2.5\times10^9\text{Hz}}{3\times10^8\text{m/s}}\right) = 416.67\text{ Hz}$$

The channel coherence time T_c is

$$T_c \approx \frac{1}{2f_D} = \frac{1}{416.67\text{ Hz}} = 0.0024\text{s} = 2.4\text{ ms}$$

The coherence time is the period of time over which the channel impulse response is approximately the same. Thus, the minimum rate of feedback has to be as fast as the rate that the channel changes. In other words, the minimum rate of feedback has to be at least once every T_c, or $1/T_c$.

$$\frac{1}{T_c} \approx 2f_D = \frac{1}{0.0024\text{s}} = 416.67\text{ Hz}$$

Because it takes at least 2 bits to represent three combinations (three antennas), the feedback channel has to transmit 2 bits of data every 0.0024 second. Thus, the minimum *data* rate of feedback[4] becomes:

$$\frac{2\text{b}}{\text{feedback}} \times \frac{1 \text{ feedback}}{0.0024\text{s}} = 833.33 \frac{\text{b}}{\text{s}}$$

At 700 MHz, the Doppler spread $2f_D$ is

$$2f_{D,700\text{ MHz}} = 2v\frac{f}{c} = 2(25 \text{ m/s})\left(\frac{700 \times 10^6 \text{Hz}}{3 \times 10^8 \text{m/s}}\right) = 116.67 \text{ Hz}$$

The channel coherence time T_c is

$$T_c \approx \frac{1}{2f_D} = \frac{1}{116.67 \text{ Hz}} = 0.00857\text{s} = 8.57 \text{ ms}$$

The minimum rate of feedback has to be at least once every T_c, or $1/T_c$, which is

$$\frac{1}{T_c} \approx 2f_D = \frac{1}{0.00857\text{s}} = 116.67 \text{ Hz}$$

The minimum *data* rate of feedback then is:

$$\frac{2\text{b}}{\text{feedback}} \times \frac{1 \text{ feedback}}{0.00857\text{s}} = 233.33 \frac{\text{b}}{\text{s}}$$

Therefore, we see that the minimum rate of feedback is lower in systems with lower carrier frequencies.

6.3.4 Transmit Diversity: Closed-Loop Precoding

Although the transmit antenna selection is simple to implement, the system does not utilize all available transmit antennas and all paths. A closed-loop transmit diversity scheme that utilizes all transmit antennas is precoding [10]. Figure 6.10 shows the general precoding scheme.

As shown in the figure, the transmitted symbol vector $\mathbf{x}$ ($M \times 1$) is first precoded by the precoding matrix $\mathbf{E}$ ($M_t \times M$). Then the precoded symbol vector $\mathbf{Ex}$ ($M_t \times 1$) is degraded by the channel matrix $\mathbf{H}$ ($M_r \times M_t$). The degraded symbol vector at the receive antennas is $\mathbf{HEx}$ ($M_r \times 1$). A postcoding matrix $\mathbf{F}$ ($M \times M_r$) is applied

4. In actual systems, the data rate of channel feedback is often several times of the minimum data rate.

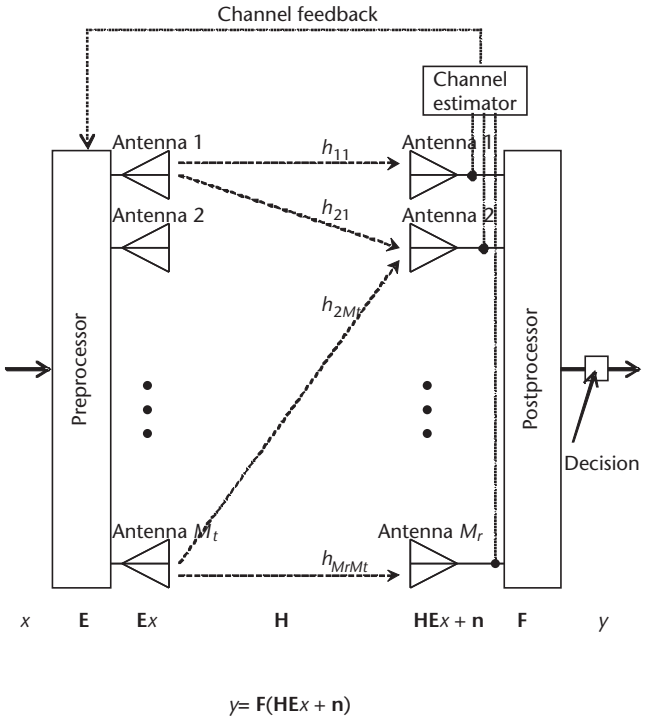


Figure 6.10 Closed-loop transmit diversity: precoding.

to the degraded symbol vector **HEx** and to the noise vector **n** ($M_r \times 1$). Thus, the received symbol vector **y** ($M \times 1$) can be written as

$$\mathbf{y} = \mathbf{F}(\mathbf{HEx} + \mathbf{n}) \tag{6.23}$$

If the system transmits one symbol at a time ($M = 1$), then (6.23) can be rewritten as:

$$y = \mathbf{F}(\mathbf{HE}x + \mathbf{n}) \tag{6.24}$$

For example, if a system has two receive antennas ($M_r = 2$) and three transmit antennas ($M_t = 3$), then (6.24) becomes

$$\begin{aligned} y &= \begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} x + \begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \\ &= \begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} h_{11}e_1 + h_{12}e_2 + h_{13}e_3 \\ h_{21}e_1 + h_{22}e_2 + h_{23}e_3 \end{bmatrix} x + \left(f_1 n_1 + f_2 n_2\right) \\ &= \left\{f_1(h_{11}e_1 + h_{12}e_2 + h_{13}e_3) + f_2(h_{21}e_1 + h_{22}e_2 + h_{23}e_3)\right\} x + \left(f_1 n_1 + f_2 n_2\right) \end{aligned} \tag{6.25}$$

The process is as follows: In each symbol period,

- The receiver estimates the channel matrix **H** (by using the transmitted pilot symbols or some other means).
- Based on the channel matrix **H**, the receiver computes the optimal precoding matrix **E** (i.e., coefficients e_i) and the optimal postcoding matrix **F** (i.e., coefficients f_i) such that the resultant SNR of received signal y is maximized.
- The receiver communicates the optimal precoding matrix **E** (i.e., coefficients e_i) to the transmitter via the feedback channel.
- The preprocessor uses **E** to precode x.

Note that the receiver cannot arbitrarily set f_i to be large because doing so will enhance the noise term [second term in (6.25)]. Also, the receiver cannot arbitrarily set e_i to be large because RF amplifiers at the transmitting end may be clipped if symbol energies are too high. Given **H**, computing the optimal **E** and **F** is an exercise in linear algebra and optimization.

The receiver estimates the channel matrix **H**. However, to save bandwidth on the feedback channel, the receiver does not send the entire channel matrix **H** back to the transmitter. Rather, the receiver itself can quickly compute **E** based on **H** and sends **E** (i.e., coefficients e_i) back to the transmitter. To further minimize bandwidth requirements on the feedback channel, the transmitter and the receiver can first agree on a finite set (or a "codebook") of **E** to use (e.g., 64 instances of **E**). After the receiver estimates **H**, the receiver computes the optimal **E** and then picks a specific **E** out of the set that is closest to the optimal **E**. The receiver then sends the "index" of the specific **E** to use back to the transmitter. If there is a total of 64 possible instances of **E** to use, then the receiver only has to send 6 bits of information on the feedback channel.

In addition, in TDD, the base station can directly estimate the channel matrix of the downlink through reciprocity. The base station does so by using *channel sounding*. In channel sounding, the mobile transmits a known sounding signal on the uplink. Based on the received sounding signal, the base station can obtain knowledge of the downlink channel through reciprocity. In IEEE 802.16e, the mobile transmits the sounding signal in a special sounding zone in the uplink subframe.

6.3.5 Remarks

By now readers recognize that open-loop 2×1 STBC and open-loop 2×2 STBC code data across both space (antennas) and time (symbol periods). To be sure, these spatial diversity techniques (e.g., STBC) leverage not only diversity in space, but also diversity in *time* by introducing redundancy across time through the structure of the code [4, 8].

In addition to diversity in time, spatial diversity may also take advantage of diversity in *frequency*. In particular, adjacent subcarriers can be used for diversity (instead of adjacent symbol periods) because adjacent subcarriers are orthogonal and have correlated channels (in OFDM). The resulting technique is called *space frequency block code* (SFBC) [11]. A simple example of space frequency coding is to adopt the Alamouti code over two subcarriers for two transmit antennas, all in one OFDM symbol [12]. Other spatial techniques that use diversity in frequency include space-frequency interleaving [6].

6.4 Spatial Multiplexing

Rather than focusing on link reliability, spatial multiplexing, also known as multiple input/multiple output (MIMO),[5] focuses on increasing the bit rate. Similar to OFDM's parallel transmission in frequency, spatial multiplexing is a method of increasing bit rate by transmitting parallel, unique symbol streams in space using multiple transmit antennas and multiple receive antennas. In other words, multiple transmit and receive antennas create parallel transmission paths over the air to increase bit rate. The advantage, of course, is that the end-to-end bit rate can be increased by up to $\min(M_t, M_r)$.

The motivation for the growing adoption of spatial multiplexing is the requirement of higher bit rates in broadband wireless systems. In theory, one can increase the bit rate by increasing bandwidth (hertz) and/or increasing bandwidth efficiency (bps/Hz). However, in wireless systems, bandwidth has always been a constraint due to regulatory and other reasons, and increasing bandwidth efficiency necessitates higher-order modulation and large constellation size that require higher SNR, which is also a constraint in wireless systems. Spatial multiplexing has emerged as a dominant way of increasing bit rate without the need for additional bandwidth and SNR [3, 13], but it does require additional antennas to implement.

Figure 6.11 shows the general spatial multiplexing scheme. At the transmitter, the serial-to-parallel converter converts the original symbol stream (at the symbol rate R_s) into M symbol streams, each running at the symbol rate R_s. The M symbol streams are transmitted by the M_t transmit antennas. This way an aggregate symbol rate of MR_s can be supported by the channels.

At the receiver, there are M_r receive antennas. As a result, there are M_tM_r paths between the transmitter and the receiver. Each path has a transfer function h_{ij}, which denotes the transfer function from transmit antenna j to receive antenna i. The postprocessor attempts to recover the symbol streams, and the parallel-to-serial converter converts the M received symbol streams into a single symbol stream.

More specifically in Figure 6.11, the serial-to-parallel converter converts the original symbol stream into M symbol streams. The transmitted symbol vector $\mathbf{x}$ $(M \times 1)$ is first preprocessed by the preprocessing matrix $\mathbf{E}$ $(M_t \times M)$. Then the preprocessed symbol vector $\mathbf{Ex}$ $(M_t \times 1)$ is sent by the M_t transmit antennas and is degraded by the channel matrix $\mathbf{H}$ $(M_r \times M_t)$. The degraded symbol vector at the receive antennas is $\mathbf{HEx}$ $(M_r \times 1)$. A postprocessing matrix $\mathbf{F}$ $(M \times M_r)$ is applied to the degraded symbol vector $\mathbf{HEx}$ and to the noise vector $\mathbf{n}$ $(M_r \times 1)$. Thus, the received symbol vector $\mathbf{y}$ $(M \times 1)$ can be written as

$$\mathbf{y} = \mathbf{F}(\mathbf{HEx} + \mathbf{n}) \tag{6.26}$$

For example, if a system uses three parallel symbol streams and has three receive antennas ($M_r = 3$) and three transmit antennas ($M_t = 3$), then (6.26) can be rewritten as

5. MIMO can also be more generally defined as a system with multiple transmit antennas ("input") and multiple receive antennas ("output"). The terms *input* and *output* are with respect to the wireless channel.

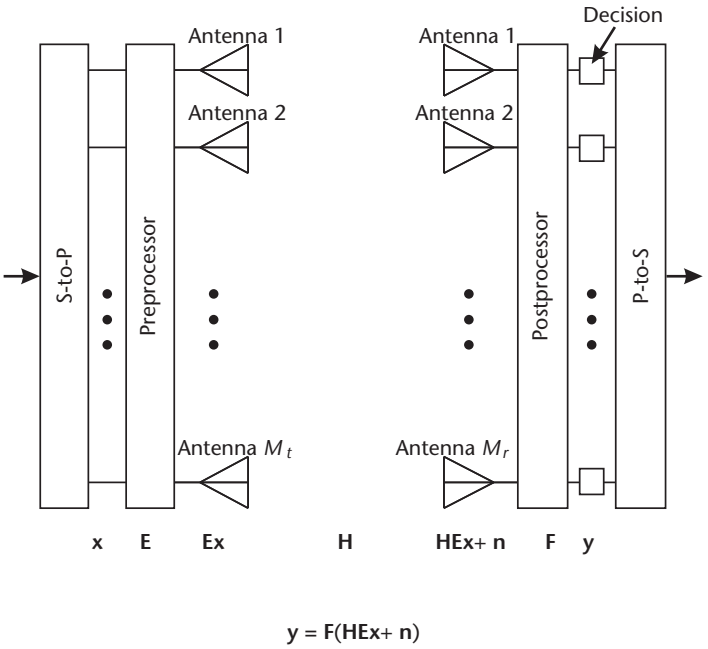


Figure 6.11 Spatial multiplexing.

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} e_{11} & e_{12} & e_{13} \\ e_{21} & e_{22} & e_{23} \\ e_{31} & e_{32} & e_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \quad (6.27)$$

So the question now becomes: how do we derive **E** and **F** in order to turn all the M_tM_r paths into a bank of M parallel paths that can simultaneously transmit M symbol streams? It turns out that, under certain conditions, a matrix can be "diagonalized" by using singular value decomposition (SVD). Diagonalizing the channel matrix **H** ($M_r \times M_t$) yields

$$\mathbf{H} = \mathbf{UDV}^{\mathrm{H}} \quad (6.28)$$

where **D** ($M \times M$) is a diagonal matrix, i.e., (i.e., **D** = diag[d_{11}, d_{22},...d_{MM}]). $\mathbf{V}^H$ denotes the Hermitian transpose (i.e., conjugate transpose) of **V**, and **V** ($M_t \times M$) is a unitary matrix such that

$$\mathbf{VV}^{\mathrm{H}} = \mathbf{V}^{\mathrm{H}}\mathbf{V} = \mathbf{I} \quad (6.29)$$

U ($M_r \times M$) is also a unitary matrix such that

$$\mathbf{UU}^{\mathrm{H}} = \mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I} \quad (6.30)$$

If the channel matrix **H** can be diagonalized as above, then all we have to do is to set the preprocessing matrix **E** to **V** and the postprocessing matrix **F** to $\mathbf{U}^{\mathrm{H}}$. Then the M_tM_r paths can be transformed into a bank of M parallel paths. In other words,

$$\mathbf{y} = \mathbf{F}(\mathbf{HEx} + \mathbf{n}) = \mathbf{U}^{\mathrm{H}}(\mathbf{UDV}^{\mathrm{H}}\mathbf{Vx} + \mathbf{n}) = \mathbf{U}^{\mathrm{H}}\mathbf{UDV}^{\mathrm{H}}\mathbf{Vx} + \mathbf{U}^{\mathrm{H}}\mathbf{n} = \mathbf{Dx} + \mathbf{U}^{\mathrm{H}}\mathbf{n} \tag{6.31}$$

Equation (6.31) constitutes the basic principle behind spatial multiplexing. Because the transmit symbol vector **x** is now only multiplied by the diagonal matrix **D**, the diagonal matrix **D** now serves as a bank of M parallel paths for the transmitted symbol vector **x**. For example, if a system has three receive antennas ($M_r = 3$) and three transmit antennas ($M_t = 3$), then (6.31) can be rewritten as

$$\mathbf{y} = \mathbf{Dx} + \mathbf{U}^{\mathrm{H}}\mathbf{n} \tag{6.32}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \mathbf{U}^{\mathrm{H}} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{6.33}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} d_{11}x_1 \\ d_{22}x_2 \\ d_{33}x_3 \end{bmatrix} + \mathbf{U}^{\mathrm{H}} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{6.34}$$

Because **D** is a diagonal matrix, SVD eliminates all the cross terms in the channel matrix so that **y** has the benefit of M separate, parallel paths over the air scaled by the path weights d_{ii} (shown in Figure 6.12).

Note that this arrangement is necessarily closed-loop because the receiver has to communicate **V** back to the transmitter. The process is as follows: In each symbol period,

- The receiver estimates the channel matrix **H**.

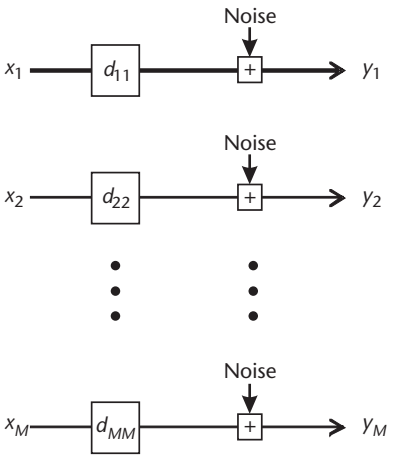


Figure 6.12 Equivalent parallel paths in spatial multiplexing.

- Based on the channel matrix **H**, the receiver performs SVD and obtains **U**, **V**, and **D**.
- The receiver communicates the preprocessing matrix **V** to the transmitter via the feedback channel.
- The preprocessor uses **V** to preprocess **x**.

In addition, if the receiver can communicate **D**, which has the path weights (i.e., d_{ii}) of the M parallel paths, back to the transmitter, the transmitter can use this knowledge to allocate bits and power to each of the M parallel paths in order to maximize bit rate. The transmitter can do so by allocating more bits and more power to those paths that have larger gains [14].

Therefore, the overall result is that the system can use closed-loop feedback to: (1) deconstruct the channel matrix and effectively reconstruct a set of parallel streams over the air that can carry more data, and (2) dynamically allocate bits and power to each parallel path to maximize the bit rate.

6.5 MIMO-OFDM

It turns out that MIMO works well with OFDM because an underlying assumption of spatial multiplexing is a frequency flat channel. Recall from Chapter 2 that multipath contributes to delay spread in the time domain, and delay spread in the time domain translates into a frequency selective channel in the frequency domain. The terrestrial cellular environment, especially in urban areas, is often characterized by frequency selective channels. OFDM converts a frequency selective channel into a series of frequency flat channels by dividing a wideband channel into many narrowband channels. Thus, each narrowband subcarrier experiences a frequency flat channel. When the system combines both MIMO and OFDM in this fashion, it is called MIMO-OFDM.

The reason why frequency flat channels underlie the assumptions behind spatial multiplexing is due to the channel matrix **H** in (6.26). The elements of the channel matrix **H** are the h_{ij}, and each h_{ij} is a single number. This means that a path represented by a specific h_{ij} has a single value characterizing the entire channel response of that path (i.e., frequency flat). Thus, in order to compute (6.26), the channels have to be sufficiently flat so they can be represented by the h_{ij} in **H**—OFDM's narrow subcarriers make that possible [9]. In contrast, MIMO is relatively more difficult to implement in traditional DSSS because there is ISI over such a wide bandwidth. MIMO-OFDM is an area of ongoing research [4, 15].

6.6 Beamforming

Adaptive beamforming is a spatial technique by which multiple antennas are used to focus the antenna beam (i.e., directivity) to discriminate between the desired signal and interfering signals. Figure 6.13 shows M_r antennas that are separated by distance d. Incident rays impinge upon the antennas at an angle θ.

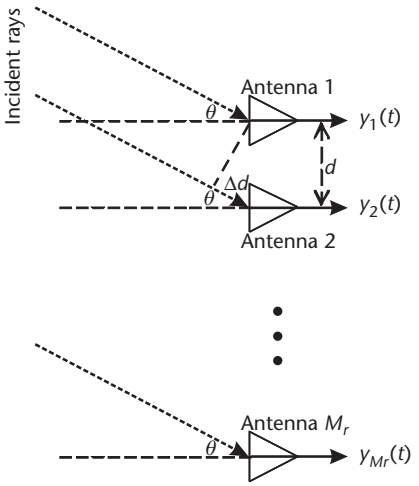


Figure 6.13 An array of M_r antennas used for beamforming.

Examining the geometry, one recognizes that the second ray travels an extra distance of Δd as compared to the first ray, that is,

$$\Delta d = d\cos\left(\frac{\pi}{2} - \theta\right) = d\sin\theta \tag{6.35}$$

and is delayed by Δt as compared to the first ray, that is,

$$\Delta t = \frac{\Delta d}{c} = \frac{d\sin\theta}{c} \tag{6.36}$$

If $y_1(t)$ is the received signal arriving at the first antenna and $y_2(t)$ is the received signal arriving at the second antenna, then the second received signal can written in terms of the first received signal as

$$y_2(t) = y_1(t)e^{-j2\pi f\Delta t} = y_1(t)e^{-j2\pi\left(\frac{c}{\lambda}\right)\left(\frac{d\sin\theta}{c}\right)} = y_1(t)e^{-j2\pi\left(\frac{d\sin\theta}{\lambda}\right)} \tag{6.37}$$

This expression is based on the narrowband assumption, which states that the bandwidth of the received signal is narrow so that it stays constant during Δt (i.e., $W << 1/\Delta t$). In general, if there are M_r equally spaced (by d) receive antennas arranged in a linear fashion, then the mth received signal collected by the mth antenna is

$$y_m(t) = y_1(t)e^{-j2\pi\left(\frac{d\sin\theta}{\lambda}\right)(m-1)} \tag{6.38}$$

Looking at (6.38), we observe that $y_1(t)$, $y_2(t)$,... $y_{Mr}(t)$ are the same except for the extra phase shift, which depends on the antenna separation d and the angle of

arrival θ of the rays [3]. d is fixed based on the physical placement of the antennas. Thus, if the angle of arrival θ of the desired signal is known, it is conceivable that the system can artificially steer the beam (generated by the multiple antennas) toward θ by compensating for the phase shifts in the received signals after the antennas.

Figure 6.14 shows an example. Suppose there are four antennas separated by $d = \lambda/2$ in an array. A desired signal is coming in at $\theta = \pi/4$. In order to steer the receive beam in the direction of $\theta = \pi/4$, the system needs to compensate for the following phase shifts in the four received signals $y_1(t)$, $y_2(t)$, $y_3(t)$, and $y_4(t)$, after the antennas:

$$y_1(t) = y_1(t)(1) \tag{6.39}$$

$$y_2(t) = y_1(t)e^{-j2\pi\left(\frac{\sin\pi/4}{2}\right)(2-1)} = y_1(t)e^{-j\pi\left(\frac{\sqrt{2}}{2}\right)} \tag{6.40}$$

$$y_3(t) = y_1(t)e^{-j2\pi\left(\frac{\sin\pi/4}{2}\right)(3-1)} = y_1(t)e^{-j\pi\sqrt{2}} \tag{6.41}$$

$$y_4(t) = y_1(t)e^{-j2\pi\left(\frac{\sin\pi/4}{2}\right)(4-1)} = y_1(t)e^{-j\pi\left(\frac{3\sqrt{2}}{2}\right)} \tag{6.42}$$

Note that changing the phase shifts of the received signals only changes the *direction* of the beam, not the shape of the beam. If the amplitudes of the received signals are also changed, then the *shape* of the beam can be changed. To change both phases shifts and amplitudes of the received signals, the system can multiply the received signals by a set of complex weights $w_1, w_2,\ldots, w_{Mr}$—an operation that is similar to multiplication by the postprocessing matrix **F** discussed previously [5].

In addition to steering the beam in the direction of the desired signal, adaptive beamforming can also "null" the beam in the direction of the undesired signal (e.g., cochannel interferers). The nulling can be done again by optimizing the complex weights to enhance the beam in the direction of the desired signal and to null the beam in the direction of the undesired signal. As a result, the SINR of the desired signal can increase. The same concept can be applied not only to the receive beam

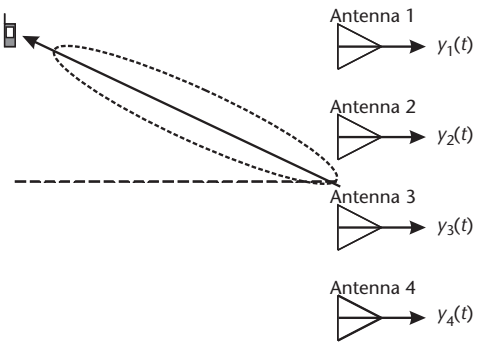


Figure 6.14 An illustration of beamforming.

but also to the transmit beam. By performing similar multiplication operations, the transmit beam can be steered toward the direction to the desired user and be nulled in the direction of other cochannel users.

In order to enhance the beam and null the beam in different directions, the angles of arrival of desired and undesired signals have to be estimated. Popular algorithms such as the Multiple Signal Classification (MUSIC) algorithm can be used to estimate the angle of arrival of a desired signal or an interfering signal [16, 17].

6.7 System Design Issues

The three types of spatial techniques discussed in this chapter, spatial diversity, spatial multiplexing, and beamforming, can each achieve different types of gains (e.g., spatial diversity gain, spatial multiplexing gain, and cochannel interference reduction), but these techniques are not meant to operate in all wireless environments. System designers often encounter two common considerations: bit rate versus reliability and low spatial correlation versus high spatial correlation. We address these two considerations in the context of frequency flat channels, which are typically the case in OFDM and OFDMA.

In terms of bit rate versus reliability, if high reliability at an average bit rate is important (e.g., for public safety systems), then spatial diversity schemes are preferred. If high bit rate and high bandwidth efficiency are the design goals (e.g., for fixed broadband wireless systems), then spatial multiplexing schemes are preferred [15]. Such a constraint exists because there is a fundamental tradeoff between how much spatial diversity gain and how much spatial multiplexing gain can be extracted by any one technique; maximizing one gain may not necessarily maximize the other [12]. In other words, one cannot simultaneously perform spatial multiplexing and achieve *full* spatial diversity gain [18].

In terms of spatial correlation of paths, low spatial correlation of paths works well with both spatial diversity and spatial multiplexing. The paths typically have low spatial correlation in NLOS situations with rich scattering. In addition, sufficient antenna separation is also important to attain low spatial correlation. In fact, with rich scattering (and sufficient antenna separation), the channel matrix $\mathbf{H}$ is i.i.d. for all frequencies [3]. Rich scattering is particularly beneficial to spatial multiplexing because it tends to increase the angle spread, which raises the rank[6] of the channel matrix and thus the spatial multiplexing gain [4].

On the other hand, high spatial correlation of paths does not work well with spatial multiplexing because of the difficulty of deriving M parallel paths out of paths that have dependencies on each other. High spatial correlation of paths can work with spatial diversity if SNR is high [15]. The paths typically have high spatial correlation in LOS situations (which tend to result in high SNR). Figure 6.15 summarizes the different situations with regard to spatial correlation.

In addition, beamforming can be used to reduce cochannel interference and hence increase SINR. However, by focusing the beam in a particular direction, beamforming tends to perceive less scattering. Less scattering can reduce delay

6. Recall that, in linear algebra, the rank of a matrix is the number of linearly independent columns (or rows) present in that matrix.

	Low spatial correlation of paths (e.g., NLOS)	High spatial correlation of paths (e.g., LOS)
Spatial multiplexing gain	High	Low
Spatial diversity gain	High	OK if SNR high

Figure 6.15 Spatial correlation of paths, spatial multiplexing gain, and spatial diversity gain.

spread but has impact on spatial multiplexing and/or spatial diversity—another tradeoff that needs to be considered.

Moreover, mobility becomes an important consideration in choosing between open-loop systems and closed-loop systems. In IEEE 802.16e, for example, the feedback delay is two frames; if the frame duration is 5 ms, then the feedback delay is 10 ms [19]. At high speeds, variations in the channel become fast and the coherence time of the channel becomes short; the channel changes so quickly that it is difficult for the transmitter to obtain timely knowledge of the channel through channel feedback. Thus, open-loop techniques can be considered in mobility scenarios because they do not require the transmitter to have knowledge of the channel to operate [19].

References

[1] Sternad, M., et al., "Towards Systems Beyond 3G Based on Adaptive OFDMA Transmission," *Proceedings of the IEEE*, Vol. 95, No. 12, 2007, pp. 2432–2455.

[2] Freeman, R. L., *Radio System Design for Telecommunications*, New York: John Wiley & Sons, 2007.

[3] Paulraj, A. J., et al., "An Overview of MIMO Communications—A Key to Gigabit Wireless," *Proceedings of the IEEE*, Vol. 92, No. 2, 2004, pp. 198–218.

[4] Bolcskei, H., "MIMO-OFDM Wireless Systems: Basics, Perspectives, and Challenges," *IEEE Wireless Communications*, Vol. 13, No. 4, 2006, pp. 31–37.

[5] Mietzner, J., et al., "Multiple-Antenna Techniques for Wireless Communications—A Comprehensive Literature Survey," *IEEE Communications Surveys and Tutorials*, Vol. 11, No. 2, 2009, pp. 87–105.

[6] Salvekar, A., et al., "Multiple-Antenna Technology in WiMAX Systems," *Intel Technology Journal*, Vol. 8, No. 3, 2004, pp. 229–239.

[7] Alamouti, S. M., "A Simple Transmit Diversity Technique for Wireless Communications," *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 8, 1998, pp. 1451–1458.

[8] Tarokh, V., N. Seshadri, and A. R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction," *IEEE Trans. on Information Theory*, Vol. 44, No. 2, 1998, pp. 744–765.

[9] Stuber, G. L., et al., "Broadband MIMO-OFDM Wireless Communications," *Proceedings of the IEEE*, Vol. 92, No. 2, 2004, pp. 271–294.

[10] Yaghoobi, H., "Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN," *Intel Technology Journal*, Vol. 8, No. 3, 2004, pp. 201–212.

[11] Lee, B. G., and S. Choi, *Broadband Wireless Access and Local Networks: Mobile WiMAX and WiFi*, Norwood, MA: Artech House, 2008.

[12] Zhang, W., X. -G. Xia, and K. B. Letaief, "Space-Time/Frequency Coding for MIMO-OFDM in Next Generation Broadband Wireless Systems," *IEEE Wireless Communications*, Vol. 14, No. 3, 2007, pp. 32–43.

[13] Bolcskei, H., D. Gesbert, and A. J. Paulraj, "On the Capacity of OFDM-Based Spatial Multiplexing Systems," *IEEE Trans. on Communications*, Vol. 50, 2002, pp. 225–234.

[14] Ha, J., et al., "LDPC Coded OFDM with Alamouti/SVD Diversity Technique," *Wireless Personal Communications*, Vol. 23, 2002, pp. 183–194.

[15] Yang, H., "A Road to Future Broadband Wireless Access: MIMO-OFDM-Based Air Interface," *IEEE Communications*, Vol. 43, No. 1, 2005, pp. 53–60.

[16] Chryssomallis, M., "Smart Antennas," *IEEE Antennas and Propagation*, Vol. 42, No. 3, 2000, pp. 129–136.

[17] Codara, L. C., "Application of Antenna Arrays to Mobile Communications, Part II: Beam-Forming and Direction-of-Arrival Considerations," *Proceedings of the IEEE*, Vol. 85, No. 8, 1997, pp. 1195–1245.

[18] Zheng, L., and D. N. C. Tse, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple-Antenna Channels," *IEEE Trans. on Information Theory*, Vol. 49, No. 5, 2003, pp. 1073–1096.

[19] Li, Q., et al., "Advancement of MIMO Technology in WiMAX: From IEEE 802.16d/e/j to 802.16m," *IEEE Communications*, Vol. 47, No. 6, 2009, pp. 100–107.

Selected Bibliography

Li, Q., et al., "MIMO Techniques in WiMAX and LTE: A Feature Overview," *IEEE Communications,* Vol. 48, No. 5, 2010.

Paulraj, A. J., R. U. Nabar, and D. A. Gore, *Introduction to Space-Time Wireless Communications*, Cambridge, U.K.: Cambridge University Press, 2003.

CHAPTER 7

Medium Access Control: Architecture and Data Plane

7.1 MAC Architecture

In a generic communication architecture (e.g., Open System Interconnection or OSI [1]), the medium access control (MAC) layer regulates a higher layer's access to the services provided by the physical layer. Thus, if a higher-layer protocol (i.e., Internet Protocol or IP) wants to transmit a packet, it would send such a transmission request to MAC. MAC would process the request and employ the physical layer's transmission service to deliver the packet. Of course, in readying the packet for its physical transmission, MAC needs to perform a number of functions, including segmenting the packet from the higher layer and scheduling the segments for transmission. At the receiver, the reverse processes takes place. In a larger sense, MAC ensures that multiple users all have access to the limited resources provided by the physical layer.

In a broadband wireless network, the MAC entity performs similar functions. Figure 7.1 shows a high-level view of MAC as specified by IEEE 802.16 [2]. In particular, the MAC layer is structurally divided into three sublayers:

- *Convergence sublayer:* This sublayer is the main interface to higher-layer protocols such as IP and Ethernet.
- *Common part sublayer:* This sublayer performs functions such as MAC Protocol Data Unit (PDU) assembly, scheduling, and network entry.
- *Security sublayer:* This sublayer performs security functions such as encryption.

As shown in Figure 7.1, a PDU is a unit of data that is logically delivered from one peer entity to another peer entity at the same protocol layer, whereas a service data unit (SDU) is a unit of data that is physically exchanged between adjacent protocol layers [2].

Figure 7.2 shows a more detailed view of the structure of the MAC layer. In this figure, the three sublayers of MAC are still present, but the functions are further divided into two planes: the data plane and the control plane. The *data plane*

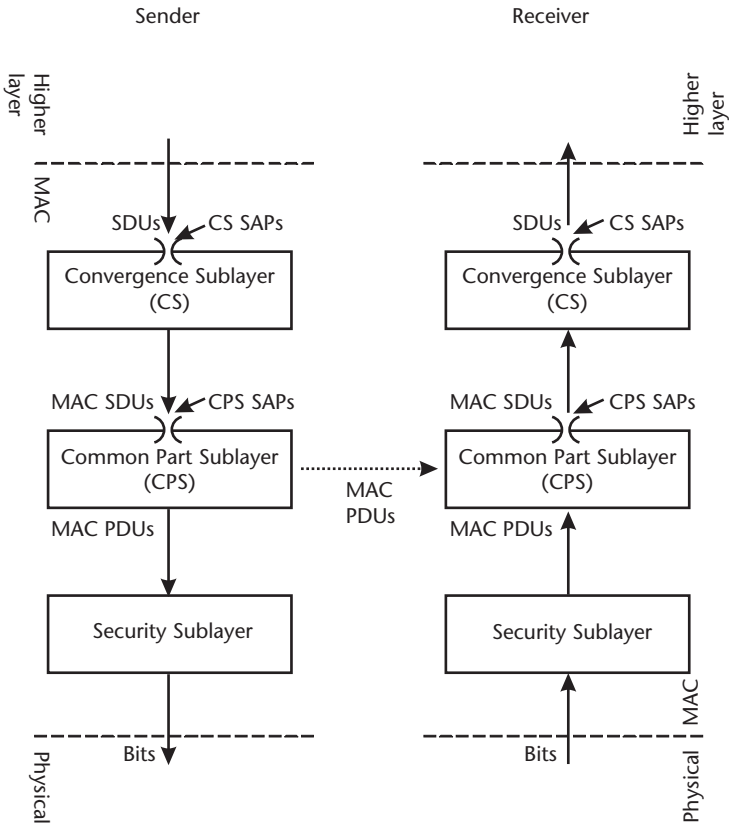


Figure 7.1 MAC sublayers. The term SAP stands for service access point.

contains those functions involved in creating and processing MAC PDUs in the data path. The *control plane* has all the other functions involved in the control, signaling, and management of radio resources. The activities carried out on the data plane are sometimes called fast path activities because they have to be performed in real time, whereas the activities on the control plane are sometimes called slow path activities. The control plane functions have to do with policies, while the data plane functions relate to execution [3]. In addition, these MAC functions can be categorized into upper and lower. The upper functions include network discovery and entry, connection management, mobility management, multicast and broadcast service (MBS) [4], and others. The lower functions include the more traditional functions of MAC such as physical layer control functions (e.g., ranging, power control, etc.), scheduling, control signaling, QoS, and others [5, 6].

In this chapter, we discuss these functions on the data plane and how data (in the form of SDUs and PDUs) traverse through the different sublayers of MAC. Chapters 8 and 9 discuss selected control and supervisory functions on the control plane. Specifically, Chapter 8 examines those functions in the lower part of MAC, and Chapter 9 examines those functions in the upper part of MAC.

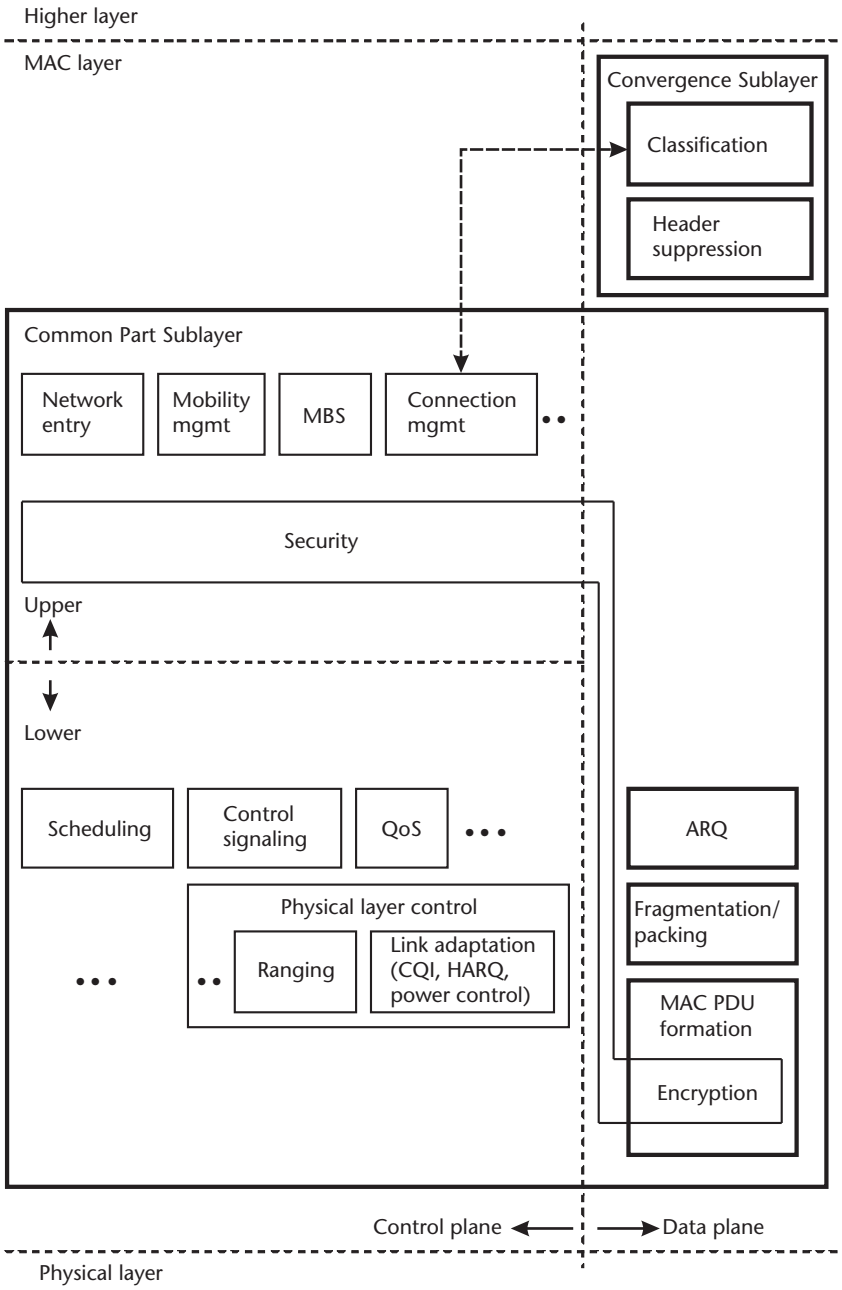


Figure 7.2 MAC protocol structure. Note that the boldfaced parts constitute the data plane. (*After:* [5, 6].)

7.2 Convergence Sublayer

The convergence sublayer serves as the main interface to higher-layer protocols. At the transmitter, the convergence sublayer accepts a data packet, called the SDU, from a higher layer, performs some initial processing on it, and then passes it on to the common part sublayer for further processing. At the receiver, the reverse is performed.

The convergence sublayer is also known as the *service-specific* convergence sublayer in that its processing functions depend on which higher-layer *service* (i.e., IP or Ethernet) sent the SDU. In fact, different convergence sublayer specifications exist to communicate with different higher-layer protocols. The main higher-layer protocols supported by the IEEE 802.16 suite standards are Ethernet [7] and IP. Depending on which higher-layer protocol sent the SDU, the convergence sublayer performs two functions: address mapping (classification) and header suppression.

7.2.1 Address Mapping (Classification)

At the transmitter, the convergence sublayer maps the original (higher-layer network) addresses used by the SDU to the system-specific addresses used by MAC. For example, if the higher-layer protocol is IP, then the IP packet would have a source IP address and a destination IP address. Using these IP addresses and their host protocol, the convergence sublayer maps them to the addresses specifically used by MAC. Two addresses are important in this respect: the connection identifier (CID) and service flow identifier (SFID).

A *connection* is a logical, unidirectional association between the base station and the mobile. It is important to note that a connection associated with the downlink is different from the connection associated with the uplink. This is because a connection is present in one direction only. There are two types of connections: transport connections (for data) and management connections (for control/signaling), and a connection identifier (CID) is used to define a single transport connection. An exception is that a CID can define a pair of uplink and downlink management connections because signaling exchanges are assumed to be bidirectional. A CID is an actual number that is 16-bits long.

A *service flow*, on the other hand, is a unidirectional stream of MAC SDUs. This flow takes place on a specific connection between the base station and the mobile. In fact, there is one connection for each service flow. The most important aspect of a service flow is that a service flow has a set of relevant QoS parameters associated with it. To identify a single service flow, the system uses the service flow identifier (SFID). The SFID is an actual number that is 32-bits long.

It is important to recognize that a single mobile can have several service flows provisioned. The reason may be due to the fact that the mobile has several different applications open, each requiring a different set of QoS parameters. As a result, that same mobile would have several transport connections and would request bandwidth on a per connection basis [2].

7.2.2 Header Suppression

To reduce excessive amounts of overhead, the MAC layer may implement a feature called payload header suppression (PHS). PHS is essentially a data compression scheme that is applied to the headers of successive SDUs entering the MAC layer. The scheme takes advantage of the fact that successive SDUs coming from a higher-layer protocol often have repetitive parts in them. For example, a series of unicast video IP packets will have identical source IP addresses, as well as identical destination IP addresses. At the transmitter, the convergence sublayer uses PHS to delete (suppress) these repetitive parts; at the receiver, the convergence sublayer uses the

same feature to reinstate (restore) the suppressed parts and deliver the complete packets to the higher layer. The goal is, of course, to minimize the amount of overhead traveling over the physical layer.

In performing PHS, the convergence sublayer at the transmitter accepts SDUs from the higher layer, and based on the specific characteristics of the SDUs, the convergence sublayer retrieves an applicable "PHS rule" for the SDUs. The applicable PHS rule provides a set of parameters needed to perform PHS. These parameters include:

- *PHS index* (PHSI): This is an index that references a specific PHS rule in use. The PHSI is an 8-bit number.
- *PHS field* (PHSF): This specifies those parts of the SDU header to be suppressed.
- *PHS mask* (PHSM): This is a mask showing which bytes in the PHSF *not* to suppress (to provide additional flexibility for PHS operation).

If the convergence sublayer performs PHS, it proceeds with suppressing all the bytes specified by the PHSF except those masked by the PHSM [2]. Then the convergence sublayer forwards the PHSI, which is the index of the specific PHS rule used, along with the processed packet to the common part sublayer.

At the receiver, the convergence sublayer receives the PDU from the common part sublayer. Using the CID and the PHSI, the convergence sublayer retrieves an identical PHS rule, which in turn provides the needed parameters (i.e., PHSF, PHSM, etc.) to reconstruct the payload header.

Needless to say, the PHS rule sets at both the base station and the mobile must be identical in order for PHS to operate properly. The rule sets are synchronized by using a series of dynamic service change (DSC) messages, including the DSC request (DSC-REQ) message, the DSC response (DSC-RSP) message, and the DSC acknowledge (DSC-ACK) message. These DSC messages are MAC management messages.

7.3 Common Part Sublayer

On the data plane, the common part sublayer carries out functions such as ARQ and fragmentation/packing—those functions involved in the assembly and processing of the MAC PDU. The reason why this sublayer is called "common part" is because its functions are the same (i.e., *common*) regardless of which higher-layer service forwarded the SDU. (Recall that it is the convergence sublayer above that performs interface functions *specific* to different higher-layer protocols.)

7.3.1 ARQ

The ARQ function performs the processing involved in ARQ. For ARQ-enabled connections, this function converts MAC SDUs into ARQ blocks and attaches sequence numbers to them [5]. The general ARQ process is discussed in Chapter 3.

7.3.2 MAC SDU and MAC PDU

A major function performed by the common part sublayer is MAC PDU assembly. Through packing and/or fragmentation, the common part sublayer assembles MAC SDUs into MAC PDUs that have a proper format for handling by the physical layer [8].

On the *downlink*, a MAC PDU assembled by the common part sublayer has three distinct sections, shown in Figure 7.3. The first section, generic MAC header, has 48 bits and contains important control information such as:

- The CID for the MAC PDU;
- The length of the entire MAC PDU;
- The types of subheaders and payload that are in the subsequent payload section;
- A flag bit showing if the payload is encrypted;
- The index of the key and initialization vector used to encrypt the payload if the payload is encrypted [2].

Although encryption-related information is appended here in the common part sublayer, the actual encryption and decryption are performed by the security sublayer.

The second section, payload, is variable in length and is optional. The payload section can carry three types of payloads. They are:

- User data from the convergence sublayer.
- MAC management messages for communicating control and management information (e.g., DL-MAP message and UL-MAP message).
- Subheaders for additional control and management information (e.g., the FAST-FEEDBACK allocation subheader).

The third section, cyclic redundancy check (CRC), is 32 bits in length and is used to check for errors in both the generic MAC header and the payload. If the payload is encrypted, then the CRC is calculated using the encrypted payload. The CRC is also discussed in Chapter 3.

On the *uplink*, there are two kinds of MAC PDUs shown in Figure 7.4. The first kind of uplink MAC PDU is like the downlink MAC PDU described above,

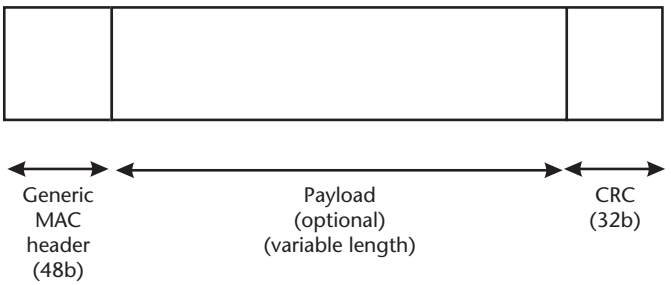


Figure 7.3 Downlink MAC PDU.

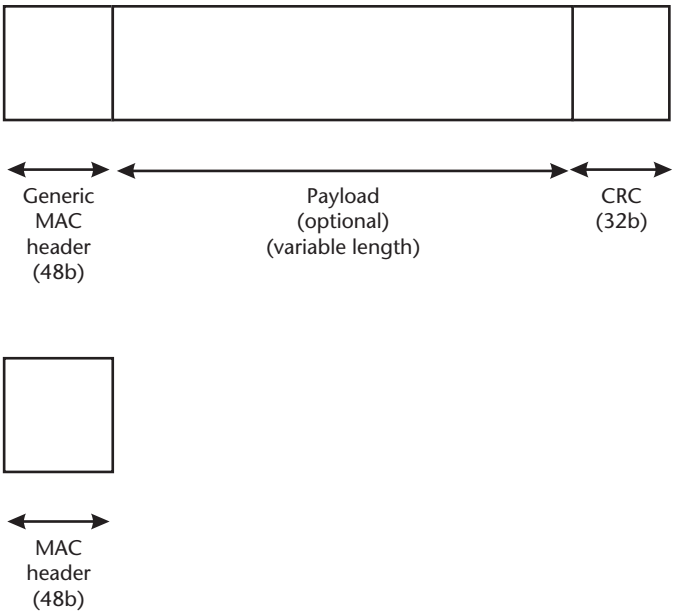


Figure 7.4 Uplink MAC PDUs.

consisting of three sections: generic MAC header, payload, and CRC. Similarly, the payload section of this kind of uplink MAC PDU can carry user data from the convergence sublayer, MAC management messages, and/or subheaders (e.g., the grant management subheader).

The second kind of uplink MAC PDU contains only a MAC header with no payload and no CRC (see Figure 7.4). The MAC header with no payload and no CRC affords an even faster way of communicating requests (e.g., bandwidth request), reports (e.g., Tx power report), and feedback (e.g., MIMO channel feedback) back to the base station. In fact, using a MAC header is faster than using a generic MAC header plus subheader because in using a MAC header, the mobile only has to communicate 48 bits (length of MAC header) back to the base station. This is in contrast to using a generic MAC header plus subheader, which costs 48 bits (length of generic MAC header) plus the number of bits of the subheader (which can be as long as another 48 bits).

7.3.3 Fragmentation/Packing

In carrying user data from the convergence sublayer, the payload section of a single MAC PDU can squeeze in multiple MAC SDUs if the SDUs are small in size. This process is called *packing* in that multiple MAC SDUs are packed inside a single MAC PDU. If packing is used, then the common part sublayer puts a specific subheader called the packing subheader in front of every MAC SDU packed. In this case, a packing subheader would contain additional information related to the MAC SDU immediately after it, such as the length of the packed MAC SDU (that comes immediately after the said packing subheader). The advantage of packing is that it decreases the number of overhead bits to be sent. This is so because, in

packing, the common part sublayer does not have to encapsulate each MAC SDU with a lengthy generic MAC header (48-bits long) and a full-blown CRC (32-bits long). Instead, each MAC SDU is preceded only by a shorter packing subheader (38 bits). Figure 7.5 shows an example of packing three MAC SDUs into a single MAC PDU. Note that the three original MAC SDUs can be different in size.

There are also situations where a MAC SDU is large in size and has to be broken into smaller fragments (to facilitate transport). This is called *fragmentation* in that a single MAC SDU is broken into multiple fragments,[1] which are then carried by one or more MAC PDUs. Here, the common part sublayer is capable of packing multiple fragments into a single MAC PDU. If a MAC PDU needs to pack multiple fragments, then the common part sublayer puts a specific subheader called the fragmentation subheader in front of every fragment packed. A fragmentation subheader would contain additional information related to the fragment immediately after it, such as the sequence number of the fragment (that comes immediately after the fragmentation subheader). Using the sequence numbers of the fragments, the peer common part sublayer at the receiver can later reassemble the original MAC SDU and deliver it to the sublayer above. Figure 7.6 shows an example of packing three MAC SDU fragments into a single MAC PDU. The three fragments can also have different sizes.

In fact, the common part sublayer is even capable of packing *both* MAC SDUs and MAC SDU fragments in the same MAC PDU. Figure 7.7 depicts an example where two MAC SDU fragments and two whole MAC SDUs are packed into the same MAC PDU. In this case, the common part sublayer uses packing subheaders to carry information (e.g., sequence numbers and lengths) related to SDU fragments/SDUs that come after the corresponding packing subheaders.[2] Being able to pack both MAC SDU fragments and MAC SDUs in a single MAC PDU permits a more flexible allocation of physical layer resources [2]. However, it is important to

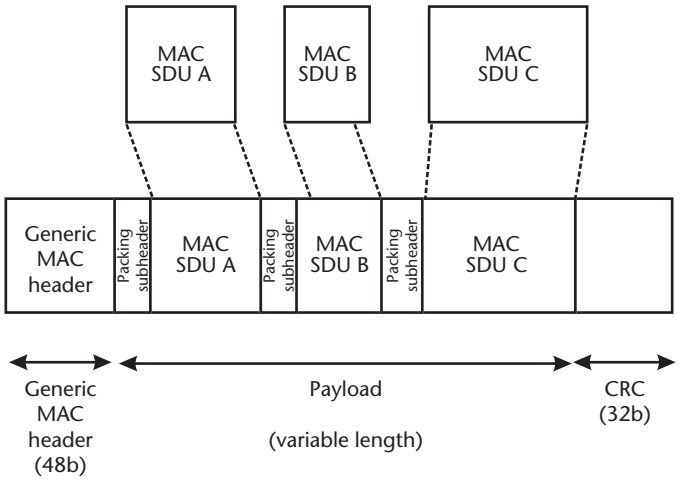


Figure 7.5 Packing multiple MAC SDUs.

1. Note that a MAC management message may be fragmented also.
2. These packing rules apply only to non-ARQ connections. For ARQ-enabled connections, similar but different rules apply.

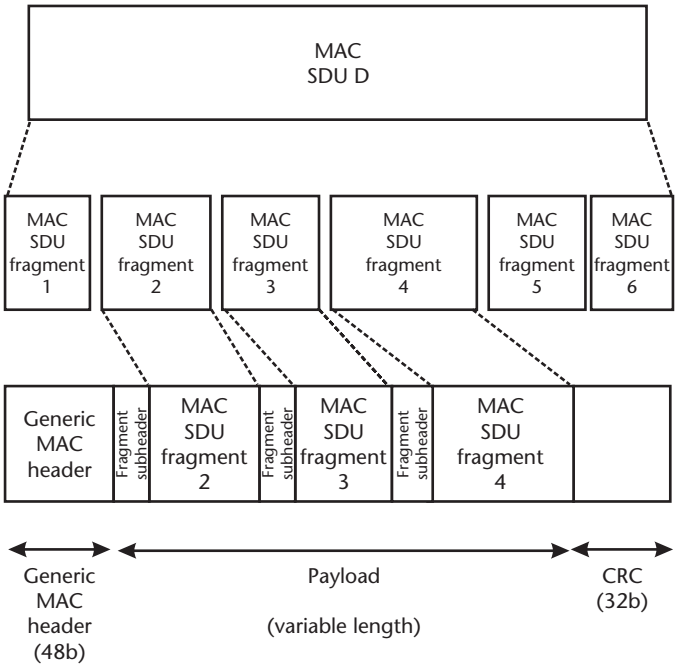


Figure 7.6 Packing multiple MAC SDU fragments.

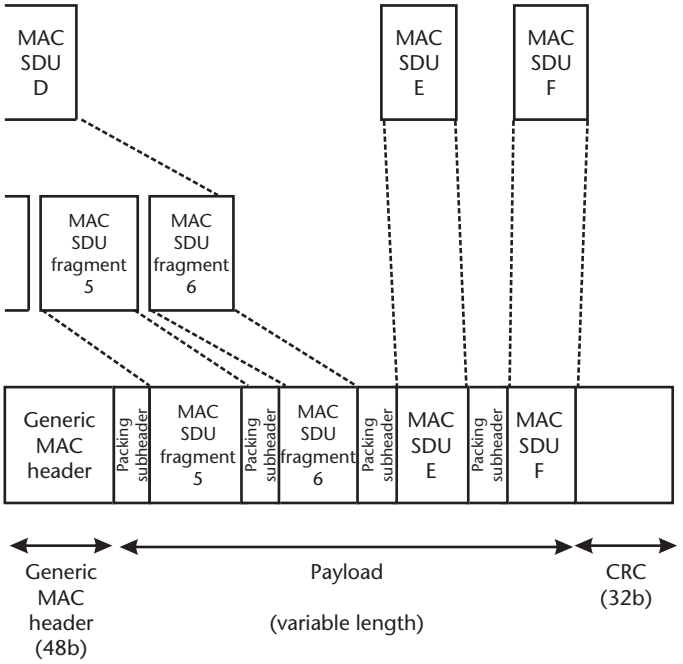


Figure 7.7 Packing both MAC SDU fragments and MAC SDUs.

recognize that each MAC PDU only transports user data for one connection (i.e., one CID) only. This is because only the generic MAC header specifies the CID in one of its fields; neither the packing subheader nor the fragmentation subheader specifies the CID.

Simultaneous fragmentation and packing enable an efficient use of the air interface [9]. Note that because the payload section of a MAC PDU can have a variable length, the length of any MAC PDU carrying a payload is also variable.

7.4 Security Sublayer

The security sublayer performs security functions such as authentication and encryption, as well as the enabling function of distributing key materials [2]. On the data plane, the encryption function encrypts the MAC PDU. If a MAC PDU is to be encrypted, then encryption is only applied to the payload section of the MAC PDU. The generic MAC header is not encrypted.

Encryption is not applied to the generic MAC header because the header itself contains important control information not only about the payload (e.g., CID), but also about the encryption process itself (i.e., encryption control and encryption key sequence). Specifically in the generic MAC header, encryption control (EC) is a one-bit flag showing whether or not the payload is encrypted, and the encryption key sequence (EKS) is a two-bit field showing the sequence number of the key (i.e., traffic encryption key or TEK) and initialization vector used for encrypting the payload. Security functions are examined in more detail in Chapter 12.

References

[1] ISO/IEC 7498-1, "Information Technology—Open Systems Interconnection—Basic Reference Model: The Basic Mode," Geneva: ISO, 1994.

[2] IEEE Standard 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York: IEEE, October 1, 2004.

[3] Nair, G., et al., "IEEE 802.16 Medium Access Control and Service Provisioning," *Intel Technology Journal*, Vol. 8, No. 3, 2004, pp. 213–228.

[4] Etemad, K., and L. Wang, "Multicast and Broadcast Multimedia Services in Mobile WiMAX Networks," *IEEE Communications*, Vol. 47, No. 10, 2009, pp. 84–91.

[5] Etemad, K., "Overview of Mobile WiMAX Technology and Evolution," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 31–40.

[6] Ahmadi, S., "An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Communications*, Vol. 47, No. 6, 2009, pp. 84–98.

[7] Riegel, M., "Ethernet Services over Mobile WiMAX," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 86–93.

[8] Nakamura, M., T. Chujo, and T. Saito, "Standardization Activities for Mobile WiMAX," *Fujitsu Scientific and Technical Journal*, Vol. 44, No. 3, 2008, pp. 285–291.

[9] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

CHAPTER 8

Medium Access Control: Lower Control Plane

8.1 Introduction

This chapter investigates those selected functions in the lower part of MAC on the control plane. Figure 8.1 shows that the control plane has functions involved in the control, signaling, and management of radio resources. The lower part of the control plane has the more traditional functions of MAC such as physical layer control (e.g., ranging and power control), scheduling, control signaling, QoS, and others [1, 2]. In this chapter, we discuss scheduling, control signaling, and physical layer control functions including ranging and power control. QoS is examined in more detail in Chapter 10.

8.2 Scheduler

After a MAC PDU is assembled, the scheduler assigns the MAC PDU to be physically transmitted using a burst provided by the physical layer (see Chapter 5). Then the MAC PDU is passed to the physical layer for actual transmission. Note that a single burst can physically transport several MAC PDUs.[1]

The scheduler is responsible for planning and assigning a MAC PDU for actual transmission; it also provides QoS differentiation. As such, the scheduler must efficiently allocate available bandwidth resources provided by the physical layer to maximize system performance. For example, in allocating available bandwidth resources on the downlink and the uplink, the scheduler in the base station considers the following factors:

- The size of the MAC PDU (or the size of the bandwidth request);
- The MAC PDU's QoS parameters (retrieved based on the MAC PDU's SFID);
- Competing requests from other MAC PDUs with their own QoS parameters;

1. The process of fitting more than one MAC PDU into one burst is called concatenation.

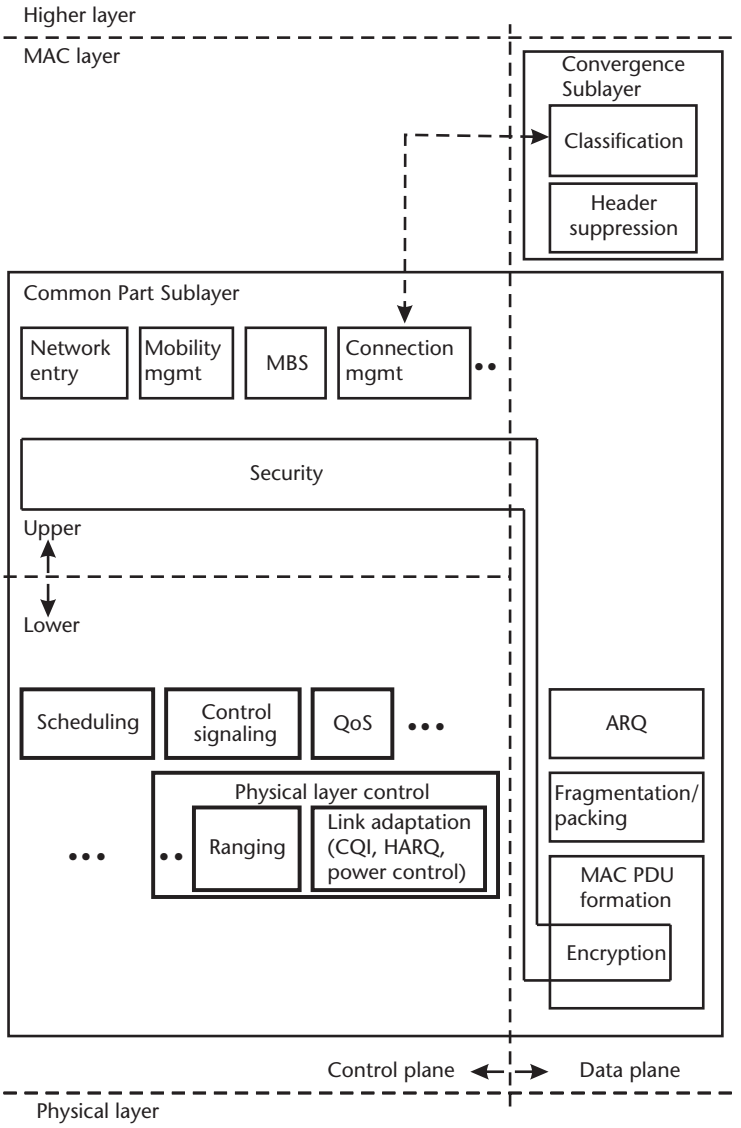


Figure 8.1 MAC protocol structure. Note that the boldfaced parts constitute the lower part of the control plane. (*After:* [1, 2].)

- Channel conditions (through the CQICH or reciprocity for the downlink and through direct measurement for the uplink).

Taking into account these factors, the scheduler produces an optimal assignment of MAC PDUs to physical-layer bursts. In addition, the scheduler also decides on the modulation and coding scheme to be used by each mobile (i.e., AMC) and performs ARQ to maximize throughput while meeting the target frame error rate. The decision on which modulation and coding scheme to use is based on the channel quality measurements (e.g., feedback sent by the mobile). Note that this operation of the scheduler constitutes crosslayer optimization in that the MAC layer (scheduler) makes decisions based on information from the physical layer, and the physical layer responds to decisions made by the MAC layer [3]. Moreover,

because the scheduler generally communicates bandwidth allocations in DL-MAP and UL-MAP (which are at the beginning of every frame), it can dynamically allocate bandwidth quickly—once every frame—in response to changing bandwidth demands and changing channel conditions [4]. QoS in the context of IEEE 802.16 is discussed in Chapter 10.

There are several different types of scheduling algorithms [4, 5]:

- *Round robin:* This is a simple scheduling algorithm that has mobiles take turns to be served in order. The round robin scheduler is fair in that it does not discriminate among mobiles, but it does not exploit the opportunity to transmit in a good RF condition [6].
- *Weighted round robin:* It is the round robin scheduler that incorporates static weights.
- *Maximum SINR:* It gives priority of bandwidth allocation to those mobiles that have the highest SINRs. This scheduler tends to maximize aggregate throughput for a base station but punish those mobiles that have poor SINRs, thus it has the worst fairness control among mobiles [6].
- *Temporary removal scheduler*: It temporarily skips bandwidth allocations for those mobiles that have SINRs that are less than some threshold [3].
- *Proportional fairness*: It gives priority to those mobiles that have the highest ratios of current throughput to average throughput. This algorithm offers a tradeoff between aggregate throughput and fairness. A mobile with a high SINR has a high priority, and the algorithm increases aggregate throughput because it gives priority to mobiles that have high current throughputs. At the same time, if a mobile's SINR has been low, then that mobile's average throughput is low, but low average throughput increases the mobile's ratio of current throughput to average throughput, thus a higher priority is given to that mobile.

The actual scheduler protocol is not specified in the IEEE 802.16 suite of standards, so every vendor is free to design its own proprietary scheduling and reservation algorithms. This arrangement is similar to other broadband wireless standards such as EV-DO [7]. A survey of scheduler algorithms for general broadband wireless networks can be found in [8], and a similar survey of schedulers for OFDM-based broadband wireless networks can be found in [9].

8.3 Bandwidth Request

On the uplink, a mobile transmits its request for uplink bandwidth to the base station. On the downlink, the base station is already aware of all the MAC PDUs, which are waiting to go to their respective destinations (mobiles), thus the base station can directly make scheduling decisions for MAC PDUs on the downlink. However, on the uplink, the mobiles have to let the base station know of their bandwidth requests [3]. Once a mobile enters the network and has a connection established, it can start asking the base station for bandwidth allocation on that connection.

(Network entry is discussed in Chapter 9.) A mobile needs a bandwidth allocation to actually transmit user data on the uplink.

The mobile always makes its bandwidth request in terms of the number of bytes required, not the number of subchannels and the number of OFDM symbols. This is because the burst profile being used may change on the fly, and when the burst profile changes, the number of bits carried by each data symbol changes as well. Therefore, the mobile requests for some number of bytes that need to be sent, and the scheduler determines the actual uplink bandwidth needed while taking into account the burst profile used.

There are several techniques that the mobile can use to request uplink bandwidth. The mandatory techniques include:

- Request in existing uplink allocation;
- Unicast polling;
- Multicast and broadcast polling;
- Contention-based request for OFDMA.[2]

8.3.1 Request in Existing Uplink Allocation

If the mobile has an existing uplink allocation for transmitting user traffic, then it can use that allocation to send a MAC header (i.e., the bandwidth request header) to request bandwidth. Optionally, the mobile can piggyback a bandwidth request on a subheader (i.e., the grant management subheader) in an existing uplink MAC PDU to convey its bandwidth request; this is known as the piggyback request (PBR). Since the bandwidth request header is 48-bits long whereas the grant management subheader is 32-bits long, using the piggyback request is more efficient (if there is an existing MAC PDU to transmit on the uplink).

For a mobile that has an existing uplink unsolicited grant service (UGS) connection, a special request mechanism exists for that mobile to request extra bandwidth on that UGS connection. To request extra bandwidth on an UGS connection, the mobile can use the slip indicator (SI) bit in the grant management subheader to do so. By setting the SI bit in the grant management subheader it transmits on the uplink, the mobile indicates that its transmit queue buffer (for the UGS connection) is overflowing and it needs more bandwidth to relieve the queue. Upon receiving the grant management subheader and reading the set SI bit, the base station may grant a small additional bandwidth (up to 1%) to the mobile [10, 11]. See Chapter 10 for more details on UGS.

8.3.2 Unicast Polling

If the mobile does not have an existing uplink allocation for transmitting user traffic, then the base station can assign that mobile a bandwidth in which the mobile can convey its bandwidth request. This technique is called *unicast polling*, which is essentially a process by which the base station "polls" *each mobile* (hence the

2. The contention-based request for the OFDM implementation is also specified by the standard.

term unicast) to ask if it has any user traffic to send. In doing so, the base station normally assigns a data grant information element (IE)[3] to that mobile (or more specifically, to that mobile's basic CID) in the UL-MAP message. The data grant assigns a bandwidth in which the mobile may respond with a bandwidth request header. If the mobile needs to request uplink allocation for transmitting user traffic, then it sends a bandwidth request header in that assigned bandwidth. If the mobile has no request, then it pads the assigned bandwidth.

In addition, a mobile can proactively make the base station poll the mobile. It does so by using the poll-me (PM) bit in the grant management subheader. By setting the PM bit in the grant management subheader it transmits on the uplink; the mobile indicates that it needs to be polled because it needs to request bandwidth.[4] Upon receiving the grant management subheader and reading the set PM bit, the base station proceeds with unicast polling (described above) to poll the mobile [10, 11].

8.3.3 Multicast and Broadcast Polling

Polling all mobiles individually obviously costs a lot of capacity. If the base station does not have enough capacity, then it can use the *broadcast polling* technique. This technique is similar to unicast polling in that the base station assigns a bandwidth, which mobiles use to convey requests. However, the major difference is that *all mobiles* served by that base station can transmit their requests in the same assigned bandwidth; the base station assigns that bandwidth to a broadcast CID in the UL-MAP message. Needless to say, because all mobiles may use that bandwidth to convey their requests, the requests can collide and thus are contention-based. If a collision occurs, a mobile uses a contention resolution protocol (specified in the IEEE 802.16 suite of standards) to back off and retransmit its request.

In addition to broadcast polling, the technique of *multicast polling* is also available. Similar to broadcast polling, multicast polling lets the base station assign a bandwidth in which requests are conveyed, except that only a *group of mobiles* (i.e., multicast group) can transmit their requests in this assigned bandwidth; in this case, the base station assigns that bandwidth to a multicast CID in the UL-MAP message. Multicast polling normally costs more capacity than broadcast polling.

8.3.4 Contention-Based Request for OFDMA

For OFDMA, the standard mandates an additional technique for requesting bandwidth. This technique is based on transmitting a PN code for bandwidth request, which is similar to transmitting the PN code for ranging (also used by OFDMA). Basically, there is a ranging channel specified in the uplink subframe, and a mobile may initiate its request process by modulating a PN code onto this ranging channel and transmitting the code. If the base station receives this PN code, the base station

3. An information element (IE) is an organized set of fields (parameters) sent inside a PDU. The organized set of fields or parameters is intended for one mobile, a group of mobiles, or all mobiles to adopt. Note that more than one IE may be sent simultaneously.

4. The mobile can only do this if it already has an existing uplink UGS connection (so it can transmit the grant management subheader) and it needs to request bandwidth for non-UGS connections.

still has no idea which mobile transmitted the PN code, but that is OK. The base station would broadcast an UL-MAP message, which contains the PN code (and parameters) used by that mobile, as well as specifies a bandwidth in which that mobile may formally send its bandwidth request. Upon receiving the UL-MAP message, the mobile would know that the base station heard its PN code, and the mobile can use the associated bandwidth assigned to transmit its bandwidth request. In doing so, the mobile transmits a bandwidth request header in the assigned bandwidth. Note that the reason for using the PN code is that if two PN codes transmitted by two mobiles collide, then the base station may still detect each PN code by using the correlation process.

If a mobile's bandwidth request results in a bandwidth allocation, the base station communicates that allocation in a subsequent UL-MAP message or messages sent to the mobile.

8.4 Control Signaling

The control signaling function generates resource allocation messages including DL-MAP, UL-MAP, and other control signaling messages [2]. Out of these, the MAC messages DL-MAP and UL-MAP are essential. In the downlink subframe, the base station transmits DL-MAP followed by UL-MAP. DL-MAP contains information on the subchannels and OFDM symbols that are assigned to each mobile in the downlink subframe, and UL-MAP has information on the subchannels and OFDM symbols assigned to each mobile in the uplink subframe. UL-MAP also specifies the uplink subframe's ranging channel, which is used for initial ranging, periodic ranging, and contention-based bandwidth requests [6].

On the downlink, the base station is the only one that transmits the downlink subframe. The base station uses the DL-MAP message to let the mobiles know when they should listen for and receive their respective bursts. On the uplink, the base station also determines the slots in which each mobile may transmit in the uplink subframe. Here, the base station uses the UL-MAP message to let the mobiles know when they should transmit. In particular, UL-MAP contains an IE that specifies the transmission opportunities (i.e., slots in which the mobile may transmit), and each mobile transmits in the predefined slots specified by the IE [6].

The MAP message has a fixed part and a variable part, and the length of the variable part depends on the number of mobiles being scheduled in the frame [4]. With a 5-ms frame, five scheduled mobiles per frame are optimal for delay-tolerant traffic such as File Transfer Protocol (FTP) and Hypertext Transfer Protocol (HTTP). For voice-over-IP (VoIP), the number of scheduled mobiles per frame may be 15 to 20 users [12].

Because DL-MAP and UL-MAP are typically meant for all mobiles in a cell/sector, they are sent using low-order, more robust, modulation and coding. However, a low-order modulation carries fewer bits on a data symbol. So a MAP message would have to be longer in length to carry a given amount of information. To reduce MAP overhead, the base station may also transmit SUB-DL-UL-MAP ("Sub-MAP" message) in addition to DL-MAP and UL-MAP. The system can encode SUB-DL-UL-MAP using a different modulation and coding scheme. For those users that are closer to the base station and have higher CINRs, the base station may

transmit multicast SUB-DL-UL-MAP to them using higher-order, more efficient modulation and coding and hence reduce MAP overhead [12].

Moreover, in IEEE 802.16m, an advanced MAP allocation IE can be addressed to one user or to multiple users. The IE also contains information on resource allocation, but now each unicast IE is encoded separately. This scheme further improves efficiency [2].

8.5 Ranging

8.5.1 Initial Ranging

The first ranging performed as part of network entry is called initial ranging, which typically requires more than one request-response cycle to complete. In the OFDMA implementation, the initial ranging process is as follows: The mobile transmits a PN code in the ranging channel, which is inside the uplink subframe (see Figure 5.10). After sending the PN code, the mobile receives a RNG-RSP message. If that RNG-RSP message contains the parameters of the mobile's own PN code, then the ranging process is not complete. In this case, the mobile needs to adjust the timing offset per what is specified in the RNG-RSP message. Then it transmits another PN code (with the adjusted timing offset) and waits for a response.

Note that in the OFDMA implementation, the mobile ranges by first sending the PN code (not a RNG-REQ message as in the OFDM implementation). By just reading the PN code, the base station has no idea which mobile sent it. Thus, the only way the base station can respond is by broadcasting a RNG-RSP message that contains the received PN code. This way, the mobile that sent the PN code can identify the RNG-RSP message (by its own PN code) intended for it and implements the timing offset specified by that RNG-RSP message.

A successful initial ranging occurs when the following take place: the mobile receives an UL-MAP message containing the parameters of the PN code it previously used to range, and the mobile sends a RNG-REQ message using the allocation specified in that UL-MAP message. Afterwards, if the mobile receives an RNG-RSP message containing the mobile's MAC address and a "success" ranging status, then initial ranging is successful; the RNG-RSP message would contain the basic and primary management CIDs and RF-related parameters for the mobile to use.

PN codes are used because mobiles' transmissions can collide with one another on the ranging channel. If two PN codes transmitted by two mobiles collide, then the base station can still detect each PN code by using the correlation process. The process is similar to that used by cellular CDMA, which detects each user even though all users are transmitting at the same time in the same frequency band [13].

In the OFDM implementation, initial ranging follows a somewhat different process, which is described here to contrast the OFDMA implementation: The mobile sends a RNG-REQ message on a contention basis during a period called the initial ranging interval. The initial ranging interval is at the beginning of the uplink subframe. After sending the RNG-REQ message, the mobile receives a RNG-RSP message. If that RNG-RSP message contains only the frame number of the previous RNG-REQ message transmitted, then the ranging process is not complete. In this case, the mobile needs to adjust the timing offset per what is specified in the

RNG-RSP message. Then the mobile transmits another RNG-REQ message (with the adjusted timing offset) and waits for a response. In OFDM, a successful initial ranging occurs when the mobile receives a RNG-RSP message containing the mobile's MAC address and a "success" ranging status. In this case, the RNG-RSP message would contain the basic and primary management CIDs assigned to the mobile, as well as other more specific RF-related parameters for the mobile to use.

After the successful completion of initial ranging, the base station assigns the basic and primary management CIDs to the mobile. The basic CID is for sending those MAC management messages that cannot tolerate any delay, such as the subscriber station basic capability request (SBC-REQ) message and the handover messages. On the other hand, the primary management CID is for sending those MAC management messages that can tolerate some delay, such as the registration request (REG-REQ) message and the dynamic service addition request (DSA-REQ) message.

8.5.2 Periodic Ranging

In addition to carrying out *initial ranging* during network entry, the system also performs *periodic ranging* on a regular basis. As mentioned above, ranging is the process by which the mobile obtains adjustments to timing offset (and other uplink parameters) so that its uplink transmissions are aligned with the specified frame [10]. Ranging is needed especially on the uplink because it is on the uplink where mobiles' transmissions experience different delays, and these delays must be corrected. In ranging, the base station receives a mobile's transmission, calculates the necessary timing offset required, and sends the offset adjustment information back to the mobile. Periodic ranging is needed because the channel changes as a mobile moves around in its environment. Hence updated adjustments are necessary.

In the OFDMA implementation, the mobile keeps a ranging timer (i.e., *T4*). When the timer runs out, the mobile chooses a ranging slot (using a randomized algorithm to minimize the probability of collision) and uses that ranging slot to send a PN code to the base station. After the mobile transmits the PN code, the base station sends back a RNG-RSP message containing the PN code of the mobile. By reading its own PN code in the RNG-RSP message, the mobile knows that the RNG-RSP message is intended for it. If the RNG-RSP message shows a "continue" ranging status, then the mobile proceeds to implement the adjustments per what are specified in the RNG-RSP message and sends another PN code. The process ends when the mobile receives an RNG-RSP message with a "success" ranging status.

As one can see, the above process (in OFDMA) is initiated by the mobile, but the base station may also prompt the mobile to start the periodic ranging process. The base station does so by sending an unsolicited RNG-RSP message, with a "continue" status, to the mobile.

In the OFDM implementation, periodic ranging also follows a somewhat different process in contrast to the OFDMA implementation. Here the base station regularly checks the uplink signal transmitted by the mobile for its quality (e.g., timing alignment). If the quality is within acceptable range and there is no RNG-REQ message in the uplink signal, then the base station does nothing. If the quality is within an acceptable range but there is a RNG-REQ message in the uplink

signal, then the base station sends a RNG-RSP message with a "success" ranging status (indicating that there is no need to range). On the other hand, if the quality is outside the acceptable range, then the base station sends a RNG-RSP message with a "continue" ranging status; in this case, the mobile implements the adjustments specified in the RNG-RSP message. If the number of ranging attempts becomes excessive but the uplink signal quality is still outside the acceptable range, then the base station sends a RNG-RSP message with an "abort" ranging status and drops the mobile.

8.5.3 Handover Ranging

Ranging is also performed during handover between two base stations. Here, the mobile performs essentially the same ranging process with the *target* base station. Handover is discussed in more detail in Chapter 9.

8.6 Power Control

Power control seeks to adjust the transmit power[5] to a level that is just enough to sustain the radio link. From a user perspective, (uplink) power control helps conserve the device's battery power, which is a nontrivial issue with small form-factor, high-speed, converged devices.

From a system perspective, there are two main reasons for power control in an OFDMA-based system: minimizing interference to nearby cochannel cells and minimizing the effect of intercarrier interference. On the uplink, both of these reasons apply. First, by commanding a mobile to power to a level that is just enough to maintain the uplink, the base station is preventing that mobile from spewing excess power into the receiver of a neighboring base station. This is especially important if a system does not adopt frequency reuse and use the same frequency carrier for all base stations. Even if a system does adopt frequency reuse, uplink power control would still help decreasing the amount of interference directed at the receiver of a cochannel base station nearby.

Second, more importantly, uplink power control minimizes the effect of intercarrier interference if a mobile's transmission is misaligned in frequency (due to an oscillator drift or Doppler shift). This is especially applicable when different users are assigned different subcarriers at the same time in an OFDM symbol (as in the case of OFDMA). Recall from Chapter 1 that if subcarriers are misaligned in frequency, then subcarriers are no longer orthogonal to each other. In OFDMA, different subcarriers are assigned to different users. If one user's received power is unnecessarily higher than another's, then the higher received power will exacerbate the problem of intercarrier interference if there is misalignment in frequency.

Figures 8.2 and 8.3 illustrate the situation. In Figure 8.2, both user 1 and user 2 have identical transmit power P_T, and there is no power control on the uplink. Because user 1 is closer to the base station, user 1's received power $P_{R,1}$ is higher than user 2's received power $P_{R,2}$. As a result, Figure 8.3 shows that user 1 would contribute more correlated interference to user 2 if there is misalignment in frequency.

5. In the context of OFDMA, transmit power refers to power per subcarrier.

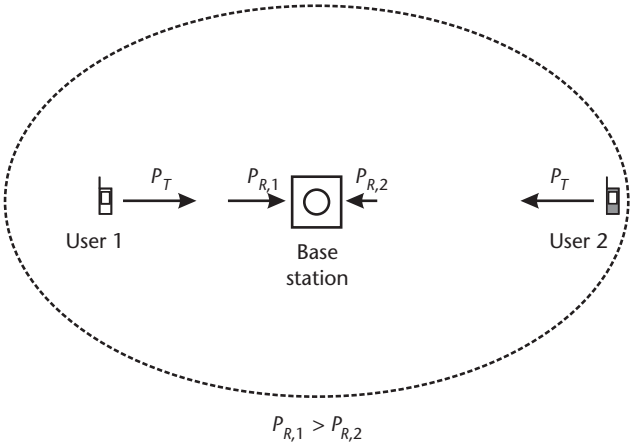


Figure 8.2 There is no uplink power control, and both user 1 and user 2 have identical transmit power P_T.

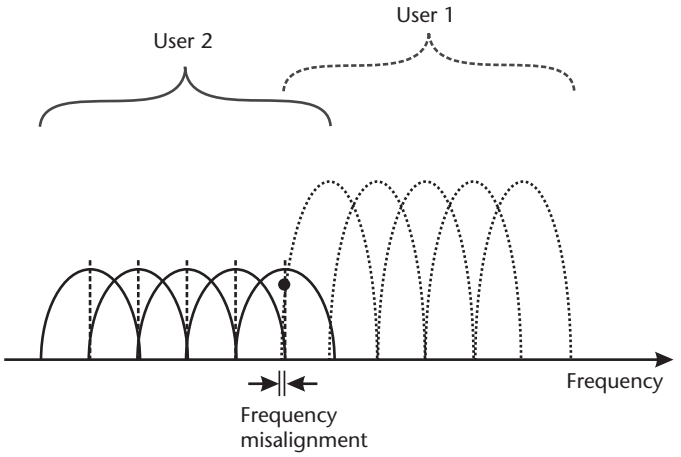


Figure 8.3 User 1 introduces higher intercarrier interference to user 2. In this hypothetical OFDMA example, user 1 and user 2 are each assigned five contiguous subcarriers.

If there is power control as shown in Figure 8.4, then the base station will power control each user's transmit power levels $P_{T,1}$ and $P_{T,2}$ such that the received power levels will be identical (P_R), assuming users 1 and 2 have the same modulation and coding scheme. As a result, user 1 would contribute less correlated interference to user 2 if there is frequency misalignment (see Figure 8.5).

On the downlink, the reason of minimizing intercarrier interference does not apply because all subcarriers (assigned to different users) originate from the same base station and are presumably already aligned in frequency. However, the reason for minimizing interference to nearby cochannel cells still applies. In other words, a base station should not arbitrarily increase its transmit power to a faraway mobile because such an increase may interfere with another mobile in a nearby cochannel cell.

Because the uplink has the additional requirement of minimizing intercarrier interference in case of frequency misalignment, uplink power control is especially important. As such, the standard specifies power control mechanisms for the uplink. Note that even though power control is important on the uplink, this requirement

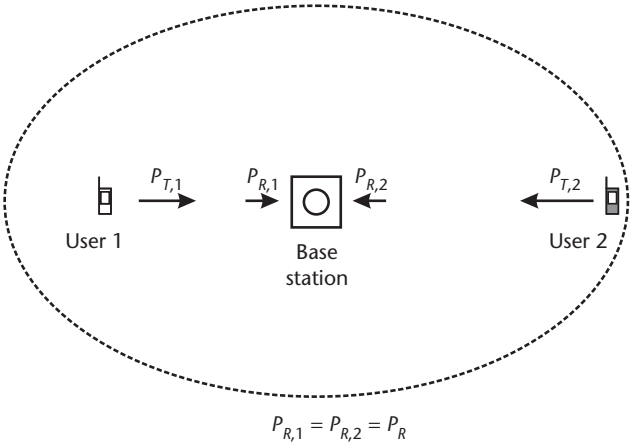


Figure 8.4 There is uplink power control, so both user 1 and user 2 have identical receive power P_R (if they both use the same modulation and coding scheme).

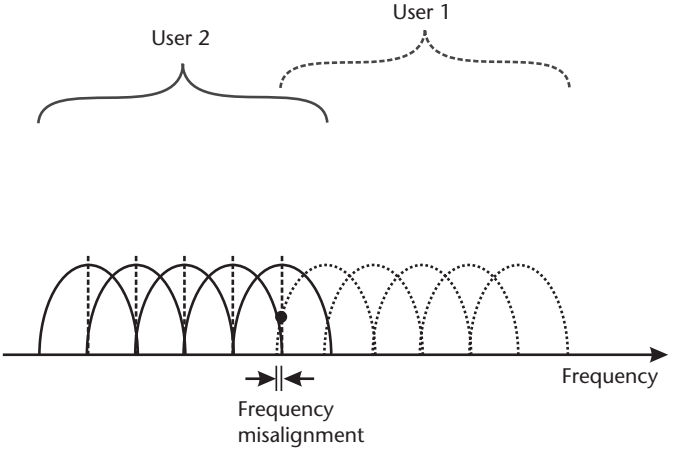


Figure 8.5 User 1 introduces lower intercarrier interference to user 2.

is not nearly as stringent as in cellular CDMA. In CDMA, a dominant effect of poor power control is the degradation of *capacity* (or the number of simultaneous users). In fact, CDMA2000 systems can power control both uplink and downlink at a rate of 800 times per second [7].

The standard does not specify power control mechanisms for the downlink, and the infrastructure vendors are free to incorporate downlink power control in their base station products. It is expected that at least some vendors will incorporate some downlink power control features in their base stations to minimize downlink interference to cochannel cells nearby.

The following sections discuss the uplink power control functions specified by IEEE 802.16.

8.6.1 Uplink Power Control: Closed-Loop

Through closed-loop uplink power control, the mobile changes its transmit power based on explicit commands from the base station. At the same time, the mobile has some intelligence built into it so that if some conditions change, the mobile can autonomously change its transmit power without being explicitly told to do so by the base station. Thus, the functionality of closed-loop uplink power control is effectively divided between the mobile and the base station. The next two sections describe closed-loop uplink power control from the perspectives of both the mobile and the base station.

8.6.1.1 Mobile

The mobile can change its transmit power unilaterally in the following situations:

- The burst profile (modulation and coding) changes and the repetition rate for the new burst profile changes.
- After sending a PN code for periodic ranging or a PN code for bandwidth request, the mobile does not receive a response from the base station.
- The number of subchannels allocated to the mobile changes.

First, if the burst profile (modulation and coding) changes and the repetition rate for the new burst profile changes, then the transmit power also changes. For example, if the physical-layer modulation changes from a lower order to a higher order (e.g., from QPSK to 16-QAM), then the mobile would automatically increase its transmit power since a higher-order modulation requires more transmit power (to maintain a given bit error rate). The amount of increase is based on the ratio of the new required CINR to the old required CINR, and these CINR numbers are retrieved from a table of values of default required CINR per subcarrier. For example, if the old modulation (e.g., QPSK) has a default required CINR of 9 dB and the new modulation (e.g., 16-QAM) has a default required CINR of 15 dB, then the mobile can, by itself, increases its transmit power by 6 dB. The same is true for the repetition rate. For example, if the repetition rate triples (from 2 to 6), then the mobile would decrease its transmit power by 1/3 (or by 5 dB).

The mobile retrieves the values of default required CINR per subcarrier from a table lookup, but the base station can also explicitly tell the mobile what value of default required CINR per subcarrier to use (through an UCD message). The automatic change procedure described above only applies when the mobile transmits on the FAST-FEEDBACK channel, or when it requests bandwidth or ranges.

Second, if after sending a PN code for periodic ranging or a bandwidth request, the mobile does not receive a response from the base station, then the mobile may adjust the transmit power of its next transmission of the PN code (for periodic ranging or bandwidth request). Typically, the mobile increases the transmit power of its next transmission. But the transmit power cannot exceed the maximum defined by the parameter $P_{TX\_IR\_MAX}$.

Lastly, if the number of subchannels (which determines the number of subcarriers) allocated to the mobile changes, then the total aggregate transmit power changes as well. For example, if the number of subcarriers allocated to the mobile

doubles, then the amount of total aggregate transmit power doubles also. This is because the transmit power per subcarrier is kept the same to maintain subcarrier integrity. Here the mobile increases the total aggregate transmit power by itself, without intervention from the base station [11].

8.6.1.2 Base Station

The base station has the ability to directly control the uplink transmit power of an individual mobile. The standard does not specify the power control algorithm, so vendors are free to develop and implement their own power control schemes. One model of uplink power control is shown in Figure 8.6. Here the base station receives an uplink transmission from the mobile. First, the base station measures the current SINR (after the receiver) and the current frame error rate (FER) (after the demodulator). Then the base station feeds both the current SINR and the current FER into a computational engine. The engine uses both the current SINR and the current FER measurements to quickly calculate a new SINR reference level, which is the SINR value required to maintain an acceptable FER. For example, if the current SINR is too low and the current FER is too high, then the engine adjusts the SINR reference level higher (because more signal power is clearly needed). If the SINR is too high and the FER is too low, then the engine adjusts the SINR reference level lower.

Afterwards, the new SINR reference level is compared to the current SINR measurement. If the measurement is less than the reference, then the current SINR is lower than what is necessary to maintain an acceptable FER; the base station

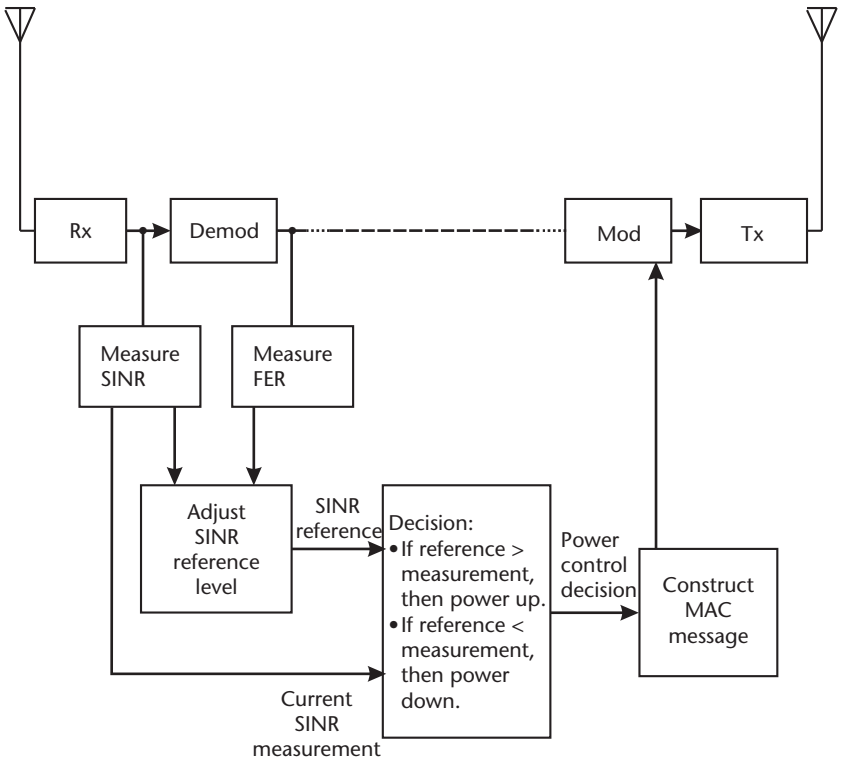


Figure 8.6 A model of uplink power control functions in the base station.

sends a power up command to the mobile. If the measurement is greater than the reference, then the current SINR is higher than what is necessary to maintain an acceptable FER; the base station sends a power down command to the mobile [7].

The way power control commands are sent is normally through periodic ranging. Power control and adjustments of mobile transmit power can also occur during initial ranging. Specifically, the RNG-RSP message can be used to communicate power control commands through its 8-bit *power level adjust* field. Recall that the RNG-RSP message is used throughout the ranging process, and the base station can embed its uplink power control command in the RNG-RSP message sent to the mobile.

In addition, the UL-MAP message can be used to command power up or power down through its power control IE. Specifically, the 8-bit *power control* field in the power control IE specifies power up or power down in multiples of 0.25 dB. The power control IE is directed at a single mobile (or more specifically, to that mobile's basic CID) in the UL-MAP message.

Transmitting power control commands costs management overhead. For example, in the UL-MAP message, the power control IE is 16-bit long in the OFDM implementation and 24-bit long in the OFDMA implementation. So the power control computational engine needs to trade off power control accuracy against management overhead. Namely, how high (or low) of an SINR is too high (or too low) to warrant adjusting the SINR reference level and thus possibly triggering a transmission of a power control command.

An optional feature made available to both OFDM and OFDMA in the standard is *fast power control*. If the base station needs to quickly command more than one mobile to power up or power down, the base station can send a broadcast fast power control (FPC) message. This message is sent to the broadcast CID, so all the mobiles served by the base station would read it. But within the FPC message, there is a list of Basic CIDs of only those mobiles to which the power control command (specified by the FPC message) applies. For example, if the base station sees that the SINRs of a group of mobiles suddenly drop (perhaps due to a localized blockage), the base station can command the uplink transmit powers of only those mobiles to go up at the same time. For those affected mobiles, this process would be much quicker than commanding power up through regular periodic ranging. Each FPC message can command power up or power down, via the *power adjust* field, in multiples of 0.25 dB.

The closed-loop power control mode is the default mode specified by the standard. In addition to closed-loop power control, the standard also specifies the open-loop power control mode, discussed next.

8.6.2 Uplink Power Control: Open-Loop

The open-loop power control mode is an optional mode. Using open-loop power control, the mobile can autonomously set its transmit power as follows (in OFDMA) [11]:

$$P_T = \frac{C}{I} + I + L - 10\log_{10} R_r + Offset\_SS_{perSS} + Offset\_BS_{perSS} \tag{8.1}$$

where

- P_T is the effective radiated power (ERP) per subcarrier in dBm.
- *C/I* is the default required CINR per subcarrier for the currently active burst profile (modulation and coding). Understandably, this number gets progressively higher for higher-order modulations (and lower for lower-rate FECs). This number is in decibels.
- *I* is the interference and noise power per subcarrier before the base station receive antenna. It is expressed in dBm.
- *L* is the sum of the current uplink propagation loss and transmit antenna gain. *L* is expressed as a positive number. Its unit is in decibels.
- R_r is the repetition rate for the currently active burst profile (modulation and coding).
- $Offset\_BS_{perSS}$ is an offset term that is controlled by the base station.
- $Offset\_SS_{perSS}$ is an offset term that is controlled by the mobile.

As one can see, (8.1) is basically the link equation (2.1) expressed in terms of the mobile's ERP. Obviously, P_T would need to increase if the required CINR increases, if the interference and noise power increases, or if the repetition rate decreases. The default required CINR per subcarrier is obtained by the mobile via table lookup but can be overridden by the base station; *I* is sent to the mobile in the DCD message; the mobile obtains *L* by subtracting its total received power from the base station EIRP (*BS_EIRP*) parameter. *BS_EIRP* is also sent to the mobile in the DCD message. As far as R_r is concerned, R_r is already known by the mobile in the demodulation process.

By comparing (8.1) to the link equation (2.1), an observant reader may be wondering where the term for the base station antenna gain is. The base station antenna gain is numerically incorporated in the offset term $Offset\_BS_{perSS}$ that is controlled by the base station.

Regarding the mobile's control of the term $Offset\_SS_{perSS}$, it has two options: the mobile can set $Offset\_SS_{perSS}$ to zero (this option is called *passive* uplink open-loop power control), or the mobile can actively change it according to some criterion (this option is called *active* uplink open-loop power control).

For the active uplink open-loop power control, the criterion specified by the standard involves the use of ARQ. Using ARQ, the base station would send an acknowledgment (ACK) message to the mobile if an uplink packet is received successfully. The base station would send a negative acknowledgment (NAK) message to the mobile if an uplink packet is not received correctly. So, the mobile may increment $Offset\_SS_{perSS}$ by a step defined by the parameter *UP_STEP* if it receives a NAK, and the mobile may decrement $Offset\_SS_{perSS}$ by a step defined by the parameter *DOWN_STEP* if it receives an ACK. Again, we see that as in many power control schemes, the bias is to the downside as far as the transmit power is concerned. In other words, if the link is in good order, gradually decrease the transmit power. In any case, $Offset\_SS_{perSS}$ cannot go above the parameter $Offset\_Bound_{upper}$ or go below the parameter $Offset\_Bound_{lower}$. These parameters used in active uplink open-loop power control can be found in the UCD message.

Note that open-loop power control, if implemented, only lets the mobile perform an estimate of the required transmit power. The fine adjustment of the mobile's transmit power comes from closed-loop power control. Of course, if the optional open-loop power control or even the active uplink open-loop control is used, the mobile can take over some power control responsibilities and thereby minimize the need for closed-loop power control overhead on the downlink.

8.6.3 Assignment of Uplink Modulation and Coding

Based on various reports sent by the mobile, the base station can make an informed decision on what burst profile (modulation and coding) the mobile should use, as well as on how many subchannels (and hence the number of subcarriers) are allocated to that mobile. Specifically, the mobile reports the value of its current transmit power via the report response (REP-RSP) message to the base station, and the mobile may report the value of its maximum available power via the subscriber station basic capability request (SBC-REQ) message back to the base station. For example, if a mobile reports that its current transmit power is already high and very close to its maximum available power, then the base station may decide to have the mobile throttle back down to a lower-order modulation (e.g., QPSK instead of 16-QAM), or the base station may decide to decrease the number of subchannels allocated to that mobile.

8.6.4 Concluding Remarks

As specified in IEEE 802.16e, either the base station or the mobile can initiate a transition from closed-loop power control to open-loop power control and vice versa (if open-loop power control is supported, of course). But the decision to effect the change ultimately lies with the base station. If the mobile wants a mode change, it sends a power control mode change request (PMC_REQ) message to the base station. If the base station approves (or does not approve) the request, then it sends a power control mode change response (PMC_RSP) message back to the mobile.

On the other hand, if the base station wants a mode change, the base station immediately sends a PMC_RSP message to the mobile, informing the mobile of the change. In addition to communicating the mode change decision from the base station to the mobile, the PMC_RSP message can also carry a power control command for the mobile. Specifically, the PMC_RSP message contains the *power adjust* field, which carries the command for power up or power down for closed-loop power control; the message also carries the $Offset\_BS_{perSS}$ field, which is the offset to the open-loop power control formula (for open-loop power control) in (8.1). Both *power adjust* and $Offset\_BS_{perSS}$ are in multiples of 0.25 dB. This means that the minimum step size of a single power control command is 0.25 dB.

The IEEE 802.16 standard does not specify power control mechanisms for the downlink. But it also does not prohibit base station vendors from incorporating downlink power control in their products. It is important to note that the mobile does provide downlink quality feedback to the base station, so the base station has at its disposal downlink quality measurements upon which it can base its downlink power control decisions.

References

[1] Etemad, K., "Overview of Mobile WiMAX Technology and Evolution," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 31–40.

[2] Ahmadi, S., "An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Communications*, Vol. 47, No. 6, 2009, pp. 84–98.

[3] Laroia, R., S. Uppala, and J. Li, "Designing a Mobile Broadband Wireless Access Network," *IEEE Signal Processing*, Vol. 21, No. 5, 2004, pp. 20–28.

[4] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

[5] Belghith, A., and L. Nuaymi, "WiMAX Capacity Estimations and Simulation Results," *Proc. IEEE Vehicular Technology Conference*, Singapore, May 11–14, 2008, pp. 1741–1745.

[6] Huang, C. Y., et al., "Radio Resource Management of Heterogeneous Services in Mobile WiMAX Systems," *IEEE Wireless Communications,* Vol. 14, No. 1, 2007, pp. 20–26.

[7] Yang, S. C., *3G CDMA2000 Wireless System Engineering*, Norwood, MA: Artech House, 2004.

[8] Cao, Y., and V.O.K. Li, "Scheduling Algorithms in Broad-Band Wireless Networks," *Proceedings of the IEEE*, Vol. 89, No. 1, 2001, pp. 76–87.

[9] Sadr, S., A. Anpalagan, and K. Raahemifar, "Radio Resource Allocation Algorithms for the Downlink of Multiuser OFDM Communication Systems," *IEEE Communications Survey and Tutorials*, Vol. 11, No. 3, 2009, pp. 92–106.

[10] IEEE Standard 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York: IEEE, October 1, 2004.

[11] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[12] Wang, F., et al., "Mobile WiMAX Systems: Performance and Evolution," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 41–49.

[13] Yang, S. C., *CDMA RF System Engineering*, Norwood, MA: Artech House, 1998.

CHAPTER 9

Medium Access Control: Upper Control Plane

9.1 Introduction

In IEEE 802.16e, the control plane has the functions necessary to establish, maintain, and tear down connectivity, and Figure 9.1 shows that the functions in the upper part of the control plane include network discovery and entry, connection management, mobility management, idle mode and paging, sleep mode, MBS, and others [1]. In this chapter, we examine two major functions in the upper control plane: network entry and mobility management.

9.2 Network Entry

After a mobile powers on, it goes through a process to be properly admitted onto the network. The process can be divided into a series of stages [3] shown in Figure 9.2. They are:

- Synchronization with the downlink of the base station and acquisition of parameters;
- Initial ranging;
- Negotiation of mobile capabilities;
- Security procedures;
- Registration;
- IP connectivity;
- Connection setup.

The following sections describe how a mobile would normally step through these stages, starting with downlink synchronization.

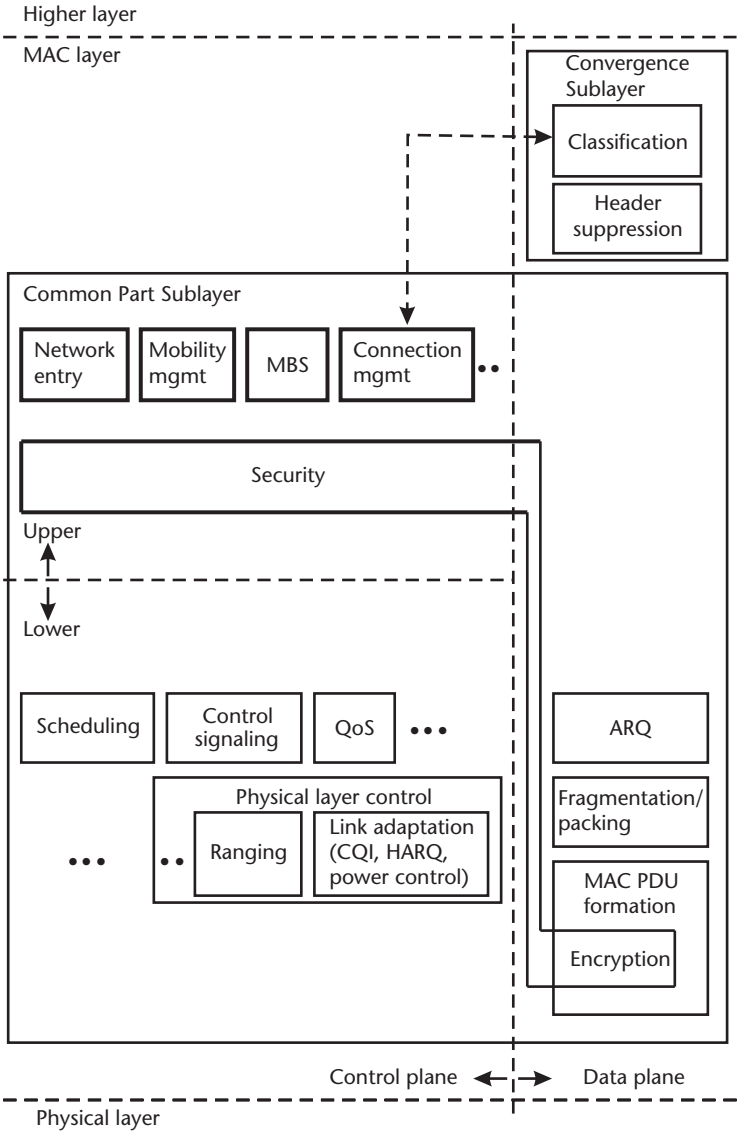


Figure 9.1 MAC protocol structure. Note that the boldfaced parts constitute the upper part of the control plane. (*After:* [1, 2].)

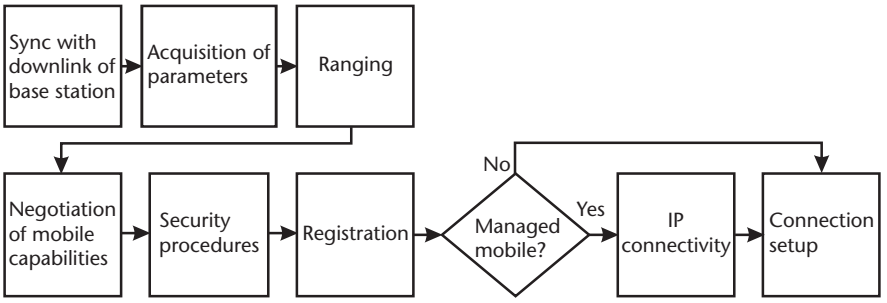


Figure 9.2 Network entry stages.

9.2.1 Synchronization with Downlink of Base Station and Acquisition of Parameters

In this stage, the mobile first scans its operating downlink frequencies to attempt to acquire a downlink channel. Once the physical layer acquires the downlink channel and achieves bit-level synchronization, it would then send the decoded bits to the MAC layer. Reading the bits, MAC would attempt to achieve frame-level synchronization (by using the preamble, for example). After it achieves frame-level synchronization on the downlink, MAC proceeds to recover the DL-MAP message. After receiving the DL-MAP message, MAC would then have in its possession the necessary downlink parameters for system operation.

After obtaining the downlink parameters, the mobile proceeds to acquire the uplink parameters. In doing so, the mobile waits for an uplink channel descriptor (UCD) message, which is a periodic broadcast message from the base station. After receiving the UCD message, the mobile determines if the associated uplink channel is usable. If it is, then the mobile loads the uplink parameters from the message.

At this time, the mobile proceeds to extract slot timing information for this uplink channel. Using the uplink slot information obtained, the mobile is now ready to perform ranging.

9.2.2 Initial Ranging

The mobile needs to perform ranging as part of network entry. In OFDM and OFDMA, ranging is the process by which the mobile obtains adjustments to important uplink transmission parameters, such as timing offset, frequency offset, and transmit power adjustments. Out of these parameters, timing offset is especially important in a TDD system, and ranging allows the mobile to adjust its timing so that its uplink transmissions are aligned with the specified frame [4].

On the downlink of a TDD system, a mobile can easily discern the start of the frame because: (1) there is a preamble at the beginning of the frame, and (2) all transmissions originate from the same source—the base station. On the uplink, however, there is a multitude of mobiles transmitting, and each mobile's transmission experiences a different delay; as such, these mobiles' delays have to be corrected. In ranging, the base station computes the correct timing offset based on a mobile's transmission and tells the mobile what adjustment to timing offset to make. In other words, ranging helps to achieve frame-level synchronization on the uplink. At a minimum, ranging needs to ensure that the mobiles' uplink transmissions arrive within a cyclic prefix of one another. Recall that cyclic prefix is the extra "guard time" that is added to the beginning of each OFDM symbol to reduce interference between successive OFDM symbols. If a mobile's uplink transmission cannot be aligned to within this criterion after a set number of attempts, the base station drops the mobile to prevent it from interfering with other mobiles.

In the OFDMA implementation, the mobile performs initial ranging in the ranging channel; it does so by sending a pseudorandom noise (PN) code. A mobile most likely needs to go through more than one request-response cycle to obtain a satisfactory timing offset. Ultimately, the initial ranging process concludes if the mobile receives the ranging response (RNG-RSP) message indicating a "success" ranging status. The RNG-RSP message contains the basic and primary management

CIDs assigned to the mobile, as well as any other more specific RF-related parameters for the mobile to use. The basic and primary management CIDs will enable the mobile to start exchanging MAC management messages with the base station. See Section 8.5 for the steps in initial ranging.

In determining the transmit power adjustments for its ranging transmission (i.e., RNG-REQ message or PN code), the mobile uses an open-loop scheme by which the maximum transmit power for initial ranging depends on the received power measured at the mobile. The lower the downlink received power measured at the mobile, the higher the mobile's maximum transmit power for initial ranging (and vice versa). The mobile's maximum transmit power for initial ranging is the $P_{TX\_IR\_MAX}$ parameter.

Ranging is an iterative process and typically requires more than one request-response cycle. Throughout this iterative process for OFDMA, the mobile also performs the appropriate closed-loop transmit power adjustment (up or down) for each ranging transmission according what is specified in the received ranging response (RNG-RSP) message. In OFDMA, the PN code is BPSK modulated onto the ranging channel.

9.2.3 Negotiation of Mobile Capabilities

Once the mobile finishes the ranging process, it proceeds to transmit a SBC-REQ message to the base station. The mobile uses the SBC-REQ message to communicate its capabilities to the base station. Upon receiving the SBC-REQ message, the base station determines what capabilities (of the mobile) should be enabled; then using the subscriber station basic capabilities response (SBC-RSP) message, the base station communicates those capabilities (that should be enabled) back to the mobile. The capabilities communicated are mostly related to the physical layer, such as which FFT sizes does the mobile support, which STBC and MIMO modes does the mobile support, and what is the mobile's maximum transmit power.

The term "negotiation" at this stage is a bit of a misnomer in that the mobile does not really negotiate with the base station at all regarding the mobile's capabilities. Instead, the mobile tells the base station what capabilities the mobile has, and the base station tells the mobile what capabilities the mobile can enable. In doing so, the base station can deny the use of any capabilities that the mobile possesses [3].

9.2.4 Security Procedures

Before being allowed onto the network, the mobile normally goes through a set of security procedures with the base station. The security procedures consist of:

- Authorization of the mobile;
- Distribution of key materials.

The security procedures are actually only performed if the base station and the mobile both support authorization policy and privacy key management (PKM) protocol. Otherwise, they are not performed [5]. Specifically, the mobile indicates

whether or not it supports authorization policy by using the *authorization policy support* field in the SBC-REQ message sent to the base station (in the negotiation stage). It is expected, however, that most network service providers and/or access providers serving paying subscribers would require their mobile devices to support authorization policy. Authorization and distribution of key materials are described in more detail in Chapter 12.

9.2.5 Mobile Registration

To register itself with the network and to gain entry onto the network, the mobile sends a registration request (REG-REQ) message to the base station. In that REG-REQ message, the mobile indicates whether or not it is a "managed" mobile. A managed mobile means that the mobile device can be administered by using standard IP management protocols (e.g., Simple Network Management Protocol or SNMP).

If the mobile is managed, the base station sends a REG-RSP message back to the mobile. Then the mobile proceeds first to IP connectivity. If the mobile is not managed, the base station sends a registration response (REG-RSP) message back to the mobile. Then the mobile proceeds directly to the connection setup.

9.2.6 IP Connectivity

If the mobile is managed, then there needs to be an extra management connection, called the secondary management connection, between the base station and that mobile. Using the secondary management connection, the mobile can exchange standard IP management traffic (e.g., DHCP, TFTP, and SNMP) with the base station. This way, the mobile effectively becomes a standard IP-based device that the base station can administer.

To obtain that secondary management connection, the mobile uses the REG-REQ message (sent previously in the registration stage) to carry out the following functions:

- The REG-REQ message requests the secondary management CID for the mobile. Once the mobile has the secondary management CID, it can establish the secondary management connection per that CID.
- The REG-REQ message may include the IP version parameter that the mobile can support (on the secondary management connection). Upon receiving this information in the REG-REQ message, the base station responds with an IP version that the mobile has to use.

In the REG-RSP message (also sent previously in the registration stage), the base station sent information such as the secondary management CID and the IP version parameter to the mobile. Using the secondary management CID, the mobile establishes the secondary management connection and uses that connection to request an IP address from the base station. If the mobile uses IPv4, then it obtains an IP address using the standard Dynamic Host Control Protocol (DHCP) from the base station (or more accurately, from a DHCP server through the base station). If

the mobile uses IPv6, then it obtains an IP address using either DCHPv6 or IPv6 Stateless Address Autoconfig [5].

After the mobile obtains an IP address, the mobile would normally proceed to download a configuration file from the base station. The configuration file contains a set of configuration settings for the mobile. The settings mostly have to do with parameters for those layers above MAC, like the TFTP server timestamp or the software (upgrade) server IP address. The configuration file is transferred over the secondary management connection using the Trivial File Transfer Protocol (TFTP).

After the mobile finishes downloading the configuration file, the mobile sends a TFTP complete (TFTP-CPLT) message back to the base station. After receiving the TFTP-CPLT message, the base station responds with a TFTP response (TFTP-RSP) message. An "OK" response in the TFTP-RSP message means that the base station confirms that the mobile has successfully downloaded the configuration file.

Note that IP connectivity is not established for those mobiles that are not managed.

9.2.7 Connection Setup

At this point, the base station proceeds to establish provisioned service flows and connections for the mobile. Because the base station is granting the request made by the mobile to enter the network here, the base station itself initiates the creation of service flows. The base station does so by sending the dynamic service addition request (DSA-REQ) message(s) to the mobile. It is important to note that each DSA-REQ message contains only one SFID, so multiple service flows require multiple DSA-REQ messages.

For each DSA-REQ message received, the mobile responds by sending a dynamic service addition response (DSA-RSP) message back to the base station. So again, multiple service flows require multiple DSA-RSP messages. For each DSA-RSP message received, the base station sends a dynamic service addition acknowledgment (DSA-ACK) message back to the mobile. After the mobile receives a DSA-ACK message, the associated service flow is established.

9.3 Mobility Management: Link Handover

To a system engineer, maximizing the handover performance at cell and sector boundaries is one of the most challenging tasks. At the link level, handover is defined as the process by which a mobile transitions its link and connection from one base station (called the *serving base station*) to another base station (called the *target base station*). See Figure 9.3. Ensuring that the connection is active and not dropped during handover is important in maintaining system quality as experienced by the user. Therefore, an understanding of the mechanics of the handover process is important.

The handover functionality at the link level is incorporated in the IEEE 802.16e standard since it pertains to a broadband *mobile* system (whereas IEEE 802.16-2004 did not specify any handover functionality). Specifically, the standard specifies three handover processes: hard handover (HHO), macro diversity handover (MDHO), and fast base station switching (FBSS). As described in the following

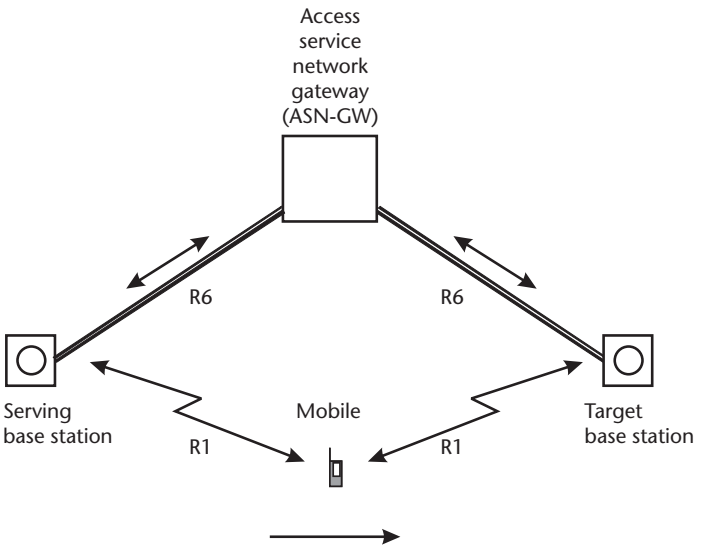


Figure 9.3 Handover.

sections, HHO is the traditional handover method where a handover decision is made and the mobile transitions from the current serving base station to the new target base station. MDHO is similar (but not identical) to soft handover in cellular CDMA systems where the mobile maintains radio links with more than one base station during handover. FBSS is functionally a faster version of hard handover.

Regardless of the specific handover process adopted, a mobile needs to perform certain prerequisite activities prior to executing an actual handover. These prerequisite activities are carried out in a stage, called the cell reselection stage, of the handover process. We first examine the activities carried out in the cell reselection stage, and then consider each one of the three handover processes.

9.3.1 Cell Reselection

Prior to executing an actual handover, both the base station and the mobile carry out regular housekeeping activities. First, the base station sends a list of its neighbors (i.e., neighboring base stations) in the neighbor advertisement (MOB_NBR-ADV) message to the mobile. In addition to the list of neighbors, the MOB_NBR-ADV message also contains channel information (the same kind of information found in DCD and UCD messages) of the neighbors. This way, the mobile does not have to specifically read the DCD and UCD messages from neighboring base stations.

Second, the mobile requests that the base station allocates a period of time (called *scanning interval*) in which the mobile can scan the transmissions of neighboring base stations to see if any of them is a good candidate for handover. The mobile may request the allocation of more than one scanning interval so it does not have to keep going back to the base station for more scanning intervals. The mobile makes the request by sending the scanning interval allocation request (MOB_SCN-REQ) message to the base station. The base station responds by sending the scanning interval allocation response (MOB_SCN-RSP) message to the mobile. Then the mobile commences scanning during the allocated scanning interval(s). If the

base station asks for the scanning results (in the MOB_SCN-RSP message), the mobile uses the scanning result report (MOB_SCN-REP) message to report the scanning results, which may include neighboring base stations' CINRs, RSSIs, or delays.

During the scanning interval, the mobile may optionally "associate" itself with neighboring base station(s). Association involves initial ranging, and the purpose of association is to perform initial ranging, acquire ranging parameters, and facilitate the eventual handover process. Three levels of association exist; they are:

- *Association level 0:* Association without coordination;
- *Association level 1:* Association with coordination;
- *Association level 2:* Network-assisted association reporting.

After a mobile performs initial ranging with a particular neighboring base station, that base station responds with a RNG-RSP message. If the RNG-RSP message indicates that the base station can provide full service (i.e., the field *service level prediction* = 2 in the RNG-RSP message) to the mobile, then the mobile records that base station in its local association table [5].

9.3.1.1 Association Level 0—Association Without Coordination

At this level, the mobile performs initial ranging with neighboring base station(s) in its allocated intervals. But the mobile can only range on a contention basis. This is because, at association level 0, there is no network coordination of the mobile's ranging attempt. As such, the probability of collision may be significant.

9.3.1.2 Association Level 1—Association with Coordination

At this level, the mobile also performs initial ranging with neighboring base station(s) in its allocated intervals. But the mobile's serving base station coordinates between the mobile and the neighboring base stations. What this means is that the serving base station makes sure that each neighboring base station assigns (to the mobile):

- A "rendezvous time" (which marks the start of an important UL-MAP message to be sent by a neighboring base station later);
- A transmission opportunity (inside a ranging region to be assigned by a neighboring base station);
- A unique PN code.

The serving base station also makes sure that the ranging region to be assigned by a neighboring base station does not conflict with those of other neighboring base stations. Next, the serving base station uses the MOB_SCN-RSP message to send these preassigned parameters (for all the neighboring base stations) to the mobile [5].

After receiving these parameters, the mobile synchronizes to the downlink of a neighboring base station and waits for the rendezvous time of that neighboring

base station. The rendezvous time is when the neighboring base station will send a UL-MAP message, and that UL-MAP message assigns a ranging region, which the mobile may use. After extracting the ranging region from the UL-MAP message, the mobile ranges by transmitting the assigned PN code at the transmission opportunity in the ranging region [5].

How is association with coordination initiated? Well, the mobile may use the MOB_SCN-REQ message to request association with coordination (by indicating the appropriate scanning type in the message). The serving base station may also initiate this type of association by unilaterally sending a MOB_SCN-RSP message to the mobile.

The idea behind association with coordination is that if the serving base station can coordinate ranging regions and transmission opportunities, then the probability of collision (in the initial ranging process) can be minimized. Note that the probability of collision is not eliminated because if a base station assigns an identical transmission opportunity (in the same ranging region) to more than one mobile, then the probability of collision is nonzero as more than one mobile may transmit in the same transmission opportunity.

9.3.1.3 Association Level 2—Network-Assisted Association Reporting

Association level 2 is similar to association level 1 discussed above with one important enhancement. Instead of collecting the RNG-RSP message from each neighboring base station with which it ranges, the mobile simply waits for a single message from the serving base station. Here the serving base station gathers, over the fixed network, all the ranging-related information from a mobile's neighboring base stations. Then the serving base station reports all that information to the mobile using a single association result report (MOB_ASC_REPORT) message.

Similar to association with coordination, the mobile may use the MOB_SCN-REQ message to request network assisted association reporting (by indicating the appropriate scanning type in the message). The serving base station may also initiate network assisted association by unilaterally sending a MOB_SCN-RSP message to the mobile.

9.3.1.4 Remarks

After carrying out scanning and/or association, the mobile (and the base station) are now ready to go into the actual handover process. The following sections examine each of the three handover processes (i.e., HHO, MDHO, and FBSS) in more detail, starting with HHO.

9.3.2 Hard Handover (HHO)

For the hard handover process, the standard specifies additional stages (in addition to cell reselection), which the mobile and the base station go through. They are [5]:

- Handover decision and handover initiation;

- Synchronization with the downlink of the target base station and acquisition of parameters;
- Ranging;
- Other network entry/reentry procedures;
- Termination with the serving base station.

Out of these stages, synchronization with the downlink of the target base station, acquisition of parameters, ranging, and other network entry/reentry procedures are actually similar to the corresponding activities carried out in the initial network entry process discussed earlier. These stages of HHO process are examined next.

9.3.2.1 Handover Decision and Handover Initiation

Based on the results of the mobile's scanning and/or association, the mobile or the base station makes a *decision to handover* (from the serving base station to a target base station) and starts the HHO process. The evidence that a handover decision has been made is the transmission of the mobile station handover request (MOB_MSHO-REQ) message or the base station handover request (MOB_BSHO-REQ) message. Specifically, if the decision originates at the mobile, the mobile can transmit the MOB_MSHO-REQ message; if the decision originates at the base station, the base station can transmit the MOB_BSHO-REQ message. The transmission of either message indicates that the handover process has started. It is important to note that the *choice of a target base station* has not been made at this point. This is because either the MOB_MSHO-REQ or the MOB_BSHO-REQ message can contain more than one possible target base station.

What if both the mobile and the base station transmit a message at the same time? The rules for resolving the conflict are:

- If the mobile receives a MOB_BSHO-REQ message after it already sent a MOB_MSHO-REQ message, the mobile ignores the MOB_BSHO-REQ message.
- If the base station receives a MOB_MSHO-REQ message after it already sent a MOB_BSHO-REQ message, the base station ignores its own MOB_BSHO-REQ message.
- If the base station receives a MOB_HO-IND message (to be discussed below) after it already sent a MOB_BSHO-REQ message, the base station ignores its own MOB_BSHO-REQ message.

As one can see, the process is biased toward the mobile because it is assumed that the mobile is the entity that can best determine its own handover disposition.

Figure 9.4 shows the normal exchange of messages at this stage. If the mobile initiates the handover process, then the mobile sends a MOB_MSHO-REQ message, which may contain a list of target base stations that the mobile is currently considering. The base station responds with a base station handover response (MOB_BSHO-RSP) message; this message may contain a list of target base stations

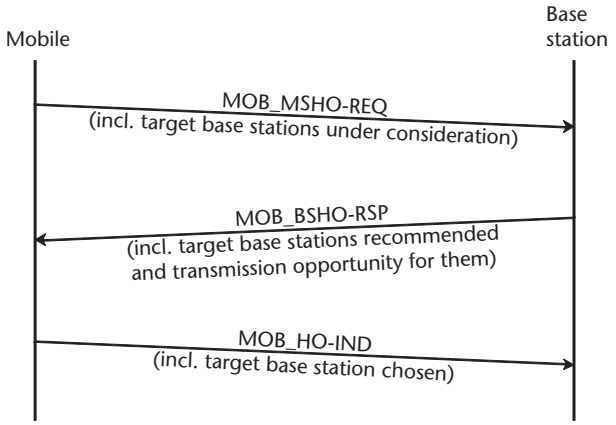


Figure 9.4 The mobile initiates handover.

that the base station is recommending, as well as these target base stations' predicted service levels. Also, the MOB_BSHO-RSP message may provide the mobile with a time of dedicated transmission opportunity for fast ranging that is common to all potential target base stations. (This time is negotiated by the serving base station with other target base stations over the fixed network.) In addition, the base station may use the MOB_BSHO-RSP message to force the mobile to handover. Taking all that information into account, the mobile makes a choice of a target base station. Note that the mobile does not have to pick a target base station from one on the recommended list (even in the case of a forced handover). When the mobile is about to perform a handover, it sends a handover indication (MOB_HO-IND) message, which contains the base station ID (i.e., *Target_BS_ID*) of the target base station [5].

Figure 9.5 shows the normal exchange of messages when the base station initiates the handover process. If the base station initiates, the base station sends a MOB_BSHO-REQ message, which may contain a list of target base stations that the base station is recommending and their respective predicted service levels. The serving base station may also negotiate (with other target base stations) the time of dedicated transmission opportunity for fast ranging common to all potential target base stations and include that time in the MOB_BSHO-REQ message. The contents of the MOB_BSHO-REQ are almost identical to those of the MOB_BSHO-RSP

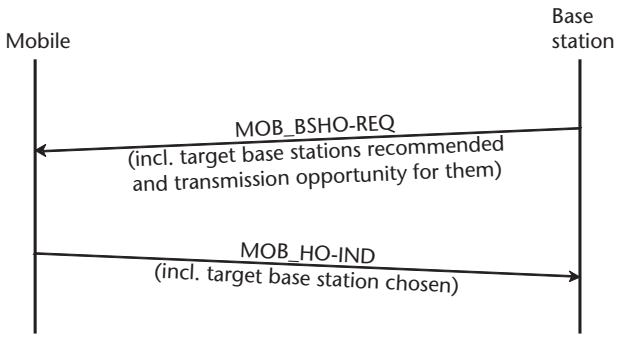


Figure 9.5 The base station initiates the handover.

message (seen above). The mobile takes into account information contained in the MOB_BSHO-REQ message and makes a choice of a target base station. When the mobile is about to perform a handover, it sends a handover indication (MOB_HO-IND) message containing the base station ID (i.e., *Target_BS_ID*) of the target base station [5]. Again, the mobile does not have to pick a target base station from one in the recommended list (even if the base station forces a handover by using the MOB_BSHO-REQ message).

Note that the serving base station may forward the mobile's context information to the potential target base stations for the purpose of facilitating the handover process.

9.3.2.2 Synchronization with Downlink of Target Base Station and Acquisition of Parameters

The mobile is now ready to start communicating with the target base station. This stage (synchronization with downlink of target base station and acquisition of parameters), the next stage (ranging), and the stage afterwards (other network entry/reentry procedures) together are similar to the network entry process discussed earlier. Under normal circumstances (assuming no handover cancellation or handover rejection), the mobile first synchronizes to the downlink of the target base station and obtains all the necessary parameters (for downlink and uplink) from the DL-MAP, UL-MAP, DCD, and UCD messages. The mobile may not have to decode as many messages from the target base station if it already has some of the parameters from the MOB_NBR-ADV message received previously.

9.3.2.3 Ranging

In this stage, the mobile conducts ranging with the target base station as it normally would. This procedure may be facilitated if the target base station has already allocated a transmission opportunity for ranging. If this is the case, then the target base station would allocate that opportunity to the mobile using its UL-MAP message.

In addition, when the mobile transmits the RNG-REQ message, the mobile would set bit #0 of the type/length/value (TLV) parameter *ranging purpose indication* to 1 and include a serving BSID TLV in the RNG-REQ message. These two actions together let the target base station know that a handover attempt is in progress.

9.3.2.4 Other Network Entry/Reentry Procedures

Theoretically, to handover to a new target base station, the mobile needs to go through all the procedures in the network entry process, which include synchronization to the downlink, acquisition of parameters, and ranging, as well as

- Negotiation of basic capabilities;
- Security procedures;
- Registration.

To save time and the amount of message transmissions, the system allows for something called *handover (HO) process optimization* by which one or more of these procedures may be omitted. Using HO process optimization, the base station can tell the mobile which procedure to omit by setting different bits in the HO process optimization TLV in the RNG-RSP message. For example, if the target base station already has all the basic capabilities information on the mobile, it can set bit #0 to 1 in the HO process optimization TLV in the RNG-RSP message. Which procedure(s) to omit really depends on how much information about the mobile the target base station already possesses (based on information received from the serving base station over the fixed network). In fact, there is a total of 13 bits that can be set to enable different combinations of HO process optimization.

For system engineers, the presence of HO process optimization makes troubleshooting layer 2 during handover slightly more complicated. This is because during handover, some management messages may appear to be missing, when in fact they are not transmitted because some aspects of HO process optimization are enabled. Thus, before diagnosing a handover, one should examine what bit is set in the HO process optimization TLV. More advanced computer-aided diagnostic tools should help in this respect.

Network reentry ends after the last procedure (i.e., registration) is finished and the provisioned connections are established.

9.3.2.5 Termination with the Serving Base Station

The mobile's relationship with the serving base station effectively concludes at the end of this stage. The mobile terminates with the serving base station by marking the field *HO_IND_type* = "serving BS release" in the MOB_HO-IND message and sending the message to the serving base station. Upon receiving the MOB_HO-IND message, the serving base station waits for a message (over the fixed network) from the target base station letting the serving base station know that the mobile has been attached to the target base station. Once the (old) serving base station receives that message from the target base station, it not only severs both the physical (radio) link and the logical connections with the mobile, but also deletes the context related to those connections, such as timers, counters, and buffered PDUs destined for the mobile.

What if the serving base station never receives the message from the target base station signifying that the mobile has been attached to the target base station? It turns out there is a timer (i.e., resource retain timer) that governs how long the context information of the mobile is retained; the timer starts running right after the serving base station receives the MOB_HO-IND message indicating release. If the timer expires without the receipt of that message from the target base station, then the serving base station still tears down the link and the connections and deletes the context of the mobile.[1]

1. The *resource retain flag* field (= 1) in the MOB_BSHO-REQ or the MOB_BSHO-RSP message tells the mobile that the base station will retain context information of the mobile for the duration of the resource retain timer.

9.3.3 Macro Diversity Handover (MDHO)

MDHO is an optional handover mode in which the mobile communicates traffic simultaneously with more than one base station during the handover. If MDHO is enabled (as indicated in the REG-REQ and REG-RSP messages), then the mobile may carry out the MDHO process, which consists of three stages [5]:

- MDHO decision;
- Diversity set selection and update;
- Anchor base station selection and update.

Before examining these three stages in more detail, we define two important terms: diversity set and anchor base station. In MDHO, a mobile may simultaneously communicate traffic with more than one base station. The group of base stations with which the mobile communicates traffic is called the *diversity set*. When the mobile is communicating with only one base station, the diversity set consists of only one base station. That single base station is known as the *anchor base station*.[2] Note that if the diversity set contains several base stations, one of them would still be labeled the anchor base station.

During MDHO on the downlink, the mobile performs diversity combining of the transmissions sent by base stations in the diversity set. Figure 9.6 illustrates

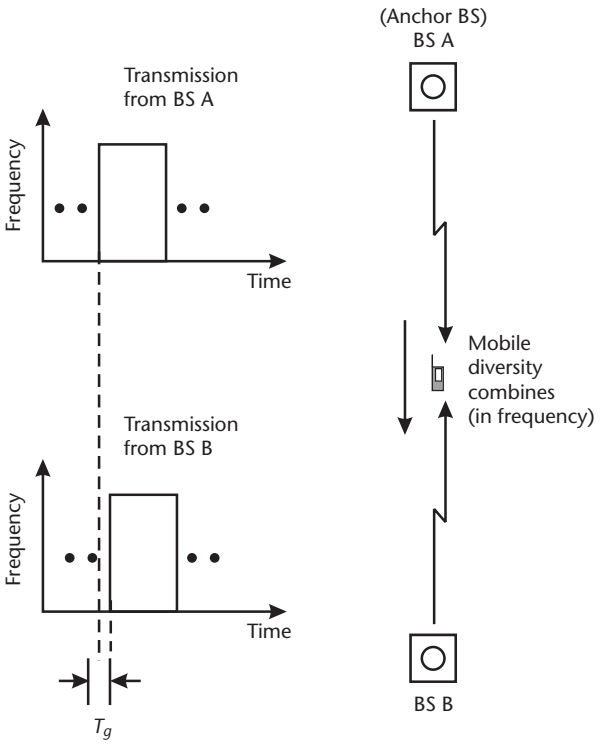


Figure 9.6 MDHO on the downlink.

2. In this case, it is also by definition the serving base station.

MDHO on the downlink. As the mobile moves from base station A (BS A) to base station B (BS B), the mobile diversity combines transmissions from both base station A and base station B. On the downlink, two important requirements exist for MDHO: (1) base stations in the diversity set synchronize their transmissions mobile; such synchronization helps ensure that (2) the base stations' frames arrive at the mobile within the interval of cyclic prefix time (T_g) [5], so the mobile can diversity combine the transmissions in real time. Note that in Figure 9.6, the diversity set consists of both base station A and base station B, but there is only one anchor base station (base station A). To enable diversity combining, the base stations in the diversity set must send identical PDUs to the mobile.

On the uplink, an entity (a controller on the fixed network or the anchor base station) performs selection combining of the transmissions received by the base stations in the diversity set.

9.3.3.1 MDHO Decision

When a mobile's diversity set consists of only one base station, then the mobile is communicating with only one base station and is not in handover. If a mobile's diversity set consists of more than one base station, then the mobile is in MDHO.

In managing the diversity set, the mobile uses two thresholds of CINR: the *H_Add* threshold and *H_Delete* threshold. If the CINR of a nearby base station goes above *H_Add*, then the mobile sends a MOB_MSHO-REQ message to request that the base station be added to the diversity set. If the CINR of a base station goes below *H_Delete*, then the mobile sends a MOB_MSHO-REQ message to request that the base station be deleted from the diversity set. These two thresholds are sent to the mobile through the DCD message. The impacts of adjusting *H_Add* and *H_Delete* thresholds are discussed later in this chapter.

Either the mobile or the base station may make the decision to commence the MDHO process. If the decision originates at the mobile, the mobile can transmit the MOB_MSHO-REQ message; if the decision originates at the base station, the base station can transmit the MOB_BSHO-REQ message. The exchange of messages in MDHO is similar to those shown previously in Figures 9.4 and 9.5. Again, if both the mobile and the base station transmit a message at the same time, then the rules for resolving the conflict still favor the mobile. Namely, the base station's MOB_BSHO-REQ message is ignored in favor of the mobile's MOB_MSHO-REQ message or MOB_HO-IND message.

Of the base stations in the diversity set, one base station is designated as the anchor base station. For control information, the mobile may monitor only the anchor base station's DL-MAP, UL-MAP, and FCH, which contain the necessary control information for the mobile to communicate traffic with all base stations in the diversity set. Alternatively, the mobile may monitor DL-MAP, UL-MAP, and FCH sent by all base stations in the diversity set.

A prerequisite for carrying out MDHO is that all the base stations in the diversity set need to have the mobile's context information (including security information). This way, the mobile is registered with and is able to communicate traffic with all base stations in the diversity set.

9.3.3.2 Diversity Set Selection and Update

If the mobile initiates the handover process, it sends a MOB_MSHO-REQ message, which can have a list of possible base stations to be included in its diversity set. (That list may include the base station whose CINR has just surpassed *H_Add*.) The mobile can generate this list of possible base stations for the diversity set based on its own CINR measurement, scanning, and/or association. It is important to recognize that the actual update of the diversity set is not yet finalized at this point. After receiving the MOB_MSHO-REQ message, the anchor base station (or base stations in the diversity set) may modify the list and send a MOB_BSHO-RSP message, which may have a list of recommended base stations to be included in the diversity set. The base station can generate this list of recommended base stations for the diversity set based on expected QoS performance of different neighboring base stations and their ability to carry out the specific handover process (MDHO or FBSS).

The mobile can accept or reject the list of recommended base stations by sending a MOB_HO-IND message with the *MDHOFBSS_IND_type* field set to "confirm diversity set update" or "reject diversity set update." If the mobile rejects the list, the base station(s) may change the list and resend the MOB_BSHO-RSP message.

If the base station initiates the handover process, the anchor base station (or base stations in the diversity set) sends a MOB_BSHO-REQ message, which may have a list of recommended base stations to be included in the diversity set, but the actual update of the diversity set is also not yet finalized. The base station can generate this list of recommended base stations for the diversity set based on expected QoS performance of different neighboring base stations and their ability to carry out the specific handover process (MDHO or FBSS).

The mobile can accept or reject the list of recommended base stations by sending a MOB_HO-IND message with the *MDHOFBSS_IND_type* field set to "confirm diversity set update" or "reject diversity set update." If the mobile rejects the list, the base station(s) may change the list and resend the MOB_BSHO-REQ message.

9.3.3.3 Anchor Base Station Selection and Update

In some ways, the selection and update of the anchor base station within the diversity set is when the "real" handover takes place. Ideally, when a mobile is near a base station, its diversity set consists of only that base station (i.e., anchor base station). As the mobile begins to leave the coverage area of the anchor base station, other base stations' signals become stronger, and the mobile's diversity set starts to admit these other base stations. At some point during MDHO, a new anchor base station is selected and updated in the diversity set. As the mobile continues to travel towards the new anchor base station, the signals from other base stations get weaker; these other base stations get deleted from the mobile's diversity set, and eventually the mobile's diversity set will again consist of only one base station (i.e., the new anchor base station).

The procedure for the selection and update of the anchor base station is similar to the selection and update of the diversity set. Based on signal measurements, the mobile picks a preferred anchor base station; it communicates that preference to

the base station by sending a MOB_MSHO-REQ message. But the actual update of the anchor base station has not yet occurred at this point. After receiving the MOB_MSHO-REQ message, the base station makes a decision on the actual anchor base station update and sends a MOB_BSHO-RSP message to the mobile. The base station makes the update decision taking into consideration signal measurements reported by the mobile. The base station can also communicate its decision to the mobile using the MOB_BSHO-REQ message.

Upon receiving the MOB_BSHO-RSP or MOB_BSHO-REQ message, the mobile can accept or reject the update of the new anchor base station. If the mobile accepts the update, then it sends a MOB_HO-IND message with the *MDHOFBSS_IND_type* field set to "confirm anchor BS update" and updates to the new anchor base station at the time shown by the action time field of the MOB_HO-IND message. If the mobile rejects the update, then it sends a MOB_HO-IND message with the *MDHOFBSS_IND_type* field set to "reject anchor base station update." If the mobile rejects the update, the base station may change the anchor base station and resend the MOB_BSHO-RSP or MOB_BSHO-REQ message.

In addition, there is another (faster) implementation of the selection and update procedure, which is summarized as follows: If the mobile has a new base station that it prefers to be the anchor (based on signal measurements), the mobile can immediately transmit an anchor switch indicator in the periodic CINR report with the fast feedback channel (CQICH) and start a switching timer. Before the expiration of the switching timer, the anchor base station may send an Anchor_BS_Switch_IE (in UL-MAP) either to acknowledge the mobile's anchor switch indicator or to cancel it. If the mobile receives no explicit cancellation from the base station before the expiration of the switching timer, then the mobile switches to the new anchor base station.

9.3.4 Fast Base Station Switching (FBSS)

FBSS is another optional handover mode. In FBSS, the mobile quickly changes its anchor base station from one to another within the diversity set. Figure 9.7 illustrates FBSS on the downlink. As the mobile moves from base station A (BS A) to base station B (BS B), the mobile quickly switches receiving from the old anchor base station (BS A) to the new anchor base station (BS B). Note that in Figure 9.7, the diversity set consists of both base station A and base station B, although there is one and only one anchor base station at any given time.

FBSS is enabled or disabled by the exchange of the REG-REQ and REG-RSP messages; similar to MDHO, the FBSS process also consists of three stages [5]:

- FBSS decision;
- Diversity set selection and update;
- Anchor base station selection and update.

In FBSS, the concept of the anchor base station is the same as that in MDHO, but the concept of the diversity set is a bit different. Whereas in MDHO the diversity set is the group of base stations with which the mobile communicates traffic, in FBSS the diversity set is the group of base stations to one of which the mobile is

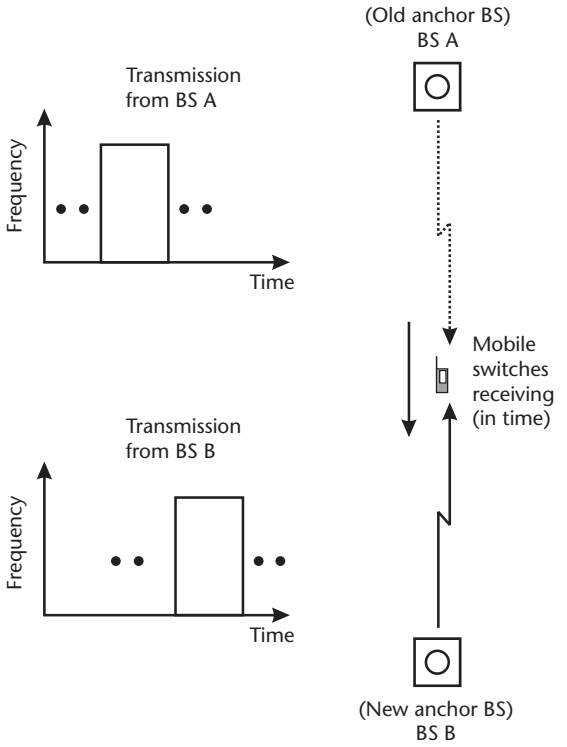


Figure 9.7 FBSS on the downlink.

ready to switch. In other words, in FBSS a mobile only communicates traffic with the anchor base station, but the mobile still maintains a (diversity) set of base stations; the mobile can quickly switch anchors to one of these base stations without having to go through the entire HHO process.

The three stages of FBSS listed above are similar to those in MDHO, so we briefly highlight the parts that are important to FBSS. First, regarding the diversity set in FBSS, the anchor base station is the one with which the mobile communicates traffic and management messages.

Second, in FBSS, the mobile again uses two thresholds: the *H_Add* threshold and *H_Delete* threshold to manage the diversity set. These two thresholds are sent to the mobile through the DCD message. If the CINR of a nearby base station goes above *H_Add*, then the mobile sends a MOB_MSHO-REQ message to request that the base station be added to the diversity set. If the CINR of a base station goes below *H_Delete*, then the mobile sends a MOB_MSHO-REQ message to request that the base station be deleted from the diversity set. After receiving the MOB_MSHO-REQ message, the anchor base station uses the MOB_BSHO-RSP message to send the updated diversity set to the mobile.

Third, a prerequisite for carrying out FBSS is again that all the base stations in the diversity set need to have the mobile's context information (including security information). This way, the mobile is registered with all base stations in the diversity set and can quickly execute switching from one anchor base station to another.

9.3.5 System Design Issue: *H_Add* and *H_Delete*

The effects of relative levels of *H_Add* and *H_Delete* thresholds are analyzed in more detail in this section. For the sake of brevity, the exchange of handover messages leading to the confirmation of the diversity set update is not emphasized here. Instead we highlight the effects of adjusting *H_Add* and *H_Delete*, and it is assumed that the diversity set is updated relatively quickly once a relevant threshold is crossed.

Figure 9.8 shows three scenarios where *H_Add* is fixed at a relatively low CINR level and *H_Delete* is changed from high (top graph) to low (bottom graph). In these scenarios, the mobile moves from the surrounding area of base station A toward the surrounding area of base station B. Thus, the CINR from base station A ($CINR_A$) gradually decreases while the CINR from base station B ($CINR_B$) gradually increases. As one can see, as *H_Delete* is set at lower and lower levels; the distance over which the diversity set has both base stations (i.e., {AB}) increases.

However, it is usually not desirable to set *H_Add* at such a low level because a low *H_Add* can easily admit (into the diversity set) base stations whose CINRs are

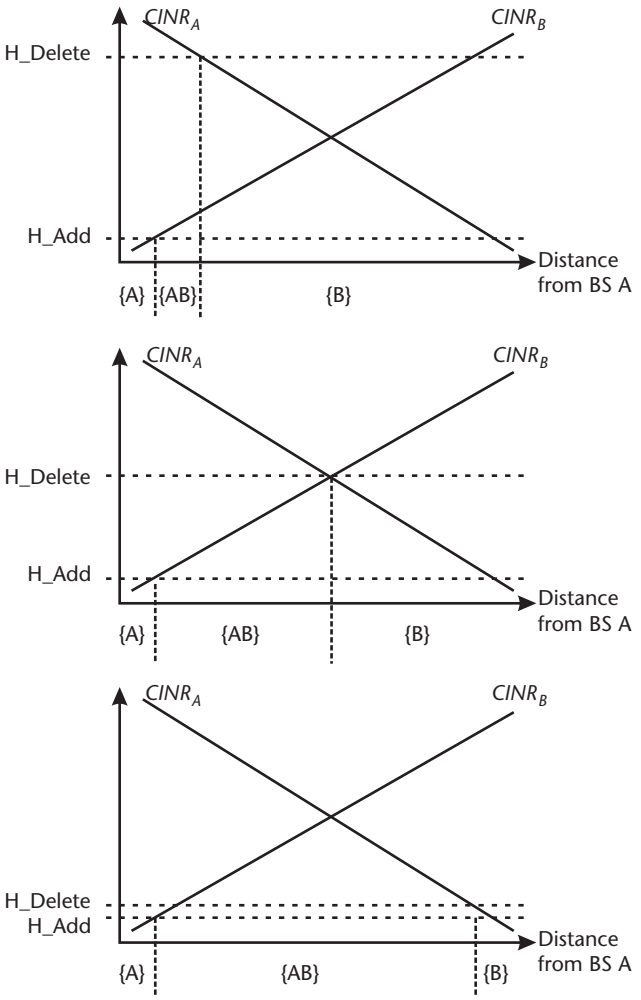


Figure 9.8 Fixing *H_Add* while changing *H_Delete*. { } denotes the diversity set.

low and unusable. As such, Figure 9.9 shows those scenarios where *H_Delete* is fixed at a relatively low CINR level and *H_Add* varies from high (top graph) to low (bottom graph). (*H_Delete* is usually set at a level that is the minimum required received CINR at the mobile plus some implementation margin.) As *H_Add* is set lower, the distance over which the diversity set has two base stations increases.

In system design, a fundamental tradeoff concerning handover exists between two goals:

- To increase reliability, one can enlarge the area (between neighboring base stations) where the diversity set has two or more base stations. But a large area means a mobile will spend more time in that area and occupy those base stations' resources longer.
- To free up resources provided by neighboring base stations, one can reduce the area in which the diversity set has two or more base stations. But a smaller such area means that the probability of a drop is higher.

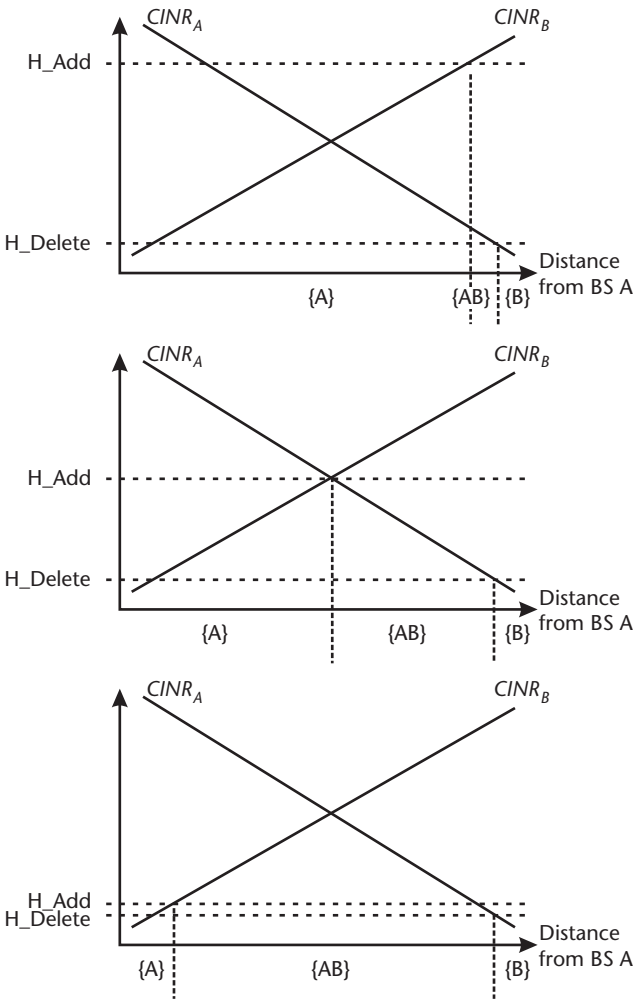


Figure 9.9 Fixing *H_Delete* while changing *H_Add*. { } denotes the diversity set.

The scenarios shown in Figure 9.8 and 9.9 illustrate that a system designer can adjust the relative levels of *H_Add* and *H_Delete* to obtain a desirable area where the diversity set has two (or more) base stations. Figure 9.10 shows two sample settings of *H_Add* and *H_Delete*, which position such area (of a suitable size) at approximately halfway between base station A and base station B.

Of course, one can also enlarge or reduce the area (over which the diversity set has two or more base stations) by adjusting the transmit powers at the neighboring base stations.

9.3.6 Concluding Remarks

System designers are well aware of the fact that while transitioning between cells/sectors, the mobile is vulnerable to drops. During handover, the mobile can indicate a drop (1) by its inability to detect and demodulate the downlink and (2) by exceeding the retry limit on RNG-REQ for periodic ranging; during handover, the base station can indicate a drop by exceeding the retry limit on inviting ranging requests for periodic ranging [5].

In first generation (1G) and early second generation (2G) systems, handover decisions were made by the base station. A controller (on the fixed network) examines a mobile's received signal strengths at different base stations; using only those pieces of *uplink* information, the base station makes the handover decision. In later 2G and 3G systems, the mobile started to participate more in the handover decision, for example in mobile-assisted handover (MAHO). More commonly, the mobile reports scanning results of its neighboring base stations to the serving base station, and the base station makes use of these pieces of *downlink* information, but the ultimate handover decision still predominantly lies with the base station.

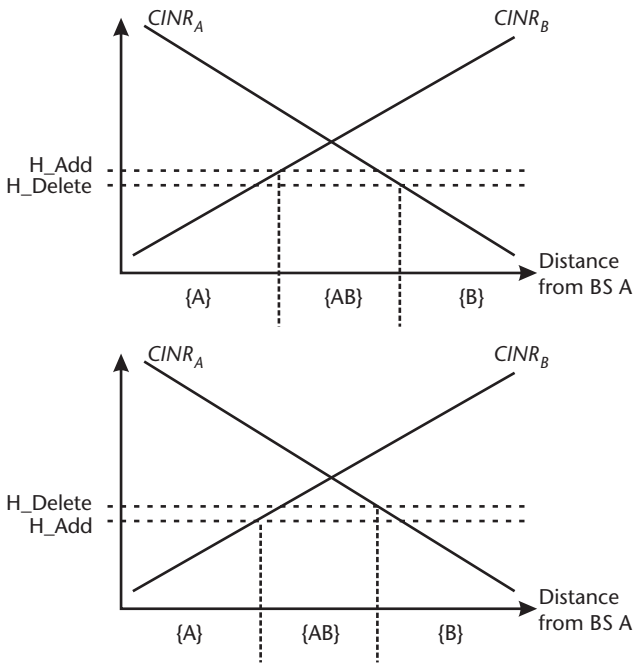


Figure 9.10 Two sample settings of *H_Add* and *H_Delete*. { } denotes the diversity set.

In IEEE 802.16e, the standard explicitly states that either the mobile or the base station may make the handover decision. Indeed, the trend with fourth generation (4G) systems is to have the mobile scan neighboring base stations, make the handover decision itself, and inform the base station of its decision. This is possible because the mobile now has more computational power than before and is capable of processing the relevant parameters and making the decision in real time.

9.4 Mobility Management: Network Handover

The procedures described in Section 9.3 enable *link-level* handover, which allows a mobile to transition its radio link from base station to base station. It is the mobility management function in layer 2 (MAC layer) that handles the base station-to-base station handover, which occurs regardless of any changes in foreign agent (in mobile IP) or IP subnet/prefix.

If there is a change in foreign agent or IP subnet/prefix (e.g., if the mobile travels to a different network), then *network-level* handover procedures are needed to transition a mobile from one network to another network. In this case, some process in layer 3 would handle network-to-network handover, and such a process is typically based on client or proxy mobile IP for IPv4 or mobile IP for IPv6 [1].

In mobile IP for IPv4 [6, 7], when the mobile travels from the *home* network to a *foreign* network, a packet destined for the mobile would travel to the mobile's home network and be intercepted by the home network's home agent. Then the home agent forwards the packet to the foreign network's foreign agent, which forwards the packet to the mobile (currently attached to the foreign network). To forward the packet to the traveling mobile, the home agent must know the mobile's current temporary IP address (i.e., *care-of address*) on the foreign network [8]. This arrangement may results in longer routes and increased delays. In mobile IP for IPv6 [9], routing is improved in that a packet destined for the mobile can go directly to the foreign agent in the foreign network [10].

References

[1] Etemad, K., "Overview of Mobile WiMAX Technology and Evolution," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 31–40.

[2] Ahmadi, S., "An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Communications*, Vol. 47, No. 6, 2009, pp. 84–98.

[3] Eklund, C., et al., "IEEE Standard 802.16: A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access," *IEEE Communications*, Vol. 40, 2002, pp. 98–107.

[4] IEEE Standard 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York: IEEE, October 1, 2004.

[5] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[6] The Internet Engineering Task Force (IETF), "IP Mobility Support for IPv4," IETF RFC 3344, August 2002.

[7] The Internet Engineering Task Force (IETF), "Mobile IPv4 Challenge/Response Extensions (Revised)," IETF RFC 4721, January 2007.

[8] Yang, S. C., *3G CDMA2000 Wireless System Engineering*, Norwood, MA: Artech House, 2004.

[9] The Internet Engineering Task Force (IETF), "Mobility Support in IPv6," IETF RFC 3775, June 2004.

[10] Pontes, A. B., et al., "Handover Management in Integrated WLAN and Mobile WiMAX Networks," *IEEE Wireless Communications,* Vol. 15, No. 5, 2008, pp. 86–95.

CHAPTER 10

Quality of Service (QoS)

10.1 Introduction

Supporting QoS in broadband wireless systems is more challenging than in wired systems because the radio link can change as a function of time (time-selectivity), frequency (frequency-selectivity), and space (multiuser). To address these issues in broadband wireless systems, the MAC layer generally manages QoS directly [1]. This is because of MAC's proximity to the physical layer and its ability to quickly respond to changes on the radio link. Thus, a key function of the MAC layer is QoS. QoS provides a means for the effective allotment of limited bandwidth resources; through QoS, the system may provide users with different levels of service. The system can use QoS to deliver a rich set of services, each with its own requirements of data rate, priority, delay, and jitter. [2]

10.2 Definitions and Fundamental Concepts

10.2.1 Service Flows and QoS Parameters

Before going into details on the QoS capabilities of IEEE 802.16-based systems, we need to define some key terminologies. In particular, the concept of service flow is central to QoS. This is because QoS capabilities essentially map packets (entering the MAC layer) to their respective service flows. Formally, a *service flow* is a service provided by MAC that transports packets with a particular QoS in one direction only (uplink or downlink). Given this description, definitions of other terms follow:

- A *service flow ID* (SFID) is a unique 32-bit identifier assigned to a service flow.
- A *QoS parameter set* is a group of QoS parameters, such as traffic rate and latency, that characterizes a service flow.

The QoS parameter set can be examined in the context of three relevant types of service flows: active service flow, admitted service flow, and provisioned service

flow. An *active service flow* is one that is currently receiving resources from the base station for transporting packets. As such, the QoS parameter set that characterizes an active service flow is not null and is called the active QoS parameter set (i.e., ActiveQoSParamSet), and the base station has assigned resources for the parameters in the active QoS parameter set. An active service flow has both an SFID and an active CID assigned.

An *admitted service flow* is one that has requested (but not received) resources from the base station for transporting packets. It nevertheless has a CID assigned. The QoS parameter set that characterizes an admitted service flow is not null and is called the admitted QoS parameter set (i.e., AdmittedQoSParamSet); here the base station has reserved resources for the parameters in the admitted QoS parameter set. As such, an admitted service flow's active QoS parameter set is null, but an admitted service flow has both an SFID and a CID assigned.

A *provisioned service flow* is one that has an SFID assigned but has deferred admission and activation of resources by the base station. The QoS parameter set that characterizes a provisioned service flow is not null and is called the provisioned QoS parameter set (i.e., ProvisionedQoSParamSet); here, the parameters in the provisioned QoS parameter set may come from an external server and be based on those specified by a mobile's tiered subscription plan. A provisioned service flow's admitted QoS parameter set and active QoS parameter set are both null. It has an SFID but no CID.

The relationship among the three sets of QoS parameters (i.e., active QoS parameter set, admitted QoS parameter set, and provisioned QoS parameter set) can be examined through a Venn diagram, shown in Figure 10.1.[1] Figure 10.1 shows that the active QoS parameter set is a subset of the admitted QoS parameter set, and the admitted QoS parameter set is a subset of provisioned QoS parameter set. When one parameter set is a subset of another, it means that the parameters in the first set always require fewer or the same resources than as those in the second set. For example, the maximum sustained traffic rate in an active QoS parameter set is always less than or equal to that in the admitted QoS parameter set (of the same service flow, of course).

As we will see in the object relationship model later, a QoS parameter set is really an attribute of a service flow. The actual values of the QoS parameters (in the QoS parameter set) then govern the QoS behavior of a transport connection that is associated with the service flow.

10.2.2 Connections

The connection is another important concept. More specifically:

- A *connection* is a logical link between MAC peers in the base station and the mobile in one direction only (uplink or downlink). A connection that carries user traffic is called a transport connection; a connection that carries signaling is called a management connection.

1. For the provisioned authorization model.

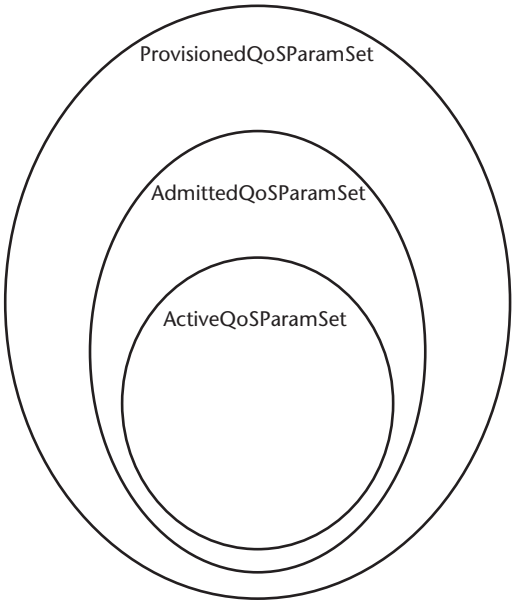


Figure 10.1 Relationships among different QoS parameter sets. (*After:* [3].)

- A *connection ID* (CID) is a unique 16-bit identifier assigned to a single transport connection (uplink or downlink). A CID can also identify a pair of management connections (uplink and downlink) for a mobile.

A transport connection is related to exactly one service flow, while a service flow is related to zero or one transport connection depending on the type of service flow. Thus, the set of QoS parameters that characterizes a service flow defines the transmission order and scheduling of that connection over the air interface. Because the air interface is typically the bottleneck, this connection-oriented QoS at layer 2 can enable end-to-end QoS [4].

10.3 Object Relationship Model

Having described the fundamental concepts above, we can now examine the relationships among these concepts. The object relationship model [5] uses diagrams to formally document objects (i.e., object classes), their attributes, and relationships among them [6]. Given that there are many different concepts in use, it would be helpful to have a diagram that depicts the relationships among them. Figure 10.2 shows the object relationship model.

Note the additional concepts depicted in the model [3]:

- A *service class* is an extra object used to represent a set of QoS parameters and their values. It is identified by the attribute (i.e., an ASCII string) "service class name."

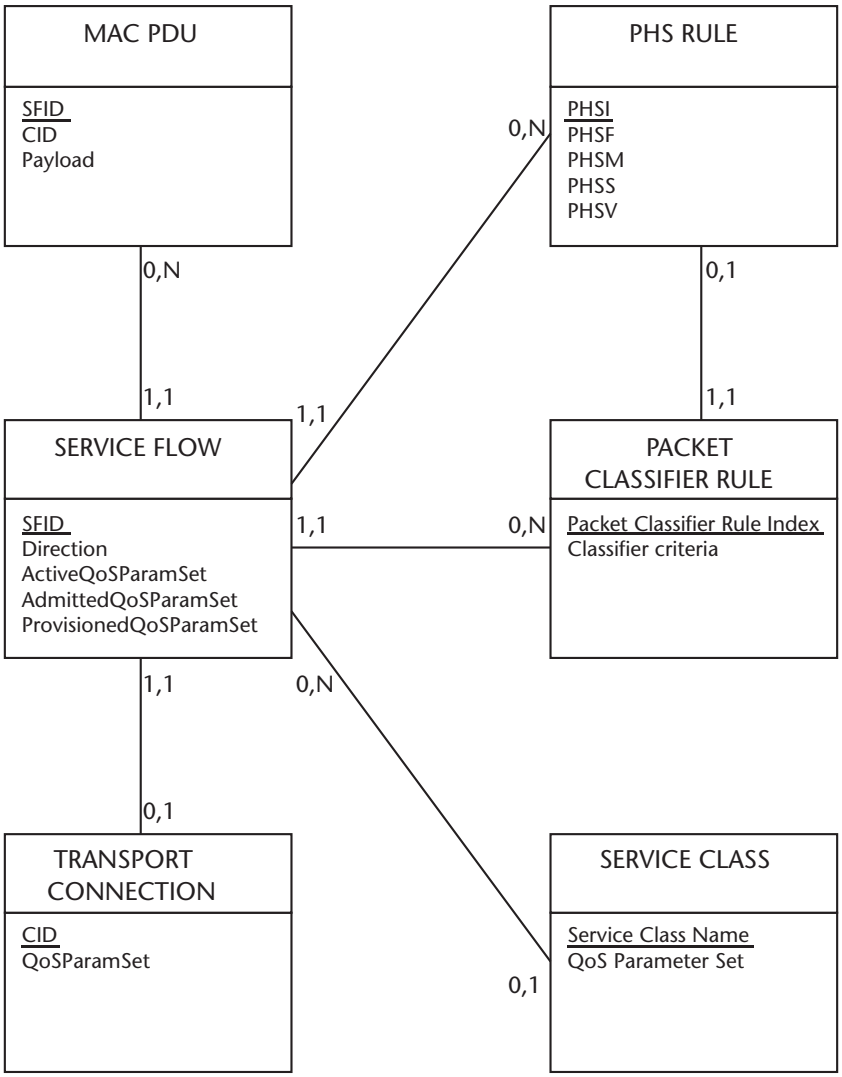


Figure 10.2 Object relationship model. (*After:* [3].)

- A *packet classifier rule* maps a packet to its service flow; a packet classifier rule also maps a packet to its PHS rule, if one exists. It is identified by the attribute "packet classifier rule index" that is 16-bit long.
- A *PHS rule* provides PHS-related parameters (e.g., PHSF, PHSM, and PHSI) and thus defines a suppressed header in a service flow. It is identified by the attribute "PHSI" (i.e., PHS index) that is 8-bits long. Each PHS rule is related to exactly one packet classifier rule and to exactly one service flow.

In the object relationship model, each rectangle represents an object. The name of the object is written near the top of the rectangle and is capitalized. The attributes are written below the name, and an attribute that identifies the object is underlined. The relationships are shown as lines connecting the objects, and the numbers (*m*,*n*) next to a line show, respectively, the minimum cardinality and maximum

cardinality of each object in the relationship. For reference, the relationships in the model are enumerated here:

- Each transport connection is related to exactly one service flow, while each service flow is related to zero or one transport connection.
- Each MAC PDU is related to exactly one service flow, while each service flow is related to zero or more MAC PDUs.
- Each service class is related to zero or more service flows, while each service flow is related to zero or one service class.
- Each packet classifier rule is related to exactly one service flow, while each service flow is related to zero or more packet classifier rules.
- Each PHS rule is related to exactly one service flow, while each service flow is related to zero or more PHS rules (because PHS is optional).
- Each PHS rule is related to exactly one packet classifier rule, while each packet classifier rule is related to zero or one PHS rule.

When a service flow is said to be provisioned, it means that an instance of the service flow object (or simply, a service flow) is created. When an instance of the service flow object is created, that instance is assigned an SFID, and its ProvisionedQoSParamSet attribute is populated with provisioned QoS parameters. Alternatively, the service flow can be related to a service class that would contain a QoS parameter set associated with the service flow.

10.4 Service Flow Transactions

10.4.1 Creating a Service Flow

If the base station initiates the creation of a service flow, the base station would first check if the service flow can be supported. If it can be, then the base station creates the service flow and the SFID. (It is the base station, not the mobile, that can generate an SFID.) The base station transmits a DSA-REQ message. This message would contain:

- An SFID for an uplink service flow or an SFID for a downlink service flow;
- AdmittedQoSParamSet or ActiveQoSParamSet.

The mobile accepts or rejects the request and responds with a DSA-RSP message. (For example, the mobile may reject because it cannot support a stated QoS parameter.) Afterwards, the base station acknowledges the receipt of the DSA-RSP message by sending a dynamic service addition acknowledgment (DSA-ACK) message. Figure 10.3 shows the exchange of messages.

Optionally, the mobile may initiate the creation of a service flow. If the mobile initiates, then the mobile transmits a DSA-REQ message. This message would contain: AdmittedQoSParamSet or ActiveQoSParamSet.

Note that, when the mobile initiates, the DSA-REQ message may not contain an SFID. This is because the base station may not have yet provisioned the

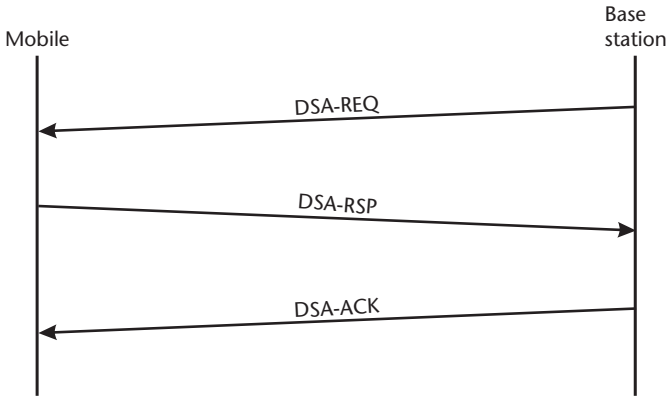


Figure 10.3 The exchange of messages when the base station initiates the creation.

requested service flow. Upon receiving the DSA-REQ message, the base station first authenticates the DSA-REQ message. If authentication is successful, then the base station sends a DSx Received (DSX-RVD) message to tell the mobile that it received the DSA-REQ message. The base station then checks if the service flow can be supported. If the base station accepts, it adds the service flow and creates the SFID. Then the base station sends a DSA-RSP message showing that it accepts (or rejects) the request. The last message of the exchange (shown in Figure 10.4) is the DSA-ACK message sent by the mobile acknowledging the receipt of the DSA-RSP message.

One important point: The exchange of messages shown in Figure 10.3 or Figure 10.4 involves only one single uplink service flow or one single downlink service flow. This is because the DSA-REQ message contains only one SFID.

10.4.2 Changing a Service Flow

The exchange of dynamic service change request (DSC-REQ) and dynamic service change response (DSC-RSP) messages is used to change a service flow. In essence, changing a service flow means updating the relevant QoS parameter sets.

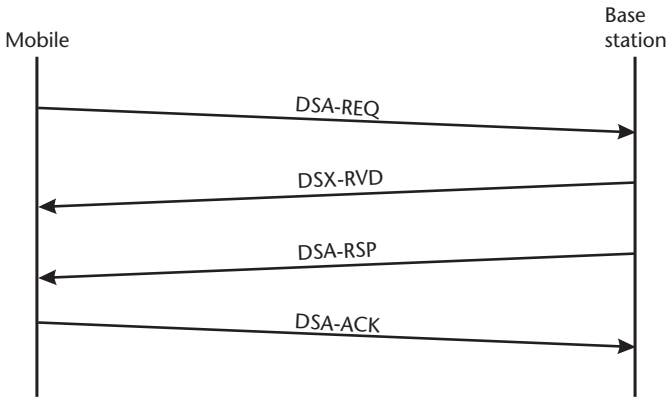


Figure 10.4 The exchange of messages when the mobile initiates the creation.

If the base station initiates the change, the base station transmits a DSC-REQ message. After receiving the DSC-REQ message, the mobile would check if the change to the service flow can be made. If it can be, the mobile changes the service flow; the mobile responds with a DSC-RSP message indicating that it accepts (or rejects). Upon receiving the DSC-RSP message, the base station changes the service flow at its end and sends a dynamic service change acknowledgment (DSC-ACK) message to acknowledge the DSC-RSP message. The exchange of messages is shown in Figure 10.5.

Optionally, the mobile may initiate the change of a service flow. The exchange of messages in this case is shown in Figure 10.6. If the mobile initiates the change, the mobile sends a DSC-REQ message. Upon receiving the DSC-REQ message, the base station first responds with a DSX-RVD message after authenticating the DSC-REQ message. The base station checks if the requested change to the service flow can be made. If it can be, the base station changes the service flow and sends a DSC-RSP message indicating that it accepts (or reject) the request. After receiving the DSC-RSP message, the mobile changes the service flow at its end and sends back a DSC-ACK to acknowledge.

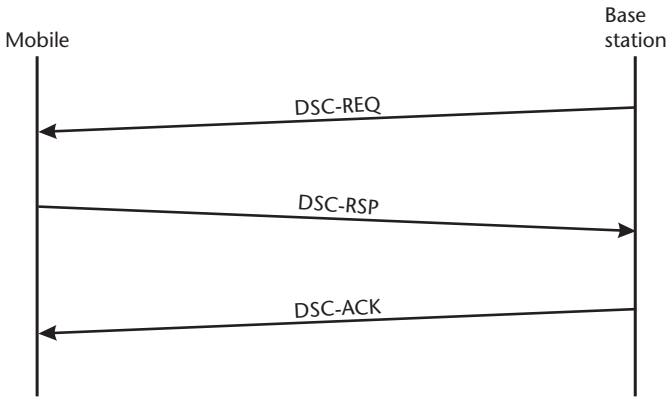


Figure 10.5 The exchange of messages when the base station initiates the change.

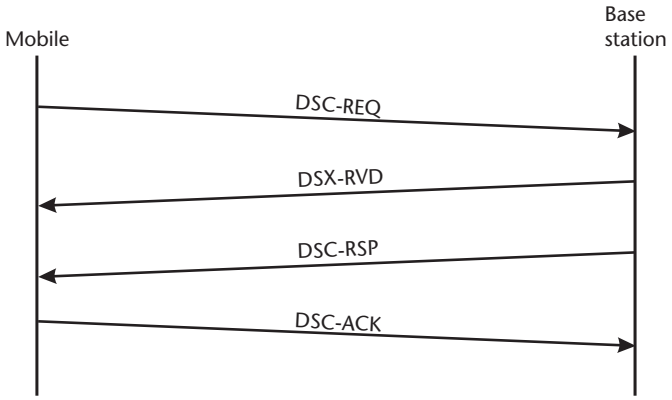


Figure 10.6 The exchange of messages when the mobile initiates the change.

In changing a service flow, eight scenarios are possible and are coded by the three functioning bits in the *QoS parameter set type* parameter, which is a parameter communicated by the message. Figure 10.7 shows the meaning of the bits in this parameter. Bit 0, when set (to 1), applies to the ProvisionedQoSParamSet; bit 1, when set, applies to the AdmittedQoSParamSet, and bit 2, when set, applies to the ActiveQoSParamSet. Table 10.1 separately shows the eight scenarios.

For reference, each of the eight scenarios is examined in more detail here:

- QoS parameter set type = "000": Not applying parameters to any set. At the same time, both the AdmittedQoSParamSet and the ActiveQoSParamSet are set to null. As a result, the service flow is both deadmitted and deactivated. In this case, the message contains neither the AdmittedQoSParamSet nor the ActiveQoSParamSet.
- QoS parameter set type = "001": Applying parameters to the ProvisionedQoSParamSet only.
- QoS parameter set type = "010": Applying parameters to the AdmittedQoSParamSet. This means that the system needs to perform admission control before applying the parameters. As a result, the service flow is de-

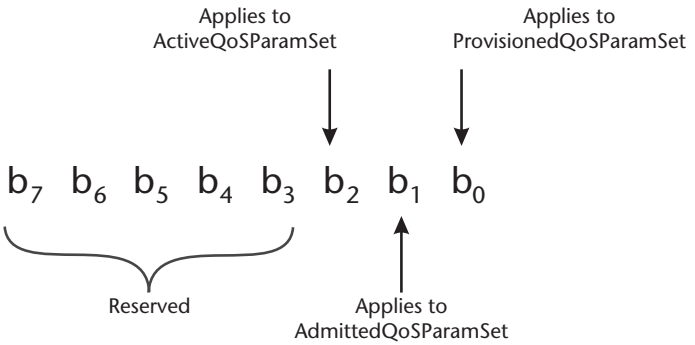


Figure 10.7 The QoS parameter set type parameter.

Table 10.1 Different Updates as Shown by QoS Parameter Set Type

Value	*Meaning*
000	Does not apply parameter to any set.
001	Apply parameters to ProvisionedQoSParamSet only.
010	Apply parameters to AdmittedQoSParamSet only.
100	Apply parameters to ActiveQoSParamSet only.
110	Apply parameters to both AdmittedQoSParamSet and ActiveQoSParamSet.
011	Apply parameters to both ProvisionedQoSParamSet and AdmittedQoSParamSet.
111	Apply parameters to ProvisionedQoSParamSet, AdmittedQoSParamSet, and ActiveQoSParamSet.
101	Apply parameters to ProvisionedQoSParamSet and ActiveQoSParamSet.

activated but remains admitted. In this case, the message contains only the AdmittedQoSParamSet, and the ActiveQoSParamSet is set to null.

- QoS parameter set type = "100": Applying parameters to the ActiveQoSParamSet. The system checks the ActiveQoSParamSet to see if it is a subset of AdmittedQoSParamSet or maybe even performs admission control if needed. If the check passes, then the service flow is activated. In this case, the message contains only the ActiveQoSParamSet.
- QoS parameter set type = "110": Applying parameters to both the AdmittedQoSParamSet and the ActiveQoSParamSet. This means that the system first performs admission control to check the AdmittedQoSParamSet. If the admission control checks out, the system then checks the ActiveQoSParamSet (to see if it is a subset of the AdmittedQoSParamSet). If the check passes, then both the AdmittedQoSParamSet and the ActiveQoSParamSet are applied. As a result, the service flow is activated (and by definition admitted). In this case, the message contains both the AdmittedQoSParamSet and the ActiveQoSParamSet. If either check fails, then both the old AdmittedQoSParamSet and the old ActiveQoSParamSet stand.
- QoS parameter set type = "011": Applying parameters to both the ProvisionedQoSParamSet and the AdmittedQoSParamSet. Thus, the system needs to perform admission control. If the check passes, the service flow is admitted. In this case, the message contains both the ProvisionedQoSParamSet and the AdmittedQoSParamSet.
- QoS parameter set type = "111": Applying parameters to the ProvisionedQoSParamSet, the AdmittedQoSParamSet, and the ActivatedQoSParamSet. Thus, the system performs admission control and checks the ActiveQoSParamSet. If the checks pass, the service flow is activated. In this case, the message contains the ProvisionedQoSParamSet, the AdmittedQoSParamSet, and the ActiveQoSParamSet.
- QoS parameter set type = "101": Applying parameters to the ProvisionedQoSParamSet and the ActiveQoSParamSet. The system here performs admission control and checks the ActiveQoSParamSet. If the checks pass, the service flow is activated. In this case, the message contains both the ProvisionedQoSParamSet and the ActiveQoSParamSet.

Similarly, the exchange of messages in Figure 10.5 (or Figure 10.6) involves only one single service flow. This is because the DSC-REQ message contains only one SFID.

10.4.3 Deleting a Service Flow

The exchange of dynamic service deletion request (DSD-REQ) and dynamic service deletion response (DSD-RSP) messages is used to delete a service flow. Upon deletion of a service flow, the system releases all resources dedicated to that service flow [3].

If the base station initiates the deletion, the base station first deletes the service flow and then transmits a DSD-REQ message. After receiving the DSD-REQ

message, the mobile deletes the service flow at its end and responds with a DSD-RSP message. There is no formal acknowledgment message to send. The exchange of messages is shown in Figure 10.8.

If the mobile initiates the deletion, the mobile first deletes the service flow and then sends a DSD-REQ message. After receiving the DSD-REQ message, the base station also deletes the service flow at its end and responds with a DSC-RSP message. Again there is no formal acknowledgment message. See Figure 10.9 for the exchange of messages in this case.

Because each DSD-REQ (and the corresponding DSD-RSP) message can only carry one SFID, a single exchange of these messages can only delete one single service flow. Note that either the base station or the mobile can just go ahead and delete the service flow first before sending the DSD-REQ message. No confirmation from the counterparty is required before deleting the service flow. This specification is adopted to quickly free up resources for other users in the system.

10.5 QoS Parameters

In this section, we examine some actual, salient QoS parameters used to provide scheduling services. To start, the following are the bit rate–related parameters:

- *Maximum sustained traffic rate* (bits per second) is the peak information rate of the service flow [3]. This parameter is concerned with just the SDUs entering the MAC layer (i.e., entering the convergence sublayer).
- *Minimum reserved traffic rate* (bits per second) is the minimum information rate promised to the service flow [3]. In other words, this parameter is the bandwidth reserved for the service flow, but the system only sticks to this

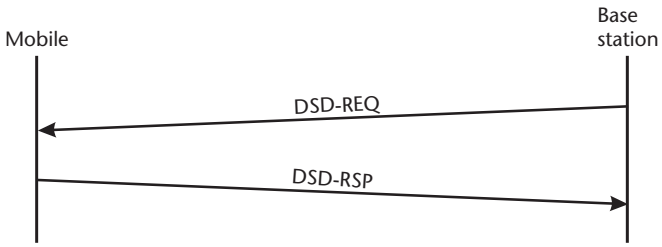


Figure 10.8 The exchange of messages when the base station initiates the deletion.

Figure 10.9 The exchange of messages when the mobile initiates the deletion.

minimum information rate when the service flow has enough traffic to send. Again, this parameter pertains to the SDUs entering the MAC layer.

The following are delay-related parameters [3]:

- *Maximum latency* (in milliseconds) is the maximum time elapsed or delay between when an SDU enters the MAC layer and when the SDU enters the air interface.
- *Tolerated jitter* (in milliseconds) is the maximum variation in delay.

The following are policy-related parameters:

- *Traffic priority* (from 0 to 7 with 7 having the highest priority) is the priority given to a service flow [3]. If two services have identical QoS parameters, preference is given to the service flow that has the higher priority. In practice, the higher-priority service flow is given a lower delay and a higher buffering preference [7]. Also, the base station uses this parameter to prioritize the request and grant of service flows (on the uplink).
- *Request/transmission policy* is a parameter that specifies certain policy-related attributes for a service flow [3]. This parameter includes 8 functioning bits, each of which specifies one attribute. Figure 10.10 depicts these bits and the associated attributes, and it is instructive to take a look at these policy-related attributes [3]:
  - Bit 0, when set (to 1), specifies that the service flow not use broadcast bandwidth request opportunities on the uplink.
  - Bit 1, when set, specifies that that the service flow not use multicast bandwidth request opportunities on the uplink.

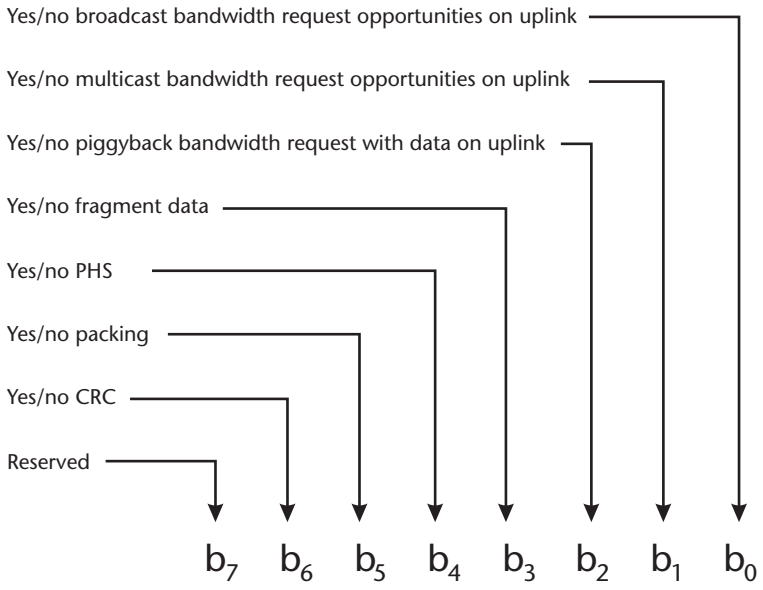


Figure 10.10 The request/transmission policy parameter.

  - Bit 2, when set, specifies that the service flow not piggyback bandwidth requests with data on the uplink.
  - Bit 3, when set, specifies that the service flow not fragment data.
  - Bit 4, when set, specifies that the service flow not utilize PHS.
  - Bit 5, when set, specifies that the service flow not utilize packing.
  - Bit 6, when set, specifies that the service flow not append CRC to the MAC PDU.
- *Unsolicited grant interval* (in milliseconds) is the desired time interval between consecutive bandwidth grant opportunities for the uplink service flow [3]. In reality, the actual time interval may be lengthened by the tolerated jitter.
- *Unsolicited polling interval* (in milliseconds) is the largest desired time interval between consecutive polling grant opportunities for the service flow [3].

10.6 Scheduling Services

Similar to the different QoS services provided by the asynchronous transfer mode (ATM)[2] at layer 2, IEEE 802.16-based systems at layer 2 can also provide different services, called scheduling services, with different QoS. The IEEE 802.16e standard explicitly calls out five uplink scheduling services enabled by different QoS parameters. These scheduling services are (in the order of generally decreasing demands on resources): unsolicited grant service (UGS), real-time polling service (rtPS), extended real-time polling service (ertPS), nonreal-time polling service (nrtPS), and best effort (BE). Table 10.2 gives a high-level summary of these scheduling services. A set of QoS parameters characterizes each of these scheduling services, and the set of QoS parameters (for a scheduling service) describes the guarantees required by the applications for which the corresponding scheduling service is designed [1].

Incidentally, the service-specific convergence sublayer (CS) depends on these scheduling services to operate. When the classifier in the CS classifies the packets, it selects the connection based on the type of QoS requirements associated with a scheduling service [8] (see Figure 7.2).

10.6.1 Unsolicited Grant Service (UGS)

The UGS supports uplink service flows that are real-time in nature and carry fixed-size SDUs. In essence, the UGS is used to emulate a circuit-like service. The UGS is

Table 10.2 Summary of Uplink Scheduling Services

	Real Time	*Nonreal Time*
Fixed-size packets	UGS (e.g., VoIP/no silence suppression)	BE
Variable-size packets	rtPS (e.g., MPEG) ertPS (e.g., VoIP/silence suppression)	nrtPS (e.g., FTP) BE (e.g., e-mail)

2. Also called ATM service categories.

called *unsolicited grant* because the base station periodically provides the mobile with data grant burst type IEs (which give the mobile an opportunity to transmit uplink PDUs). So the advantage is that the mobile does not have to expend overhead to request bandwidths itself and thus latency can be reduced. However, the disadvantage of the UGS is that the periodic grant for fixed-size SDUs is not efficient because there may be idle periods (e.g., conversational silence) in traffic from a higher-layer application. As such, the UGS is ideal for transporting T1 or E1 traffic and even VoIP with no silence suppression [3].

Obviously, the service flow's QoS parameter minimum reserved traffic rate determines: (1) how often the data grant burst type IEs are provided to the mobile, and (2) how large the bandwidth grants are. Because the UGS is emulating a circuit-like service, its minimum reserved traffic rate is the same as its maximum sustained traffic rate. In addition, because the base station is already providing data grant burst type IEs at periodic intervals, the QoS parameter request/transmission policy is set so that the mobile cannot send any contention-based bandwidth requests. This means that bit 0 and bit 1 of the request/transmission policy are set (i.e., the mobile cannot use broadcast or multicast bandwidth request opportunities on the uplink).

10.6.2 Real-Time Polling Service (rtPS)

The rtPS supports uplink service flows that are real-time in nature and carry variable-size SDUs. As such, the rtPS is ideal for transporting compressed video such as MPEG variable bit rate traffic [9]. The rtPS is called *real-time polling* because the base station periodically polls the mobile for the mobile's desired bandwidth grant, and the mobile can periodically communicate its bandwidth request (in a unicast fashion) to the base station. For real-time traffic, the advantage of the rtPS is that the periodic grant for variable-size SDUs can more efficiently respond to the needs of a higher-layer application while satisfying any real-time requirement, and rtPS supports variable grant sizes. The disadvantage is that the mobile does have to respond to the base station's polling and hence expend overhead to request bandwidths (through unicast polling). Also, rtPS has more latency than UGS.

The service flow's traffic rates determine how often the mobile is polled and how large the bandwidth grants are. Also, the maximum latency sets the upper bound on the waiting time experienced by a packet in the MAC layer [1]. In addition, because the base station is using unicast polling at periodic intervals, the QoS parameter request/transmission policy is set so that the mobile cannot send any contention-based bandwidth requests. This means that bit 0 and bit 1 of the request/transmission policy are set (i.e., the mobile cannot use broadcast or multicast bandwidth request opportunities on the uplink).

10.6.3 Extended Real-Time Polling Service (ertPS)

The ertPS also supports uplink service flows that are real-time in nature and carry variable-size SDUs, but it is slightly different from the rtPS. The ertPS is called *extended real-time polling* because it has a service level that is higher than rtPS but lower than UGS. The ertPS is really a balance between the UGS and rtPS and is characterized by the following:

- The base station periodically provides the mobile with bandwidth grants in an unsolicited manner (similar to the UGS) [8].
- The bandwidth grants are for variable-size SDUs (similar to rtPS).

In addition to receiving periodic grants from the base station, the mobile itself may also send bandwidth requests. In case the base station cannot provide unicast polling, the mobile here is allowed to send contention-based bandwidth requests. Moreover, the mobile may also request changes in the size of the uplink allocations.

The advantage of the ertPS is that it directly addresses the disadvantage of the UGS in the context of modern IP-based traffic. Recall that in the UGS, the periodic grant for fixed-size SDUs is not efficient because there may be idle periods in traffic from a higher-layer application. The ertPS's use of variable-size SDUs directly addresses this issue. As such, the ertPS's slightly lower QoS (as compared to UGS) enables the ertPS to transport real-time traffic that does not require a circuit-like emulation (e.g., VoIP traffic with silence suppression) [3].

Similarly, the service flow's traffic rates determine how often the bandwidth grants are given to the mobile and how large the bandwidth grants are. Because the mobile can send contention-based bandwidth requests, bit 0 and bit 1 of the request/transmission policy are not set (i.e., the mobile can use broadcast or multicast bandwidth request opportunities on the uplink).

10.6.4 Nonreal-Time Polling Service (nrtPS)

The nrtPS supports uplink service flows that are tolerant to delays and carry variable-size SDUs. The nrtPS is a nonreal-time service that requires a minimum reserved traffic rate. As such, the nrtPS is ideal for transporting delay-insensitive traffic like file transfer protocol (FTP) [3]. The nrtPS is called *nonreal-time polling* because the base station is committed to polling the mobile regularly (in a unicast fashion), but it does so at longer intervals. In addition, the mobile may also send contention-based bandwidth requests. For nonreal-time traffic, the advantage of the nrtPS is that the base station is committed to some level of regular unicast polling, while the mobile also has the flexibility of sending contention-based bandwidth requests.

The service flow's traffic rates determine how often the mobile is polled and how large the bandwidth grants are. Also, because the mobile is allowed to send contention-based bandwidth requests (in addition to unicast polling), the QoS parameter request/transmission policy is set so that the mobile can send contention-based bandwidth requests. This means that bit 0 and bit 1 of the request/transmission policy are not set.

10.6.5 Best Effort (BE)

The BE supports uplink service flows that are very tolerant to delays. It differs from the nrtPS in that the BE has a lower QoS. As such, the BE can be used to transport delay-insensitive, noncritical, background traffic. E-mail traffic can make use of the BE as well [9]. Nominally, the BE traffic is transported when there is bandwidth available. With the BE, the mobile may use both contention-based bandwidth

requests and unicast polling [3]. For nonreal-time traffic, the advantage of the BE is that it is highly efficient, consuming bandwidth when available while satisfying the low-QoS requirements. Admittedly, the BE may provide a low traffic rate.

The service flow's QoS parameter maximum sustained traffic rate determines how often the mobile is polled and how large the bandwidth grants are. By definition, the BE has no minimum service level. (That is why there is no mandatory minimum reserved traffic rate.) Also, because the mobile is allowed to send contention-based bandwidth requests (in addition to unicast polling), the QoS parameter request/transmission policy is set so that the mobile can send contention-based bandwidth requests (i.e., bit 0 and bit 1 of request/transmission policy are not set).

10.6.6 Remarks

Table 10.3 shows the five scheduling services and the QoS parameters that are important to each scheduling service. In terms of traffic rates, all scheduling services obviously have a specified maximum traffic rate (i.e., maximum sustained traffic rate) that they cannot exceed. For all scheduling services except the BE, a minimum traffic rate (i.e., minimum reserved traffic rate) is also specified to guarantee some minimum service level. The BE, by definition, does not have a minimum traffic rate.

In terms of delay, all three real-time scheduling services (i.e., UGS, ertPS, and rtPS) specify a maximum latency to satisfy the real-time requirement. Because the

Table 10.3 Different Uplink Scheduling Services and Their Salient QoS Parameters

	Unsolicited Grant Service (UGS)	*Real-Time Polling Service (rtPS)*	*Extended Real-Time Polling Service (ertPS)*	*Nonreal-Time Polling Service (nrtPS)*	*Best Effort (BE)*
Maximum sustained traffic rate	√	√	√	√	√
Minimum reserved traffic rate	√	√	√	√	—
Maximum latency	√	√	√	—	—
Tolerated jitter	√	—	—	—	—
Request/ transmission policy	√	√	√	√	√
Traffic priority	—	—	—	√	—
Unsolicited grant interval (for UL)	√	—	—	—	—
Unsolicited polling interval (for UL)	—	√	—	—	—

UGS is emulating a circuit-like service, the UGS has a mandatory jitter (i.e., tolerated jitter) requirement.

In terms of policy, a request/transmission policy is needed for all scheduling services. For the nrtPS (a nonreal-time scheduling service), a traffic priority is needed in case there are competing nonreal-time service flows vying for resources. In addition, the unsolicited grant interval is specified for the UGS because a key feature of the UGS is how often a data grant is given. Likewise, the unsolicited polling interval is specified for the rtPS because unicast polling is a key feature defining the rtPS.

References

[1] Cicconetti, C., et al., "Quality of Service Support in IEEE 802.16 Networks," *IEEE Network,* Vol. 20, No. 2, 2006, pp. 50–55.

[2] Laroia, R., S. Uppala, and J. Li, "Designing a Mobile Broadband Wireless Access Network," *IEEE Signal Processing,* Vol. 21, No. 5, 2004, pp. 20–28.

[3] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[4] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

[5] Martin, J., and J. Odell, *Object-Oriented Analysis and Design*, Englewood Cliffs, NJ: Prentice-Hall, 1992.

[6] Dewitz, S. D., *Systems Analysis and Design and the Transition to Objects*, New York: McGraw-Hill, 1996.

[7] Nair, G., et al., "IEEE 802.16 Medium Access Control and Service Provisioning," *Intel Technology Journal*, Vol. 8, No. 3, 2004, pp. 213–228.

[8] Huang, C. Y., et al., "Radio Resource Management of Heterogeneous Services in Mobile WiMAX Systems," *IEEE Wireless Communications,* Vol. 14, No. 1, 2007, pp. 20–26.

[9] Pontes, A. B., et al., "Handover Management in Integrated WLAN and Mobile WiMAX Networks," *IEEE Wireless Communications,* Vol. 15, No. 5, 2008, pp. 86–95.

CHAPTER 11

Security Fundamentals

11.1 Introduction

Before examining the functions performed by the security sublayer (in IEEE 802.16), we need to consider first the essentials of information security. This chapter looks at those aspects of security important to understanding the functions of the security sublayer. As such, those readers who are familiar with the fundamentals may wish to proceed directly to the next chapter.

In information security, the system seeks to meet three fundamental objectives [1, 2]:

- *Confidentiality:* "Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information…A loss of confidentiality is the unauthorized disclosure of information" [2]. In the context of security functions performed by a layer 2 protocol (such as the security sublayer in IEEE 802.16-based systems), confidentiality means ensuring that data remain private and are only disclosed to an authorized entity.
- *Integrity:* "Guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity…A loss of integrity is the unauthorized modification or destruction of information" [2]. In terms of security functions of layer 2, integrity means ensuring that data are not modified by an unauthorized entity, as well as ensuring non-repudiation and authenticity.
- *Availability:* "Ensuring timely and reliable access to and use of information… A loss of availability is the disruption of access to or use of information or an information system" [2]. In other words, availability means ensuring that data or system can be accessed in a timely manner.

These three objectives are also known as the *CIA triad* (see Figure 11.1), which includes the foundational concepts and goals of information security. When people refer to information security, they often mean one or more of these three concepts.

For a to-the-point treatment, this chapter examines only those topics of information security relevant to the operation of the security sublayer (in IEEE 802.16).

Figure 11.1 The CIA triad.

In particular, we focus on those fundamental concepts related to the first two objectives of information security: confidentiality and integrity, which relate more to the specific security functions performed by layer 2. More details on the functions of the security sublayer are found in Chapter 12.

11.2 Symmetric Encryption

Encryption is an essential operation in meeting the objectives of information security, including confidentiality and integrity. Symmetric encryption, in particular, helps meet the goal of confidentiality; it does so by encrypting messages to be transmitted over the air (in the case of a wireless network) and rendering them unintelligible to an eavesdropper. Figure 11.2 shows the symmetric encryption/decryption process.

In Figure 11.2, Bob is the sender, and Bob wishes to send a confidential message to Alice. Bob encrypts the *plaintext P* (i.e., the original message) by using the *encryption algorithm E* with an *encryption key K*. The encryption algorithm produces the *ciphertext C* that is transmitted over the network (where it is vulnerable to intercept). At the receiver, Alice decrypts the received ciphertext by using the *decryption algorithm D* with the *decryption key K*. If the decryption key is the correct key, then the decryption algorithm would produce the original plaintext that was sent. In symmetric encryption, *the encryption key K is the same as the decryption key K*.

This process ensures the confidentiality of the message because if an eavesdropper on the network intercepts the ciphertext, the eavesdropper cannot read the ciphertext because it is encrypted and unintelligible. The eavesdropper would not be able to read the ciphertext without the decryption key (and the knowledge of the encryption algorithm used). In general, information security does not depend on keeping the type of encryption algorithm secret. It does, however, very much depend on keeping the decryption key secret.

The advantage of symmetric algorithms is that they are generally fast and can be easily implemented in hardware. However, there are two issues with symmetric encryption/decryption:

- *Key distribution:* Because the encryption key is the same as the decryption key, care must be taken to communicate the decryption key over the network (where the key is subject to intercept) to the receiver, so the receiver can use the same key to decrypt the ciphertext. If the key itself is compromised, then

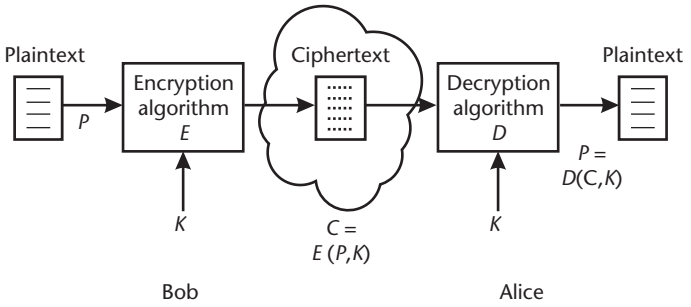


Figure 11.2 Symmetric encryption and decryption for message confidentiality.

confidentiality can no longer be ensured. Often times, the sender encrypts the key itself before sending the (encrypted) key over the network to the receiver.

- *Key management:* Because the encryption key is the same as the decryption key, receiving confidential messages from N parties requires N keys.

Examples of symmetric encryption algorithms include data encryption standard (DES), triple data encryption standard (3DES), and advanced encryption standard (AES).

11.3 Asymmetric Encryption

Asymmetric encryption also helps to meet the goal of confidentiality by encrypting a message to be transmitted. Figure 11.3 shows the asymmetric encryption/decryption[1] process, by which Bob is again sending a confidential message to Alice. As shown in Figure 11.3, the process is similar to that of symmetric encryption, except that the *encryption key is different from the decryption key.*

As the sender, Bob encrypts the plaintext P by using the encryption algorithm E with Alice's public encryption key $K_{E,public,Alice}$ (i.e., Alice's "public key"). The encryption algorithm produces the ciphertext C which is transmitted over the network. At the receiver, Alice decrypts the received ciphertext by using the decryption

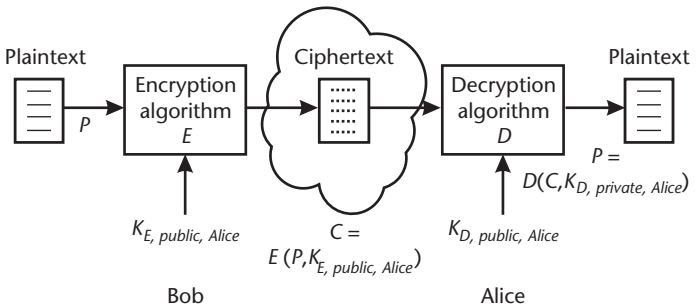


Figure 11.3 Asymmetric encryption and decryption for message confidentiality.

1. Asymmetric encryption is also known as public key encryption because of the use of the public key in addition to the private key.

algorithm D with her own private decryption key $K_{D,private,Alice}$ (i.e., Alice's "private key"). If the decryption key is the correct key, then the decryption algorithm would produce the original plaintext that was sent.

In effect, Alice keeps her encryption key public (by posting it on her Web site or by asking a certificate authority to distribute it for her). Alice is in fact saying to the whole world "if you want to send me a confidential message, use this (public) encryption key to encrypt the message to be sent to me." On the other hand, Alice keeps her decryption key private and does not let anyone know what it is. If the decryption key is kept private, then the asymmetric encryption/decryption process described above can keep a message confidential while in transit.

The advantages of asymmetric algorithms are twofold:

- *Key distribution:* Because the encryption key is different from the decryption key, the decryption key can reside with the receiver and stay unexposed.
- *Key management:* Because senders only need the receiver's public encryption key to protect their confidential messages, only one encryption key is needed to send messages. For a receiver to receive confidential messages from N parties, only one (public) encryption key and one (private) decryption key are needed. Requiring only two keys makes the task of key management much simpler.

An example of asymmetric encryption algorithm is the RSA.[2]

11.4 Digital Signature

It turns out that the asymmetric encryption/decryption process described in Section 11.3 can be used to implementing *digital signature*, which helps to meet the goal of integrity including message integrity. Figure 11.4 shows how asymmetric algorithms can implement digital signature.

In Figure 11.4, Bob encrypts the plaintext P by using the encryption algorithm E with Bob's own private encryption key $K_{E,private,Bob}$ (i.e., Bob's "private key"). The encryption algorithm produces the ciphertext C which is transmitted over the

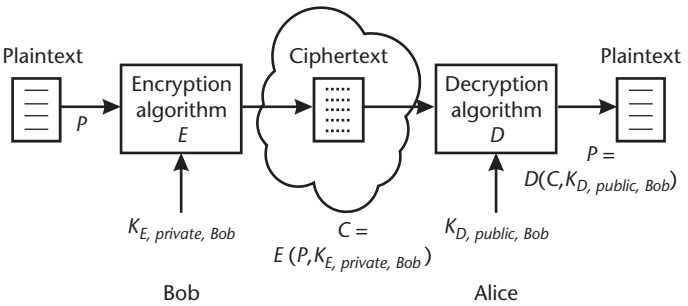


Figure 11.4 Asymm etric encryption and decryption for signing messages.

2. RSA is a public key encryption algorithm named after its three inventors: Ronald Rivest, Adi Shamir, and Leonard Adleman.

network. At the receiver, Alice decrypts the received ciphertext by using the decryption algorithm D with Bob's public decryption key $K_{D,public,Bob}$ (i.e., Bob's "public key"). The decryption algorithm would produce the original plaintext that was sent. More importantly, if the decryption algorithm can use Bob's public decryption key to successfully produce the original plaintext, then Alice can be certain that the message is really from Bob. That certainty comes about because Bob is the only one in the world who has his own private (encryption) key. Therefore, Bob has effectively "signed" the message by encrypting it with his own private encryption key.

Note that the process just described does not ensure confidentiality. Because the message can be decrypted by Bob's public decryption key, everyone in the world can read the message (because everyone in the world by definition has Bob's public decryption key). However, the process does ensure message integrity through the signing of the message. In addition, the process ensures nonrepudiation because the process mathematically proves that it was Bob who sent the message, and Bob cannot later repudiate the fact that the message came from him.

11.5 Message Authentication Using Message Authentication Code

In practice, one often does not authenticate a message by encrypting the entire message. The reason is that the encryption/decryption process is computationally intensive (especially the decryption process). In addition, if the message is long (e.g., a portable computer program), then it would take a long time to authenticate the message. The solution is to use the original message to generate a shorter "digest" and authenticate the digest, not the message [3]. One popular way to generating the digest is through the message authentication code (MAC) scheme, which is shown in Figure 11.5. The MAC scheme helps to meet the goal of integrity (i.e., message integrity).

As shown in the figure, the sender uses a MAC algorithm to produce the MAC using the following:

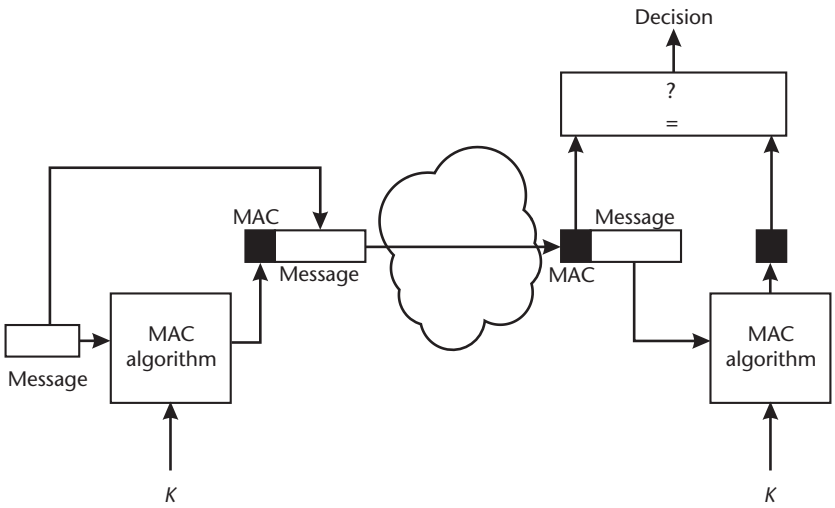


Figure 11.5 The MAC process.

- The message to be sent;
- A key.

The sender sends both the message and the MAC to the receiver. Then the receiver uses the same algorithm to calculate the MAC using:

- The received message;
- The same key.

The receiver compares the received (attached) MAC with the receiver-calculated MAC. If they are the same, then the message has not been tampered in transit. This is because an attacker cannot alter the transmitted message and the attached MAC without the key, and no one else (besides the sender and the receiver) has the key [3]. On the other hand, if the attached MAC is different from the receiver-calculated MAC, then the receiver can no longer be sure that the message is authentic. This process depends on the MAC algorithm, which requires a key to produce the MAC. Of course, the reliability of the MAC scheme depends on both the sender and the receiver having the same key and keeping it secret. Thus, the key is sometimes called the *shared secret key*.

There are two methods by which the MAC algorithm can produce the MAC: hash-based and cipher-based.

11.5.1 Hash Based

The hash-based MAC (HMAC) algorithm incorporates both a hash algorithm and a key. A hash function (or more specifically, a one-way hash function) produces a fixed-size digest as its output by using a variable-size message as its input; the fixed-size digest is meant to be a "fingerprint" of the message [1]. As such, it should be pretty difficult to find two different messages that have identical hashed digest—a property called *collision resistance*. A hash function typically does not require a key to produce a digest, but a MAC algorithm does require a key to produce a MAC. In particular, the HMAC algorithm is designed so that it not only makes use of a hash function but also requires a key to generate a MAC. Figure 11.6 shows a high-level description of the HMAC algorithm.

Mathematically, the HMAC algorithm can be summarized as:

$$HMAC(K,M) = H\left\{\left[K \oplus p_A\right] \| H\left[\left(K \oplus p_B\right) \| M\right]\right\} \tag{11.1}$$

where $\oplus$ is XOR and $\|$ is concatenation. p_A and p_B are padding sequences A and B, respectively.

One advantage of HMAC is that it can make use of a hash function already available (e.g., in the public domain). If a new, more secure hash function becomes available, then HMAC can easily incorporate that new hash function because HMAC treats the hash function as a black box.

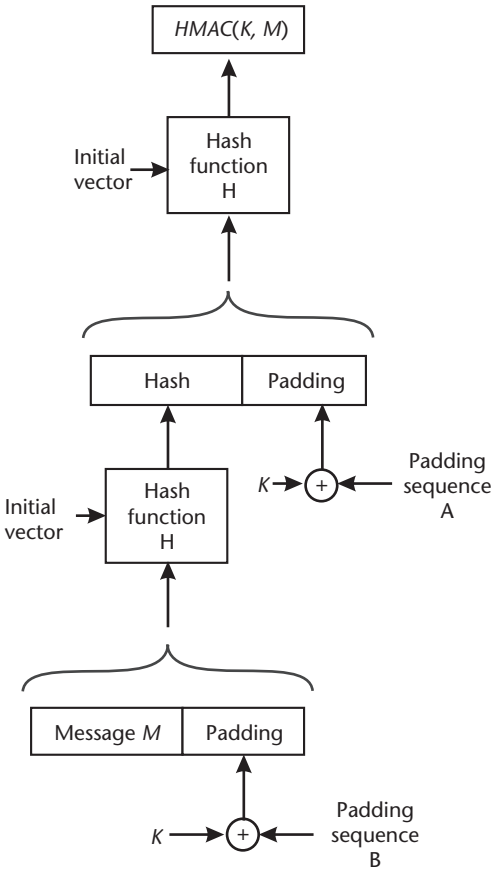


Figure 11.6 High-level description of the HMAC algorithm.

11.5.2 Cipher Based

Instead of using a hash function, a cipher-based MAC (CMAC) algorithm uses a cipher to generate the MAC. Because a cipher itself already requires a key, the resulting MAC algorithm also requires a key to generate the MAC. For the cipher, a CMAC algorithm can use the cipher block chaining (CBC) block cipher,[3] for example. In the CBC block cipher, the output ciphertext is fed back into the encryption algorithm as the initial vector, and the CMAC uses the CBC block cipher (and the corresponding key) to generate the MAC. The IEEE 802.16e standard specifies its CMAC using the AES block cipher [4].

11.6 Conclusions

This chapter examines those fundamental concepts related to the first two objectives of information security: confidentiality and integrity, which are goals to be met by the security functions of layer 2. Encryption (both symmetric and asymmetric) can help to ensure confidentiality, and digital signature and MAC can help ensure

3. A *block cipher* generates one block of ciphertext output for every block of plaintext input. On the other hand, a *stream cipher* generates ciphertext output continuously.

integrity. With this background, we are now ready to discuss, in Chapter 12, the specific security functions performed by layer 2 of IEEE 802.16.

References

[1] Stallings, W., and L. Brown, *Computer Security: Principles and Practice*, Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

[2] National Institute of Standards and Technology, *Standards for Security Categorization of Federal Information and Information Systems*, Federal Information Processing Standards (FIPS) Publication 199, February 2004, p. 2.

[3] Stallings, W., *Business Data Communications*, Upper Saddle River, NJ: Prentice-Hall, 2009.

[4] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

CHAPTER 12

Security Functions

12.1 Introduction

In information security, there are three fundamental security objectives: confidentiality, integrity, and availability [1]. Confidentiality means ensuring that data remain private and are only disclosed to an authorized entity. Integrity means ensuring that data are not modified by an unauthorized entity, as well as ensuring nonrepudiation and authenticity. Availability means ensuring that data or system can be accessed in a timely manner.

In IEEE 802.16-based systems, the security sublayer provides functions that facilitate meeting the first two objectives: confidentiality and integrity. The third objective, availability, is met by the entire system (not just the security sublayer). In particular, the following are performed to meet the security objective of *integrity:*

- The system can authorize the mobile to use its services. This way, only an authorized mobile (e.g., a subscriber) may use the system.
- The system "signs" certain messages by attaching a digest to the message. The digest would show if the message has been modified en route by an attacker. This process is also known as message authentication.

To meet the security objective of *confidentiality,* the system can encrypt messages sent over the air, thus rendering them unintelligible to an eavesdropper.

Figure 12.1 shows the above functions at a high level. First, the base station authenticates the mobile and makes sure that it really is who it says it is (e.g., a subscriber), and then the mobile is authorized to use the services provided by the system. Note that the mobile may also authenticate the base station to make sure that the base station is a legitimate one and not an "evil twin." Authorization, if performed, can be done using RSA-based authorization, extensible authentication protocol (EAP)-based authorization, or a sequence of these two authorizations.

Second, based on the results of the authorization procedure, the system distributes key materials that are needed to sign messages and to encrypt messages. The distribution of key materials (to both the mobile and the base station) is performed and managed by the privacy key management (PKM) protocol. For example, the

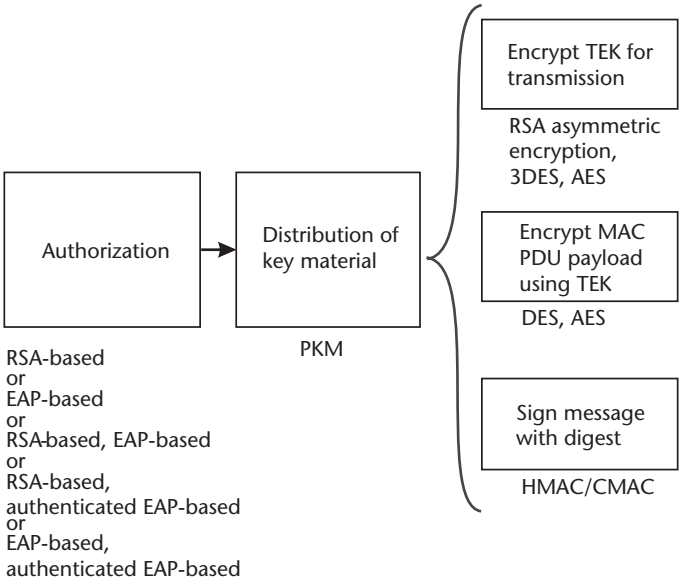


Figure 12.1 High-level functions of the security sublayer.

base station sends the traffic encryption key (TEK) to the mobile in a confidential manner by encrypting the TEK.

The distribution of key materials enables the mobile and the base station to exchange messages securely. This means that: (1) a message can be encrypted using the TEK so the message cannot be intercepted and read (message encryption), and (2) a message can be signed with a digest so the receiver can be sure of its originator (message authentication). Note that only the payload of the MAC PDU can be encrypted. In general, overhead information is not encrypted, so the generic MAC header is sent in the clear (not encrypted), and all MAC management messages are sent in the clear. However, all MAC management messages should be signed to ensure their integrity [2].[1]

This chapter emphasizes on the aspects of authorization and the distribution of key materials (shown in Figure 12.1). The basics of symmetric and asymmetric encryptions, which can be used to protect TEK and to protect the MAC PDU payload, are covered in Chapter 11; hash/cipher-based message authentication codes in general, which can be used to sign messages, are also covered in Chapter 11. This chapter focuses on security functions associated with unicast connections.

12.2 Definitions and Fundamental Concepts

Before going into details on the functions shown in Figure 12.1, we need to go over first some important definitions. A *cryptographic suite* is a set of encryption/decryption algorithms used for three functions: encrypting the TEK to transmit the TEK privately to the mobile, encrypting the payload of the MAC PDU, and signing the

1. The actual supported message authentication code (MAC) is shown by the MAC mode field in the SBC-REQ/SBC-RSP messages.

message. In other words, a cryptographic suite defines specific algorithms used to carry out these three functions (also shown on the right side of Figure 12.1). Table 12.1 shows the different cryptographic suites used as specified by IEEE 802.16e. For each cryptographic suite, Table 12.1 also shows the actual encryption/decryption algorithms used to carry out the three functions. Note that as the standard evolves, more cryptographic suites may be added to the supported list.

Each cryptographic suite used is designated by a *cryptographic suite value*. The cryptographic suite value is a concatenation of 3 bytes. The most significant byte (MSByte) indexes the algorithm used to encrypt the MAC PDU payload; the middle byte (MidByte) indexes the algorithm used to sign messages; and the least significant byte (LSByte) indexes the algorithm used to encrypt the TEK. For example, MSByte = 2 (in decimal) means that the system is using 128-bit AES with CCM for encrypting the MAC PDU payload; MidByte = 1 (in decimal) means that the system is using 128-bit AES with CCM for signing messages, and LSByte = 3 (in decimal) means that the system is using 128-bit AES with ECB for encrypting the TEK. For reference, the appropriate standards behind the actual algorithms used are also cited in Table 12.1.

A *security association* (SA) is a collection of information shared between the base station and the mobile for the purpose of carrying out the functions of the security sublayer. For example, an SA defines the cryptographic suite used for that particular SA and holds the key materials for that suite. Each SA is uniquely identified by a 16-bit *security association identifier* (SAID). As far as the relationship between connections and SAs is concerned, a transport connection is associated with an SA [2].

An *authorization key* (AK) is a secret value shared between the mobile and the base station [2]. The AK is important for two reasons. First, the generation and

Table 12.1 Cryptographic Suites

Cryptographic Suite Value					
MSByte	*MidByte*	*LSByte*	*Encrypting MAC PDU Payload*	*Signing Message*	*Encrypting TEK*
0	0	1	None	None	128-bit 3DES (EDE)
1	0	1	56-bit DES (CBC) [3, 4]	None	128-bit 3DES (EDE)
0	0	2	None	None	1,024-bit RSA [5]
1	0	2	56-bit DES (CBC)	None	1,024-bit RSA
2	1	3	128-bit AES (CCM) [6–8]	128-bit AES (CCM)	128-bit AES (ECB) [6, 9]
2	1	4	128-bit AES (CCM)	128-bit AES (CCM)	128-bit AES key wrap
3	0	3	128-bit AES (CBC) [6, 9]	None	128-bit AES (ECB)
128	0	3	128-bit AES (CTR for MBS) [6, 9]	None	128-bit AES (ECB)
128	0	4	128-bit AES (CTR for MBS)	None	128-bit AES key wrap

EDE: encrypt, decrypt, encrypt; CBC: cipher block chaining; CCM: countermode with cipher block chaining-message authentication code; CTR: countermode; MBS: multicast broadcast service; ECB: electronic code book.

the distribution of the AK to the mobile and the base station mean that the mobile has been authorized to use the service. In fact, the mobile's request for an AK is part of the authorization process. Second, the AK serves as the basis from which other important keys are derived, including key encryption key (KEK) and message authentication code (MAC) keys.

A *key encryption key* (KEK) is a key used to encrypt the *traffic encryption key* (TEK). To prevent the TEK from being intercepted and read over the air, the system uses the KEK to encrypt the TEK. A KEK is the key that is associated with a specific algorithm in the sixth column of Table 12.1. The algorithms used to encrypt the TEK can be either symmetric encryption algorithms (i.e., 3DES and AES) or an asymmetric encryption algorithm (i.e., RSA).

The message authentication code (MAC) keys are used to sign certain messages and implement message authentication. In effect, the system uses the MAC key to generate a digest based on the message to be signed, and that digest is attached to the message. If the message has been altered en route by an attacker, the message is no longer authentic, and such an alteration can be detected because the attached digest would differ from the receiver-calculated digest (calculated based on the received message). There are two types of MAC keys: hashed message authentication code (HMAC) [10, 11] keys and cipher-based message authentication code (CMAC) [12] keys. An HMAC key is used to generate the HMAC digest, and a CMAC key is used to generate the CMAC digest.[2]

A *traffic encryption key* (TEK) is ultimately what the mobile is trying to obtain. The TEK is used to encrypt the payloads of MAC PDUs. A TEK is the key associated with a specific algorithm in the fourth column of Table 12.1. Note that all the algorithms used to encrypt MAC PDUs are symmetric encryption algorithms (i.e., DES and AES). This is because one of the advantages of symmetric encryption is its fast processing speed, and fast speed imposes less delay.

12.3 Authorization

Authorization normally follows the negotiation of mobile capabilities in the network entry process (see Chapter 9). Here, the base station:

- Authenticates a mobile;
- Matches the authenticated mobile to a subscriber in good standing;
- Authorizes that mobile to use the services.

Out of these, the first step is very important for it involves ensuring that the mobile requesting service is really who it says it is and not someone who pretends to be such.

There are basically two constituent schemes by which the base station authorizes a mobile: RSA-based authorization and EAP-based authorization. Depending on the authorization policy supported, the system can authorize using one of the

2. For CMAC, the CMAC algorithm generates a CMAC "value," which becomes part of the CMAC digest.

two schemes, or it can authorize using two schemes in sequence. For example, the system can first perform RSA-based authorization and then perform EAP-based authorization. All in all, there are a total of six authorization policies specified [2]:

- No authorization;
- Only RSA-based authorization;
- Only EAP-based authorization;
- RSA-based authorization followed by EAP-based authorization;
- RSA-based authorization followed by authenticated EAP-based authorization;[3]
- EAP-based authorization followed by authenticated EAP-based authorization.

The specific authorization policy to be used is first negotiated between the base station and the mobile using the SBC-REQ and SBC-RSP messages. Authorization, if used, defends against *masquerade* attacks where an attacker pretends to be a legitimate user. Note that if both the base station and the mobile negotiate their authorization policy to be "no authorization," then the system will behave like an open system—no authorization will be performed (RSA-based or EAP-based), no SA will be assigned, no keys will be distributed (AK or TEK), and the SAID will be null [2].

To keep the discussion concise, we focus on the authentication aspect of the two constituent schemes: RSA-based authorization and EAP-based authorization. Additionally, although both PKM version 1 (PKMv1) and PKM version 2 (PKMv2) are specified by the standard for generating and distributing key materials, we explore the more recent PKMv2 in the subsequent discussion.

12.3.1 RSA Based

In the basic RSA-based technique, an entity presents a credential as proof of its identity, and that credential is the X.509 certificate. An X.509 certificate can be thought of as a small file that has several fields containing security-related information. X.509 is a standard [13] that specifies the formatting of those fields. Out of those fields, there are two that are the most important: the public key of the entity possessing the X.509 certificate and the ID of the entity possessing the X.509 certificate. In the context of an 802.16-based system, these two fields in a mobile's X.509 certificate are the mobile's public key and the mobile's medium access control (MAC) address.

Because the mobile's MAC address is readily obtainable (often by looking at the manufacturer's tag on the device) and the mobile's public key is by definition public, anyone could generate a small file and populate the fields with the mobile's public key, MAC address, and other security-related information. To put it in another way, anyone could forge a certificate and "clone" the mobile. So to prevent forgery of the X.509 certificate (a type of masquerade attacks), an X.509 certificate

3. Authenticated EAP-based authorization is a special type of EAP-based authorization. In it, an EAP message is signed using the HMAC/CMAC key generated from the EAP integrity key (EIK) derived in a previous authorization.

has to be digitally signed by the mobile device's manufacturer; this means that the X.509 certificate is encrypted using the device manufacturer's private key. Often, a trusted third party called a *certificate authority* (CA) issues these X.509 certificates in bulk to a device manufacturer; the device manufacturer uses its own private key to encrypt the X.509 certificates and incorporate them in the manufactured devices. Note that on the World Wide Web, an X.509 certificate is signed by the CA.

To acquire an AK (which is the goal of the authorization process), the mobile first needs to be authenticated. To be authenticated, the mobile transmits its X.509 certificate (encrypted by the manufacturer's private key) to the base station. The mobile sends this (encrypted) X.509 certificate using the PKMv2 RSA-Request message. After receiving the PKMv2 RSA-Request message, the base station decrypts the X.509 certificate using the manufacturer's public key (which is by definition public and readily obtainable). If the base station can indeed decrypt the X.509 certificate and read its fields, then it can be sure that the X.509 certificate presented is authentic; the certificate is proven to be authentic because the manufacturer, presumably a trusted party, is the only one in the world that has its own private key.

Once the base station is convinced that the certificate is authentic, it goes ahead and reads the mobile's public key from one of the fields in the certificate. Then the base station uses the mobile's public key to encrypt a key called the preprimary AK (pre-PAK) and sends it back to the mobile. The base station sends the (encrypted) pre-PAK using the PKMv2 RSA-Reply message. After receiving the PKMv2 RSA-Reply message, the mobile then uses its own private key to decrypt the pre-PAK and obtains the pre-PAK. The mobile also sends a PKMv2 RSA-Acknowledgment message to acknowledge the receipt of the PKMv2 RSA-Reply message. The exchange of messages is shown in Figure 12.2.

The successful outcome of RSA-based authorization is the transfer of the pre-PAK to the mobile; after possessing the pre-PAK, the mobile uses the pre-PAK to generate the primary AK (PAK). Then it uses the PAK to generate the AK (see Figure 12.3). The reason this procedure is called RSA-based is because the mobile's public key/private key pair (used to protect the pre-PAK) is based on the RSA asymmetric encryption algorithm.

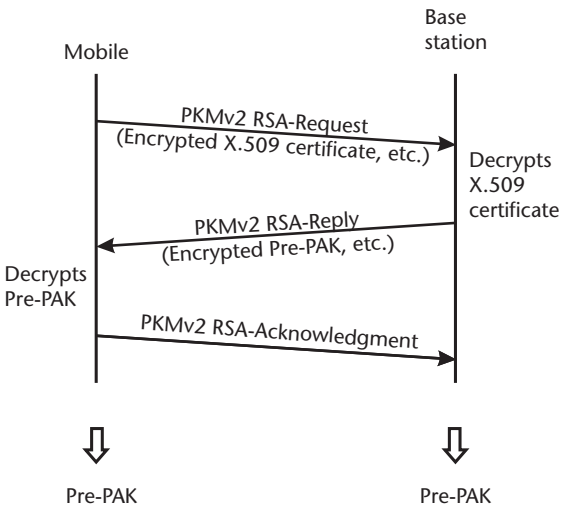


Figure 12.2 RSA-based authorization.

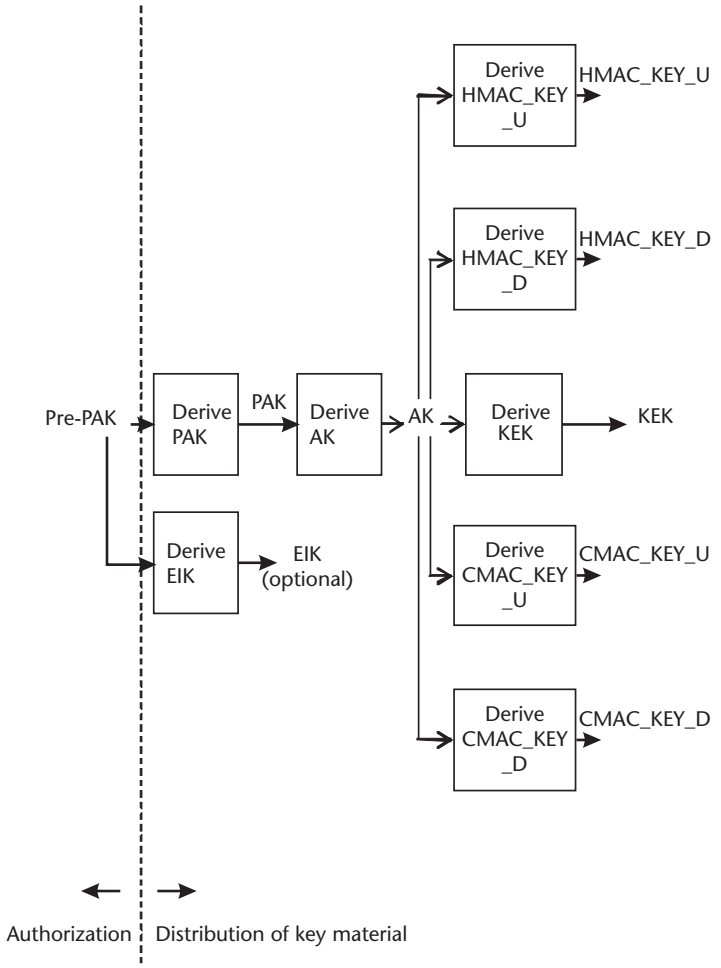


Figure 12.3 High-level key derivation scheme (for RSA-based authorization only). Note that the EAP integrity key (EIK) is used if an authenticated EAP-based authorization is used after RSA-based authorization.

12.3.2 EAP Based

The EAP-based technique is more flexible than the basic RSA-based one. In EAP, an entity also presents credentials as proof of its identity, but in EAP [14], the credentials are in the form that the network service provider or access provider specifies, such as a removable subscriber identity module (SIM). EAP is basically a protocol that enables the mobile to communicate with an authentication server (such as an authentication, authorization, and accounting, or AAA, server) on the fixed network. EAP is very flexible in that the *EAP* itself is separate from the *EAP method*. EAP is the authentication protocol, whereas the EAP method is the actual procedure used to acquire credentials. So an EAP method can be a SIM, a time-based token (e.g., SecureID), a biometric, or even an X.509 certificate. As such, EAP is called *extensible* because it can support a variety of different EAP methods.

EAP itself is a higher layer protocol (above the security sublayer) that enables the mobile and an authentication server to communicate with each other. Figure

12.4 shows the basic architecture with which EAP works. The architecture has three entities:

- The *supplicant* (e.g., the mobile), which seeks to be authenticated.
- The *authenticator* (e.g., the base station), which is the first entity on the network to which the mobile contacts regarding authentication. On a wireless network, the authenticator can be the base station or the ASN-gateway (ASN-GW). On a fixed network, the authenticator may be a gateway on the edge of the network.
- The *authentication server*, which has the database and the intelligence to make authentication decisions.

The mobile exchanges EAP messages with the authentication server through the base station. The responsibility of the security sublayer is to tunnel (encapsulate and decapsulate) EAP messages to and from the base station. In short:

- The EAP method acquires the credentials.
- EAP sends the mobile's credentials and authentication data (in EAP messages) to the authentication server through the base station.
- The security sublayer tunnels the EAP messages (in PKMv2 EAP messages).

Note that in this architecture, EAP runs on top of the remote authentication dial-in user service (RADIUS) [15] between the base station and the authentication server.

As Figure 12.4 shows, using EAP messages, the mobile and the authentication server exchange credentials and authentication data at a higher layer. These higher-layer EAP messages are tunneled using PKMv2 EAP messages (e.g., PKMv2 EAP Transfer message) at layer 2. Ultimately, it is the authentication server that makes

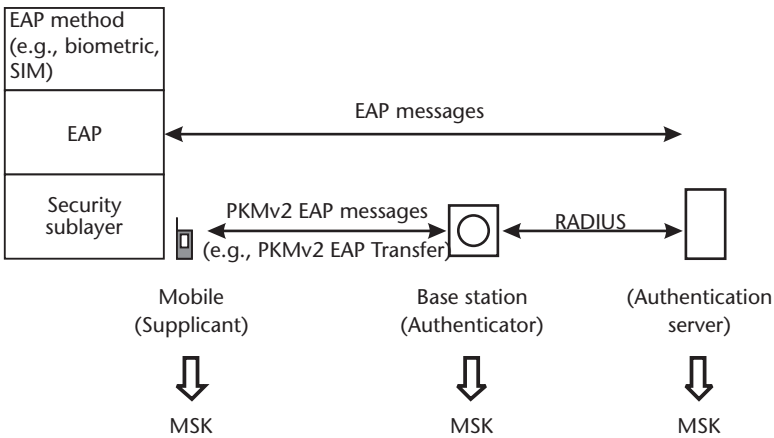


Figure 12.4 EAP-based authorization: basic architecture.

the decision on whether or not a mobile can enter the network. If the authentication process is successful, then the authenticator receives indication about the successful completion of the process, and the base station sends the EAP payload to the mobile in a PKMv2 EAP message [16].

The successful outcome of an EAP-based authorization is the transfer of the master session key (MSK) to the mobile; the MSK is used to derive the pairwise master key (PMK), and the PMK is used to derive the AK (see Figure 12.5).

12.3.3 Refresh of the AK

The AK generated for a mobile has a specified lifetime (i.e., an AK lifetime). When the lifetime expires, the AK becomes invalid. To remain connected, a mobile needs to obtain a new AK before it becomes invalid. Obtaining a new AK requires that the mobile be reauthorized. The procedure for reauthorization is similar to that of authorization. After a mobile is reauthorized, a fresh AK is generated for it.

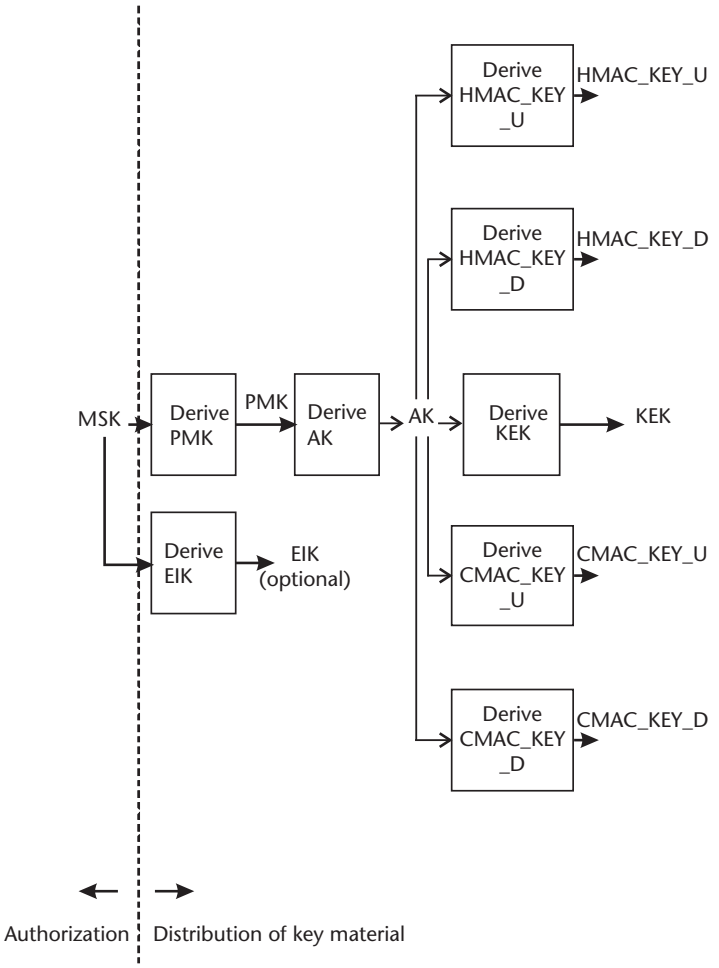


Figure 12.5 High-level key derivation scheme (for EAP-based authorization only). Note that the EAP integrity key (EIK) is used if an authenticated EAP-based authorization is used after EAP-based authorization.

12.4 Distribution of Key Materials

12.4.1 Prerequisite

After the system completes the authorization procedure (depending on the authorization policy negotiated) and a series of key derivations (also depending on the authorization policy negotiated), both the base station and the mobile now have the AK. In fact, the sharing of the AK is a prerequisite to the distribution of the TEK.

As Figures 12.3 and 12.5 show, the AK is used to derive the KEK; the AK is also used to derive the HMAC keys (i.e., HMAC_KEY_U and HMAC_KEY_D) and the CMAC keys (i.e., CMAC_KEY_U and CMAC_KEY_D). Note that HMAC_KEY_U and CMAC_KEY_U are used to sign management messages on the uplink, while HMAC_KEY_D and CMAC_KEY_D are used to sign management messages on the downlink. After these keys are derived, the system is now ready to distribute the TEK.

12.4.2 Distribution of TEK

In PKMv2, the system distributes the TEK to the mobile through a three-way handshake procedure. In a sense, the TEK is the most important as it is used to encrypt payloads of MAC PDUs. So the base station does not easily hand out the TEK; it has to be sure that it is actually handing out the TEK to the mobile that has been authorized. In other words, the three-way handshake procedure seeks to prove that the mobile in question actually has the AK.

Recall that the AK is used to derive the HMAC keys and the CMAC keys, and an HMAC or a CMAC key (depending on the MAC mode supported) is used to sign the message; the "signature" is the HMAC digest or the CMAC digest that is attached to the end of the message. In essence, the base station issues a challenge to the mobile by sending a pseudorandom number to the mobile. What the base station is looking for is that the mobile: (1) returns the same pseudorandom number back to the base station, (2) supplies the ID associated with the correct AK (i.e., AKID), and (3) signs the message containing the pseudorandom number using an HMAC key or a CMAC key derived from the correct AK. Once the base station verifies that the pseudorandom number is correct, the AKID is valid, and the attached HMAC digest or CMAC digest is generated using an HMAC key or a CMAC key derived from the correct AK, the base station can be sure that the mobile is alive and does have the correct AK. After the completion of these steps, the base station sends the TEK (encrypted using the KEK) to the mobile.

Operationally, the three-way handshake procedure consists of three messages exchanged between the base station and the mobile: the PKMv2 SA-TEK-Challenge message, the PKMv2 SA-TEK-Request message, and the PKMv2 SA-TEK-Response message (see Figure 12.6):

First, the base station sends the PKMv2 SA-TEK-Challenge message to the mobile. In this message:

- The base station provides a pseudorandom number, BS_RANDOM (generated at the base station).

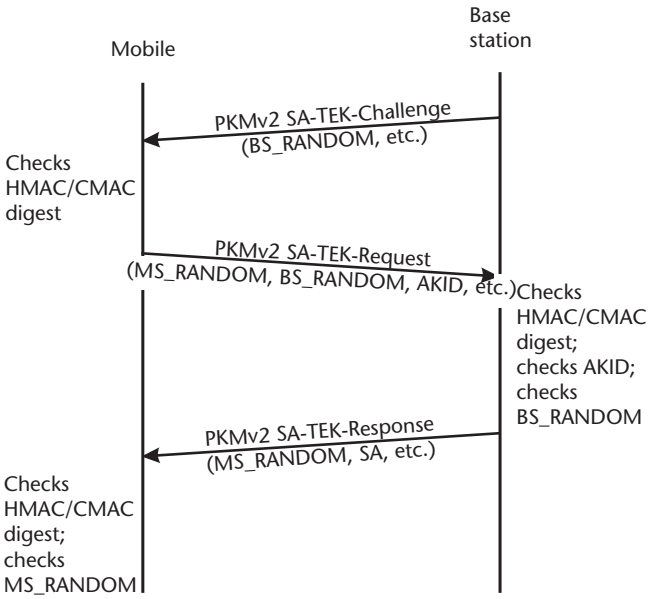


Figure 12.6 Three-way handshake.

- The base station uses the HMAC key or the CMAC key to sign the PKMv2 SA-TEK-Challenge message and attaches the HMAC digest or CMAC digest.

Second, after receiving the PKMv2 SA-TEK-Challenge message, the mobile verifies the HMAC digest or the CMAC digest. If the digest checks out, the mobile sends the PKMv2 SA-TEK-Request message to the base station. In this message:

- The mobile provides its own pseudorandom number, MS_RANDOM (generated at the mobile).
- The mobile returns the base station's pseudorandom number, BS_RANDOM.
- The mobile supplies the ID associated with its current AK (i.e., AKID). The AKID is an altered form of AK, not the AK itself. The AK itself should not travel in the clear over the air.
- The mobile uses the HMAC key or the CMAC key to sign the PKMv2 SA-TEK-Request message and attaches the HMAC digest or CMAC digest.

Third, after receiving the PKMv2 SA-TEK-Request message, the base station performs the following verifications:

- The base station verifies the HMAC digest or CMAC digest. If the HMAC digest or CMAC digest checks out, then the base station is convinced that the mobile has the correct AK.
- The base station verifies the AKID.
- The base station checks that the returned BS_RANDOM number matches the one sent by the base station in the previous PKMv2 SA-TEK-Challenge message. A matching BS_RANDOM shows that the mobile is alive.

If these items check out, the base station sends the PKMv2 SA-TEK-Response message to the mobile. In this message, the base station returns the mobile's pseudorandom number, MS_RANDOM (to show that the base station is alive). More importantly, this message contains the SA(s) that the mobile is authorized to use. Recall that an SA includes the SAID, the cryptographic suites, and other pieces of information that the mobile needs to protect a transport connection. Also, the message contains the TEK and its associated parameters issued under the SAID. (The TEK is encrypted by the KEK.)

Lastly, after receiving the PKMv2 SA-TEK-Response message, the mobile checks that the returned MS_RANDOM matches the one it sent in the previous PKMv2 SA-TEK-Request message. A match shows that the base station is alive. In addition, the mobile verifies the HMAC digest or CMAC digest of the message. If the HMAC digest or CMAC digest checks out, then the mobile is confident that the TEK and the associated parameters sent (in the PKMv2 SA-TEK-Response message) are valid and may begin to use them. Note that the TEK is a standalone key that is not derived from any other keys discussed so far. In fact, the TEK is generated randomly in the base station.

12.4.3 Refresh of TEK

It is important to recognize that the TEK also has a limited lifetime and needs to be periodically changed. If a key is not changed at all during a session, then the key becomes vulnerable to attacks. For example, using long-life keys was a weakness of wireless encryption protocol (WEP), an earlier security scheme used in the IEEE 802.11 systems.

Two messages are normally used to refresh the TEK: the PKMv2 Key-Request message and the PKMv2 Key-Reply message (see Figure 12.7).

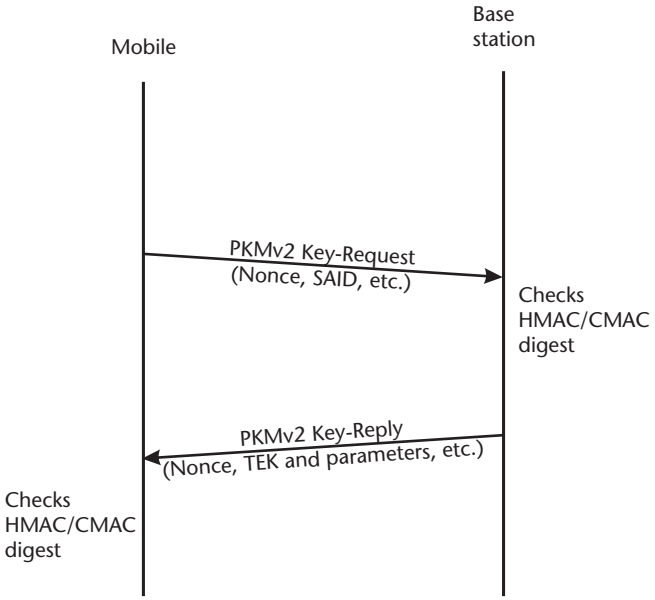


Figure 12.7 The PKMv2 Key-Request/PKMv2 Key-Reply exchange.

To refresh its TEK regularly, a mobile sends the PKMv2 Key-Request message to the base station at regular intervals. In this message:

- The mobile provides a random number called "nonce" (generated locally at the mobile).
- The mobile provides the SAID under which the mobile is authorized and issued the TEK.
- The mobile uses the HMAC key or the CMAC key to sign the PKMv2 Key-Request message and attaches the HMAC digest or CMAC digest.

After receiving the PKMv2 Key-Request message, the base station verifies the HMAC digest or the CMAC digest. If the digest checks out, the base station sends the PKMv2 Key-Reply message to the mobile. In this message:

- The base station returns the mobile's nonce.
- The base station supplies the TEK (encrypted by the KEK) and its associated parameters. One of the parameters is the remaining lifetime of this fresh TEK.
- The base station uses the HMAC key or the CMAC key to sign the PKMv2 Key-Reply message and attaches the HMAC digest or CMAC digest.

Upon receiving the PKMv2 Key-Reply message, the mobile verifies the HMAC digest or CMAC digest. If the HMAC digest or CMAC digest checks out, then the mobile accepts the TEK contained in the message. The use of nonce guards against replay attacks.

In actuality, the PKMv2 Key-Reply message contains two TEKs (and their parameters) for a single SAID. One TEK is designated the "newer" TEK and the other TEK is designated the "older" TEK. These two TEKs have overlapping lifetimes in that the newer TEK becomes active before the older TEK's lifetime ends. Using two overlapping TEKs, the system is assured that there is always a good TEK that can be used for encryption. In other words, the system minimizes the chance of connection disruption due to the delay of PKMv2 Key-Request/PKMv2 Key-Reply exchange. This is because the mobile conducts the PKMv2 Key-Request/PKMv2 Key-Reply exchange after the expiration of the older TEK but before the expiration of the newer TEK.

Each TEK actually comes with a *TEK sequence number* (as the mobile stops using old TEKs and receives new TEKs). The TEK sequence number is one of the parameters associated with a TEK. So the first TEK may have a TEK sequence number = 0, the second TEK has a TEK sequence number = 1, the third TEK has a TEK sequence number = 2, and the fourth TEK has a TEK sequence number = 3. Then the next TEK's TEK sequence number would go back to 0 and so on. The TEK sequence number is a 2-bit number (i.e., modulo 4). In fact, the TEK sequence number is shown by the encryption key sequence (EKS) field in the unencrypted generic MAC header (see Chapter 7). By showing the TEK sequence number, the generic MAC header tells the receiver which TEK to use to decrypt the payload.

A precondition for a mobile to refresh the TEK is that it remains authorized. In fact, if a mobile has to be reauthorized for any reason, then the TEK refresh

procedure already in progress would wait until the mobile is successfully reauthorized. Note that the TEK is refreshed more frequently than the AK.

12.5 Possible Vulnerabilities

12.5.1 Fixed Network

If an end user has enabled a higher-layer encryption such as Internet Protocol Security (IPSec)[4] or Secure Socket Layer (SSL),[5] he or she may assume that encryption provided by the security functions in MAC is no longer necessary. That assumption may not be valid. In the case where the IP payload is encrypted (using the IPSec transport mode) but MAC PDU payload is not encrypted, the user data would be private because IPSec encrypts the IP payload[6] (see Figure 12.8). However, without encrypting MAC PDU payload, the IP addresses in the IP header is in the clear and vulnerable to intercept over the air interface. In this case, the traffic (user data) is private, but the traffic pattern (IP addresses) is not private [17].

On the other hand, if the system only enables encryption of the MAC PDU payload, an end user may assume that user data is completely private. That assumption also may not be valid. In the case where the MAC PDU payload is encrypted but IP

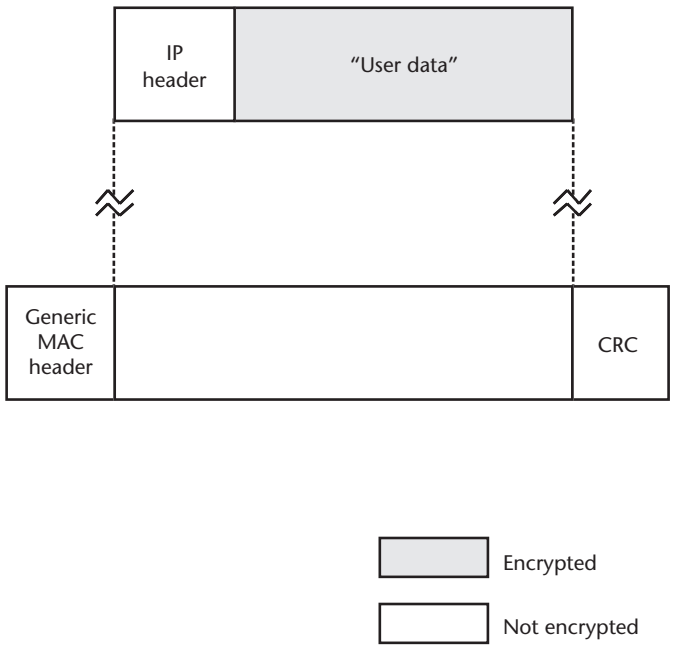


Figure 12.8 A case where IP payload is encrypted but MAC PDU payload is not encrypted. (IPSec transport mode is used for illustration.)

4. IP security protocol (IPSec) is a higher layer protocol that encrypts IP packets.
5. Secure Socket Layer (SSL) is a higher layer protocol that encrypts application packets. SSL operates below the application layer.
6. The IPSec transport mode only encrypts the IP payload, not the header.

payload is not encrypted, both the user data and the IP addresses would be private over the air interface (see Figure 12.9), but the user data is vulnerable in the nodes of the core network. This vulnerability comes about because a node in the core network must decrypt the MAC PDU payload to extract the IP addresses, which are needed to route the IP packet through the core network. As the payload of a MAC PDU is decrypted in the network, both the IP addresses (in the IP header) and the user data are exposed. In this case, the air interface is private, but the traffic (user data) is vulnerable in the nodes of the core network. Nevertheless, the nodes in the private core network operated by a service provider should be well protected; as such, the risk there should be small.

For better protection, one can enable both encryption of MAC PDU payload and encryption of IP payload. Encryption of MAC PDU payload protects the air interface, and encryption of IP payload protects the user data when an IP packet is temporarily in the clear in a node of the core network. For reference, Figure 12.10 shows a framework of vulnerabilities for different combinations of encryption.

12.5.2 Air Interface

The reader at this point probably recognizes that there is also vulnerability in the air interface. If encryption is activated in layer 2, only the MAC PDU payload is encrypted; the generic MAC header is not encrypted and is in the clear. If an eavesdropper intercepts a MAC PDU over-the-air, the eavesdropper cannot read the payload, but he or she can still read the header and extract some information regarding the link, such as the CID. However, CID information is assigned by the network access provider and is only meaningful within the provider's own private network. Hence CID information is not as useful to an attacker. This is in contrast to public IP addresses, which if intercepted, can provide much information about

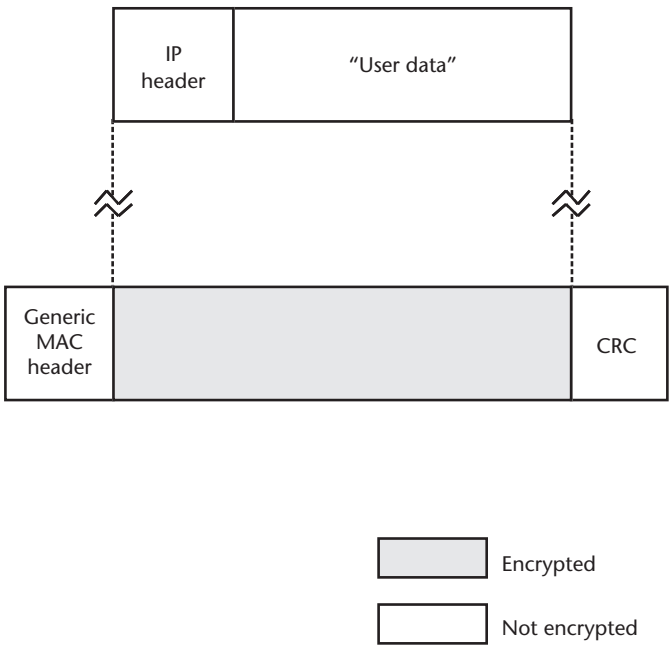


Figure 12.9 A case where MAC PDU payload is encrypted but IP payload is not encrypted.

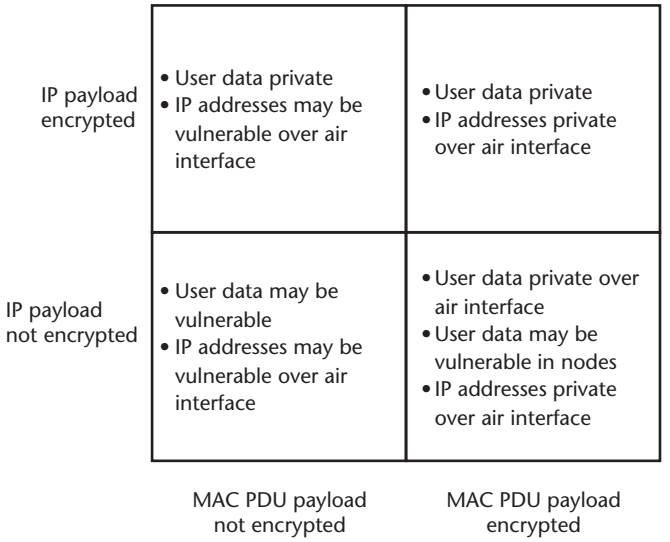


Figure 12.10 A framework of vulnerabilities for different combinations of encryption.

identities of parties and patterns of communication. In addition, all MAC management messages travel over-the-air unencrypted; hence, these messages are subject to interception as well.

References

[1] National Institute of Standards and Technology (NIST), *Standards for Security Categorization of Federal Information and Information Systems*, Federal Information Processing Standards (FIPS) Publication 199, February 2004.

[2] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[3] National Institute of Standards and Technology (NIST), *Data Encryption Standard (DES)*, Federal Information Processing Standards (FIPS) Publication 46-3, October 1999.

[4] National Institute of Standards and Technology (NIST), *DES Modes of Operation*, Federal Information Processing Standards (FIPS) Publication 81, December 1980.

[5] RSA Laboratories, "PKCS #1 v2.0: RSA Cryptography Standard," October 1998.

[6] National Institute of Standards and Technology (NIST), *Advanced Encryption Standard (AES)*, Federal Information Processing Standards (FIPS) Publication 197, November 2001.

[7] National Institute of Standards and Technology (NIST), *Recommendation for Block Cipher Modes of Operation: The CCM Mode for Authentication and Confidentiality*, NIST Special Publication 800-38C, May 2004.

[8] The Internet Engineering Task Force (IETF), "Counter with CBC-MAC (CCM)," IETF RFC 3610, September 2003.

[9] National Institute of Standards and Technology (NIST), *Recommendation for Block Cipher Modes of Operation: Methods and Techniques*, NIST Special Publication 800-38A, December 2001.

[10] The Internet Engineering Task Force (IETF), "HMAC: Keyed-Hashing for Message Authentication," IETF RFC 2104, February 1997.

[11] National Institute of Standards and Technology (NIST), *Secure Hash Standard*, Federal Information Processing Standards (FIPS) Publication 180-1, April 1995.

[12] National Institute of Standards and Technology (NIST), *Recommendation for Block Cipher Modes of Operation: The CMAC Mode for Authentication*, NIST Special Publication 800-38B, May 2005.

[13] The Internet Engineering Task Force (IETF), "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile," IETF RFC 3280, April 2002.

[14] The Internet Engineering Task Force (IETF), "Extensible Authentication Protocol (EAP)," IETF RFC 3748, June 2004.

[15] The Internet Engineering Task Force (IETF), "Remote Authentication Dial In User Service (RADIUS)," IETF RFC 2865, June 2000.

[16] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

[17] Stallings, W., *Business Data Communications*, Upper Saddle River, NJ: Prentice-Hall, 2009.

CHAPTER 13

RF System Design: Coverage

13.1 Introduction

When designing a mobile broadband wireless system, the designer needs to pay attention to its special requirements [1]. Ensuring that a system can deliver adequate service over a required coverage area is often the most important part of system design. When the system is first designed, it is initially sized to satisfy the coverage requirement. After satisfying the coverage requirement, this initial configuration forms the baseline from which the system can grow as more subscribers are added. As such, designing for coverage remains fundamentally the most critical step in the design of any system [2].

For a mobile broadband wireless system, satisfying the coverage requirement typically means that the radio link must have certain quality (i.e., meet or exceed some minimum required SINR). Because the system attempts to meet some required SINR, satisfying the coverage requirements means that: (1) the ensemble of base stations in the aggregate can deliver enough RF power to the required coverage area, (2) they can receive adequate RF power in the same area, and (3) each receiver in the system does not experience excessive interference.

13.2 Link Quality

13.2.1 SINR

The signal-to-interference and noise ratio (SINR) is the de facto figure of merit of an RF system, and this ratio can be derived for an OFDMA-based system. For the SINR, the received power S for the jth subcarrier of the ith user is:

$$S^{i,j} = P_T^{i,j} L_p^i u_g^i u_l^i G_R^i \tag{13.1}$$

where

- $S^{i,j}$ is the received power for the jth subcarrier of the ith user;
- $P_T^{i,j}$ is the EIRP for the jth subcarrier of the ith user (in the direction of the receiver for user i);

- L_p^i is the propagation loss between the transmitter and the receiver for user *i*;
- u_g^i consists of other gains provided by the system (e.g., transmit diversity gain, macro diversity gain) for user *i*;
- u_l^i consists of other propagation-related losses between the transmitter and the receiver (e.g., log-normal shadowing loss, fast fading loss, in-car penetration loss) for user *i*;
- G_R^i is the receive antenna gain in the direction of the transmitter for user *i*.

Interference can result if two users in nearby cells use subcarriers at the same frequency at the same time; such a collision of subcarriers degrades the received SINR of individual subcarriers. The received cochannel interference *I* perceived by the *j*th subcarrier of the 0th user is (assuming that user 0 is the user for whom the SINR is calculated):

$$I^{0,j} = \sum_{i=1}^{M} P_T^{i,j} L_p^i u_g^i u_l^i G_R^i \tag{13.2}$$

where

- $I^{0,j}$ is the received interference perceived by the *j*th subcarrier of user 0;
- $P_T^{i,j}$ is the EIRP for the *j*th subcarrier of the *i*th user (in the direction of the receiver for user *i*);
- L_p^i is the propagation loss between the transmitter and the receiver for user *i*;
- u_g^i consists of other gains provided by the system (e.g., transmit diversity gain, macro diversity gain) for user *i*;
- u_l^i consists of other propagation-related losses between the transmitter and the receiver (e.g., log-normal shadowing loss, fast fading loss, in-car penetration loss) for user *i*;
- G_R^i is the receive antenna gain in the direction of the transmitter for user *i*.
- *M* is the number of cochannel interfering users.

Obviously, $P_T^{i,j}$ $(i \neq 0)$ all contribute to the interference received by user 0 (at user 0's *j*th subcarrier).

The received noise *N* perceived by a single subcarrier is:

$$N = kT\Delta f \tag{13.3}$$

where

- k is the Boltzmann's constant (1.38×10^{-23} watt/Hz·K or –198.6 dBm/Hz·K);
- *T* is the system temperature;
- Δf is the frequency spacing between subcarriers.

Given the above derivations, the SINR for the *j*th subcarrier of user 0 is simply:

$$SINR^{0,j} = \frac{S^{0,j}}{N + I^{0,j}} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{kT\Delta f + \sum_{i=1}^{M} P_T^{i,j} L_p^i u_g^i u_l^i G_R^i} \tag{13.4}$$

where

- $SINR^{0,j}$ is the SINR for the jth subcarrier of user 0;
- $P_T^{0,j}$ is the EIRP for the jth subcarrier of user 0 (in the direction of the receiver for user 0);
- L_p^0 is the propagation loss between the transmitter and the receiver for user 0;
- G_R^0 is the receive antenna gain in the direction of the transmitter for user 0.

Because the variables of (13.1), (13.2), (13.3), and (13.4) are in terms of the transmitter and the receiver (not the base station and the mobile), these equations are valid for both the uplink and the downlink. Note that the SINR expression just derived is per subcarrier (i.e., for the jth subcarrier), so if the channel is frequency selective, each subcarrier would experience a different fading.

Figure 13.1 illustrates (13.4) from the perspective of user 0 on the downlink and shows user 0 in one cell and three interfering users (user 1, user 2, and user 3)

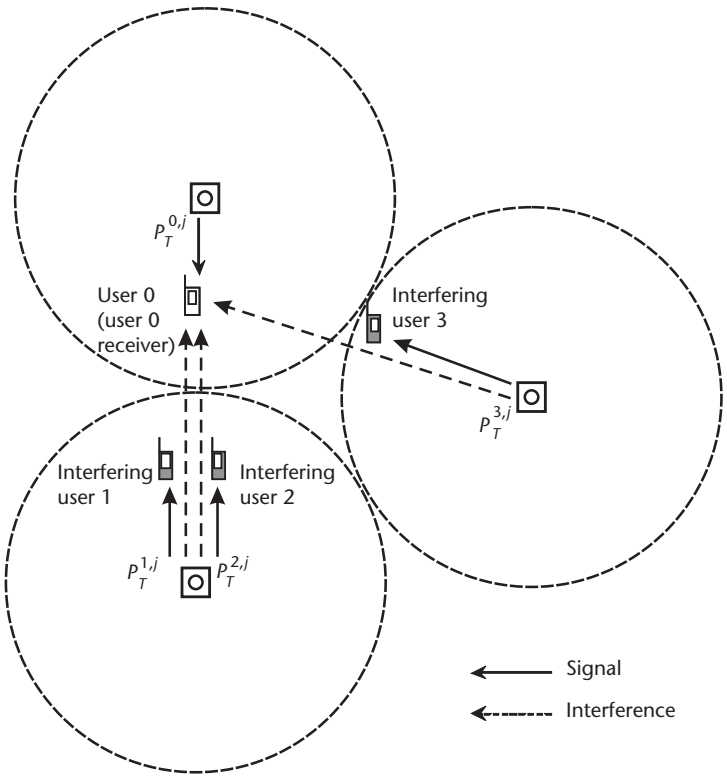


Figure 13.1 User 0 and its interferers on the downlink. User 0's receiver is at the mobile, so the SINR is measured at the mobile.

in two neighboring cells. On the downlink, user 0's receiver is at the mobile. The downlink signals meant for user 1, user 2, and user 3 become interference directed at user 0.

Figure 13.2 shows the same situation from the perspective of user 0 on the uplink. On the uplink, user 0's receiver is at user 0's home base station. Because mobiles typically have omnidirectional transmit antennas, transmissions from user 1's, user 2's, and user 3's mobiles inadvertently reach user 0's base station and constitute interference for user 0. In both Figure 13.1 and Figure 13.2, the assumptions are that the neighboring cells are reusing the same frequencies as the home cell, and that the interfering users are transmitting on the same subcarrier(s) as user 0.

13.2.2 SNR and SIR

A parameter related to the SINR is the signal-to-noise ratio (SNR), which has only the noise term in the denominator. The SNR for the *j*th subcarrier of user 0 is:

$$SNR^{0,j} = \frac{S^{0,j}}{N} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{kT\Delta f} \tag{13.5}$$

where $SNR^{0,j}$ is the SNR for the *j*th subcarrier of user 0.

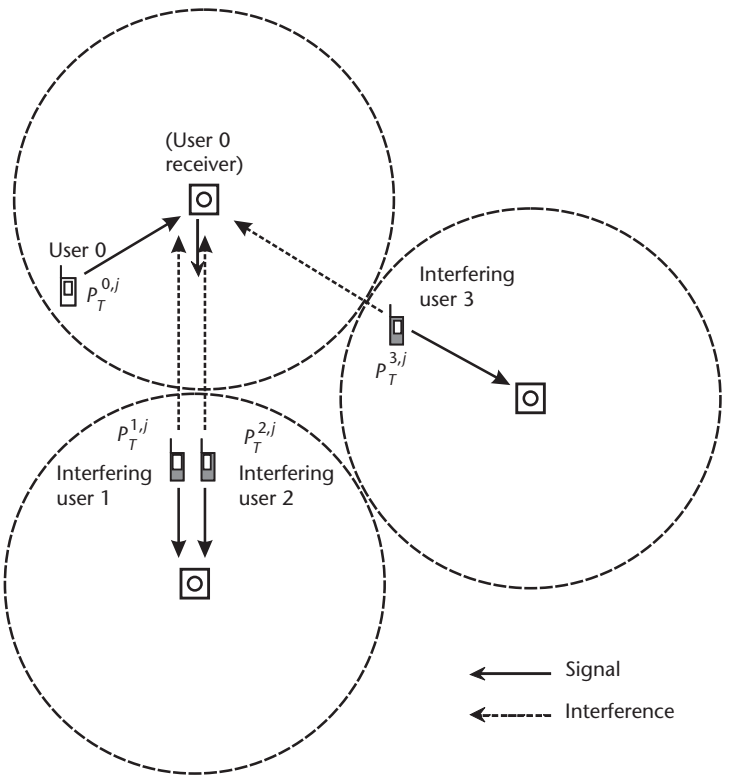


Figure 13.2 User 0 and its interferers on the uplink. User 0's receiver is at the base station, so the SINR is measured at the base station.

Because the interference term in the denominator may be much higher than the noise term (especially in a system with tight frequency reuse), another related parameter called signal-to-interference ratio (SIR) is used. The SIR, which has only the interference term in the denominator, for the *j*th subcarrier of user 0 is:

$$SIR^{0,j} = \frac{S^{0,j}}{I^{0,j}} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{\sum_{i=1}^{M} P_T^{i,j} L_p^i u_g^i u_l^i G_R^i} \tag{13.6}$$

where $SIR^{0,j}$ is the SIR for the *j*th subcarrier of user 0.

13.2.3 Interference

For the multiple-cell case, a user can experience interference from other cells (or sectors) nearby. To calculate the interference term I for a user, one has to apply (13.2) and sum up interference contributions from all other users. On the uplink, this calculation involves summing up transmissions from other cochannel mobiles nearby. On the downlink, this calculation involves summing up transmissions from other cochannel base stations nearby. In the context of system design, one way to quantify the effect of interference is by using an interference margin.

For a particular user (e.g., user 0), an *interference margin* I_M can be computed. The interference margin is the "hurdle" that the received power has to overcome to achieve an acceptable SINR. This margin is often expressed as an interference "rise" above the noise floor [2]. For user 0, the interference margin for the *j*th subcarrier is:

$$I_M^{0,j} = \frac{N + I^{0,j}}{N} \tag{13.7}$$

which is equivalent to

$$I_M^{0,j} = \frac{SNR^{0,j}}{SINR^{0,j}} \tag{13.8}$$

Given (13.8), the equation for the SINR is

$$SINR^{0,j} = \frac{SNR^{0,j}}{I_M^{0,j}} \tag{13.9}$$

which makes sense because the SINR is effectively the SNR degraded by the interference rise I_M (contributed by users in other cells). Substituting the equation for the SNR into (13.9) produces a more succinct expression for the SINR:

$$SINR^{0,j} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}} \tag{13.10}$$

For the uplink, the receiver is located at the base station, and the uplink interference rise perceived at the base station comes from other mobiles in neighboring cells. Thus, on the uplink, the interference rise seen by the base station is independent of the distance between itself and user 0's mobile. In other words, the uplink interference rise is constant as a function of distance to user 0's mobile.

For the downlink, the situation is a bit more involved. Here the receiver is located at user 0's mobile, and the downlink interference rise perceived at user 0's mobile comes from other base stations in neighboring cells. As the distance between user 0's mobile and its home base station changes, the interference received at user 0's mobile also changes and is indeed a function of distance to user 0's mobile. As the distance changes, user 0's mobile would receive less interference from some neighboring base stations and more interference from some other neighboring base stations. In practice, a system designer sometimes assumes a (worst-case) interference margin on the downlink to account for the effect of interference from other neighboring base stations.

13.2.4 Noise

Quantifying the noise term N is always an important part of system design. Because $N = kT\Delta f$, computing N requires knowing what the system temperature T is. It is well known that in a radio system, the system temperature T is the sum of the antenna temperature T_{ant} of the antenna and the composite effective temperature T_{comp} of the receiving equipment,

$$T = T_{ant} + T_{comp} \tag{13.11}$$

where the receiving equipment consists of both the cable and the radio front end. Figure 13.3 shows the configuration of a generic base station receiving system.

The composite effective temperature T_{comp} of the receiving equipment (i.e., both the cable and the radio front end) has to be derived from the composite noise

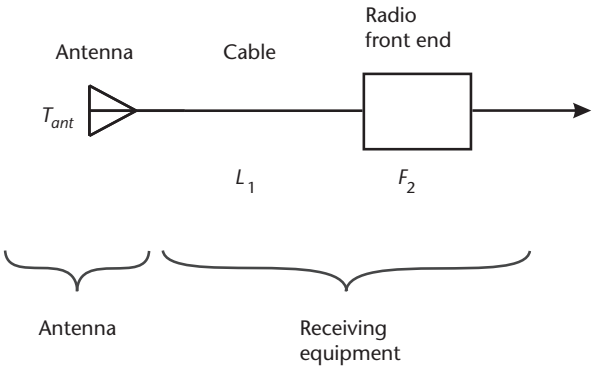


Figure 13.3 A base station receiving system including the antenna, cable, and radio front end.

figure F_{comp} of the receiving equipment. The composite noise figure F_{comp} of two components in series is

$$F_{comp} = F_1 + \frac{F_2 - 1}{G_1} \tag{13.12}$$

where

- F_1 is the noise figure of the first component in the series;
- G_1 is the gain of the first component in the series;
- F_2 is the noise figure of the second component in the series.

For the receiving equipment, the first component is the cable, and the second component is the radio front end. Recognizing that the gain G_1 of a cable is simply the inverse of its loss (i.e., $G_1 = 1/L_1$) and the noise figure F_1 of a cable is the same as its loss[1] (i.e., $F_1 = L_1$), we can rewrite (13.12) as [3]:

$$F_{comp} = L_1 + L_1(F_2 - 1) \tag{13.13}$$

For any receiving equipment, the composite noise temperature is

$$T_{comp} = (F_{comp} - 1)290\text{K} \tag{13.14}$$

Substituting (13.13) into (13.14) gives us

$$T_{comp} = [L_1 + L_1(F_2 - 1) - 1]290K = (L_1F_2 - 1)290\text{K} \tag{13.15}$$

Therefore, an equation for the system temperature T can be derived by substituting (13.15) into (13.11). If we assume that the antenna temperature T_{ant} is 290K, then

$$\begin{aligned} T &= T_{ant} + T_{comp} = 290\text{K} + (L_1F_2 - 1)290\text{K} \\ &= [1 + (L_1F_2 - 1)]290\text{K} = (L_1F_2)290\text{K} \end{aligned} \tag{13.16}$$

In other words, the system temperature T of a complete base station receiving system (antenna, cable, and radio front end) is the loss of the cable multiplied by the noise figure of the radio front end multiplied by the ambient temperature 290K [3].

1. This makes sense because the noise figure of a component quantifies how much the signal-to-noise ratio degrades as the signal goes through the component. Since a piece of cable simply introduces a loss, the noise figure F_1 of a cable is the same as its loss (i.e., $F_1 = L_1$).

13.3 Designing for Coverage

13.3.1 Fundamentals

Having developed the expressions for SNR, SIR, and SINR, we now proceed to investigate the extent of the coverage provided by a cell using OFDMA. A useful tool at the disposal of system designers is the link budget. The link budget is a way of accounting for various gains and losses of the system for the purpose of computing coverage. As such, the bottom line of a link budget (for a terrestrial wireless system) is typically the maximum allowable propagation loss that can be sustained if one is to meet a given required level of SINR.

Before proceeding, we need to emphasize the difference between calculated SINR (SINR) and required SINR ($SINR_{req}$). The *calculated SINR* is computed by using (13.4) when all the parameters on the right side of the equation are available:

$$SINR^{0,j} = \frac{S^{0,j}}{N + I^{0,j}} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{kT\Delta f + \sum_{i=1}^{M} P_T^{i,j} L_p^i u_g^i u_l^i G_R^i}$$

Once computed, the calculated SINR gives system designers an idea of the actual SINR based on the given system parameters.

The *required SINR*, on the other hand, is the minimum SINR that a link has to attain for successful demodulation to occur at the receiver. For a given required SINR, one can solve for propagation loss in (13.4):

$$L_{p,\max}^0 = \frac{SINR_{req}^{0,j}\left(kT\Delta f + \sum_{i=1}^{M} P_T^{i,j} L_p^i u_g^i u_l^i G_R^i\right)}{P_T^{0,j} u_g^0 u_l^0 G_R^0} \tag{13.17}$$

where $SINR_{req}^{0,j}$ is the required SINR for the jth subcarrier of user 0. The resulting propagation loss is the *maximum allowable propagation loss* that can be sustained given all the system parameters, including the required SINR. Using a suitable path loss model, one can then convert this maximum allowable propagation loss into a maximum coverage distance (e.g., to the cell edge) within which a service level (e.g., bit rate) can be supported [4].

Equation (13.17) is accurate for the maximum allowable propagation loss. However, for the link budget, system designers often use the more succinct SINR equation in (13.10) containing the interference margin. Specifically, solving the more succinct SINR (13.10) for the propagation loss yields:

$$L_{p,\max}^0 = \frac{SINR_{req}^{0,j}\left(kT\Delta f\right) I_M^{0,j}}{P_T^{0,j} u_g^0 u_l^0 G_R^0} \tag{13.18}$$

Equation (13.18) is often used for link budgets because the parameter interference margin I_M is simply another item that is added to the budget.

Before presenting an example of the link budget, we need to describe another commonly used parameter called *receiver sensitivity*. The receiver sensitivity is defined as the minimum received power required for signal demodulation (not considering the effect of interference) [4]. Formally, the receiver sensitivity *RS* is:

$$RS = SINR_{req}(kTW) \tag{13.19}$$

In other words, the receiver sensitivity is the minimum required SINR multiplied by (only) the noise. If there is no interference (e.g., from users in other cells), then the receiver sensitivity is the same as the minimum received power required. Specified as a function of the required SINR (part of the equipment specification) and noise of the receiving system (also part of the equipment specification—noise figure), the receiver sensitivity becomes a parameter that depends solely on the receiving equipment, not on the interference margin which may change from system to system. Note that the receiver sensitivity is typically specified for the entire bandwidth, not for a subcarrier.

Since there is interference in a real system, the effect of interference can be included by using the interference margin elsewhere in the link budget.

13.3.2 Link Budget

Based on (13.18), Table 13.1 shows a case of a link budget for both the downlink and the uplink. The case is for N_{FFT} = 1,024 in a mobility situation.

Some items in the link budget in Table 13.1 merit the following explanations:

- The boldfaced parameters denote those that appear in (13.18).
- By dividing EIRP by the number of occupied subcarriers, the link budget assumes that subcarriers have equal power (line 7).
- The interference margin is 2 dB for the downlink and 3 dB for the uplink (line 14), assuming a frequency reuse pattern of 1 × 3 × 1 (see Chapter 14). If the frequency reuse pattern changes to 1 × 3 × 3, then the interference margin reduces to 0.2 dB, but the tradeoff is the reduced amount of frequency reuse in the system [5].
- The –10 dB of log-normal shadowing loss (line 21) assumes that the standard deviation of loss variation is 6 dB and 1% of outage probability at cell edge.

Regarding the log-normal shadowing loss, if in the area around a base station, the standard deviation of loss variation is 6 dB (obtained through actual measurement, for example). Then, based on the log-normal distribution with a standard deviation of 6 dB, 1% is the probability that the actual loss will be –10 dB relative to the median. In other words, in a log-normally distributed propagation environment with a standard deviation of 6 dB:

- Achieving an outage probability of 1% requires an extra link margin of 10 dB.

Table 13.1 Link Budget

Line Number	*Gain/Loss*	*Link*		*Comments*
		Downlink	*Uplink*	
Line 1	Tx power per antenna	43 dBm	23 dBm	
Line 2	Tx line loss	−3 dB	0 dB	
Line 3	Tx antenna gain	18 dBi	0 dBi	
Line 4	No. of Tx antennas	2	1	
Line 5	**EIRP, P_T**	61.0 dBm	23.0 dBm	=L1+L2+L3+10log(L4)
Line 6	Number of occupied subcarriers	840	192	PUSC permutation
Line 7	**EIRP, P_T per subcarrier**	31.8 dBm	0.2 dBm	=L5−10log(L6)
Line 8	Rx line loss, L_1	0 dB	0 dB	
Line 9	Rx noise figure, F_2	8 dB	4 dB	
Line 10	**Noise temperature, T**	32.6 dBK	28.6 dBK	=L8+L9+10log(290K)
Line 11	Boltzmann's constant, k	−198.6 dBm/HzK	−198.6 dBm/HzK	
Line 12	Spacing between subcarriers	10,937.5 Hz	10,937.5 Hz	
Line 13	**Received noise, N**	−125.6 dBm	−129.6 dBm	=L10+L11+10log(L12)
Line 14	**Interference margin, I_M**	2 dB	3 dB	
Line 15	Modulation/coding	16-QAM, R = 1/2 CTC	QPSK, R=1/2 CTC	
Line 16	**Required SINR, $SINR_{req}$**	8.6 dB	2.9 dB	=function(L15)
Line 17	Tx diversity gain	0 dB	0 dB	
Line 18	Macro diversity gain	0 dB	0 dB	
Line 19	Rx antenna diversity gain	0 dB	3 dB	
Line 20	**Other system gains, u_g**	0 dB	3 dB	=L17+L18+L19
Line 21	Log-normal shadowing loss	−10 dB	−10 dB	99% availability with SD=6 dB
Line 22	Fast-fading loss	−5 dB	−5 dB	
Line 23	Body/in-car loss	0 dB	0 dB	
Line 24	**Propagation-related losses, u_l**	−15 dB	−15 dB	=L21+L22+L23
Line 25	**Rx antenna gain, G_R**	−1 dBi	18 dBi	
Line 26	**Max. allowable path loss, $L_{p,\max}$**	−130.8 dB	−129.9 dB	=(L16+L14+L13) − (L7+L20+L24+L25)

- Achieving an availability of 99% requires an extra link margin of 10 dB.

The link budget includes this factor (–10 dB) to account for the shadowing loss, which results in a reduction in the maximum allowable path loss and ultimately in a reduction in the size of the cell.

Several observations can be made regarding the link budget. The first observation is that the maximum allowable propagation losses shown are similar between the downlink and the uplink. In particular, the downlink can sustain a loss of up to –131 dB, whereas the uplink can sustain a loss of up to –130 dB. If the maximum allowable propagation losses are similar, then the links are said to be *balanced*. Balanced downlink and uplink are typically the case if the links are *asymmetrical*. In this case, the downlink uses 16-QAM with $R = 1/2$ convolutional turbo code (CTC), while the uplink uses QPSK with $R = 1/2$ CTC, resulting in a higher data rate on the downlink. Because the downlink can transmit at a much higher power than the uplink, the asymmetrical data rates help to balance the link as far as the maximum allowable path losses are concerned. Asymmetrical links typically support Web-oriented applications.

The second observation about the link budget is that one can use the maximum allowable propagation loss to compute the maximum coverage distance by using a loss model. Chapter 2 examined several path loss models. As an illustration, the simple empirical model of path loss (for user i) is:

$$L_p^i = \left(\frac{c}{4\pi f d_{ref}}\right)^2 \left(\frac{d_{ref}}{d_i}\right)^{\alpha} \tag{13.20}$$

where

- α is the path loss exponent;
- d_i is the distance between the transmitter and the receiver for user i;
- f is the carrier frequency;
- c is the speed of light;
- d_{ref} is the reference distance (which can be taken to be 1m).

Solving for d_i yields:

$$d_i = d_{ref}\left[\left(\frac{c}{4\pi f d_{ref}}\right)^2 \frac{1}{L_p^i}\right]^{1/\alpha} \tag{13.21}$$

On the downlink, substituting a loss of –131 dB (or 8.32×10^{-14}), a frequency of 700 MHz, and α of 3 into (13.21) produces a maximum coverage distance of 2.4 km. On the uplink, the same calculation (using a loss of –130 dB) produces a maximum coverage distance of 2.2 km. Of course, one can use other path loss models, such as Okumura-Hata or COST-231 Hata, to compute the maximum coverage distance.

In the link budget, the required SINR is a function of the modulation and coding scheme used on the link. The link budget assumes that the downlink is designed to support 16-QAM with $R = 1/2$ CTC, which requires an SINR of 8.6 dB. The higher the order of the modulation, the more bits can be transmitted in a given

bandwidth. Because a higher-order modulation requires a higher SINR, the higher bit rate requires a higher SINR. Other required SINR values can be found in refrences such as [6, 7], as well as from an infrastructure vendor's recommendations.

13.3.3 Analytical Model

Another way to examine the extent of coverage provided by a cell is through a closed-form analytical model, which allows a designer to move directly between required SINR and the maximum coverage distance. To reiterate, the SINR in (13.10) was previously derived to be:

$$SINR^{0,j} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}}$$

Again, we invoke the simple path loss model (for user 0):

$$L_p^0 = \left(\frac{c}{4\pi f d_{ref}}\right)^2 \left(\frac{d_{ref}}{d_0}\right)^\alpha \tag{13.22}$$

Substituting the simple path loss model for L_p^o in (13.10) yields

$$SINR^{0,j} = \frac{P_T^{0,j} u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}} \left(\frac{c}{4\pi f d_{ref}}\right)^2 \left(\frac{d_{ref}}{d_0}\right)^\alpha \tag{13.23}$$

Figure 13.4 shows the SINR as a function of the distance of the user according to (13.23). For illustration, the figure shows the downlink SINR as a function of the distance, and the assumptions are identical to those in the link budget in Table 13.1.

Figure 13.4 and other SINR graphs generated using (13.23) turn out to be very useful for system designers. They provide a quick way of computing what the SINR might be at a particular distance. For a certain distance, one simply has to look at the graph to obtain an approximation of the SINR achievable at that distance.

For those who would like to go from SINR to coverage distance, (13.23) can be inverted to perform that calculation. We now proceed to invert (13.23) and solve for the distance as a function of SINR. Rearranging (13.23) produces:

$$\left(\frac{d_{ref}}{d_0}\right)^\alpha = \frac{SINR^{0,j}}{\dfrac{P_T^{0,j} u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}} \left(\dfrac{c}{4\pi f d_{ref}}\right)^2} \tag{13.24}$$

or

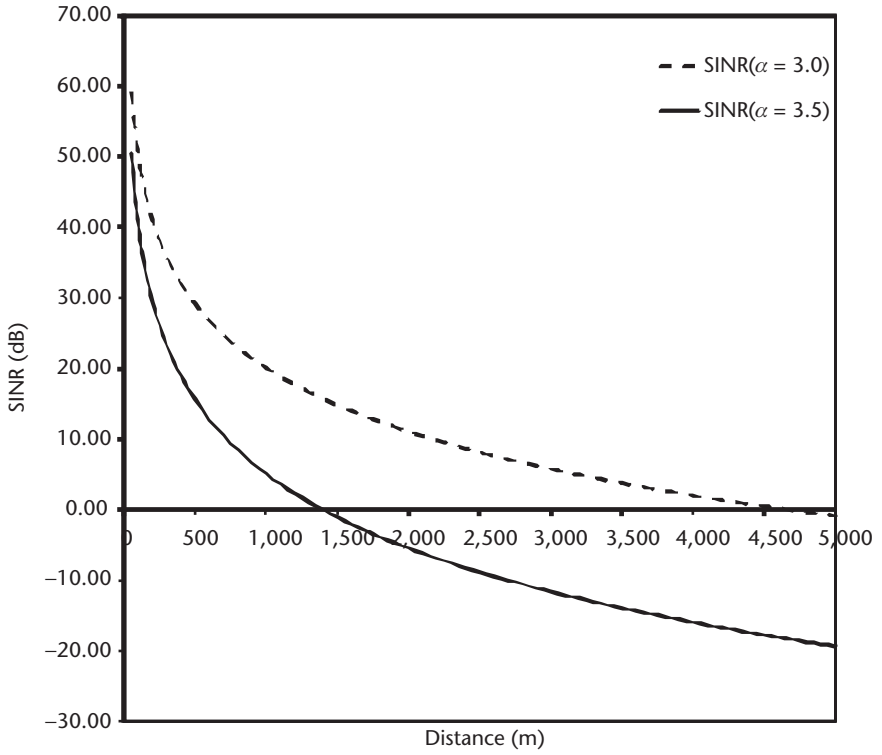


Figure 13.4 The downlink SINR as a function of the distance. The assumptions are identical to those in the link budget in Table 13.1. The operating frequency is 700 MHz, and the interference margin is kept constant as a function of the distance.

$$\left(\frac{d_0}{d_{ref}}\right)^{\alpha} = \frac{\dfrac{P_T^{0,j} u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}} \left(\dfrac{c}{4\pi f d_{ref}}\right)^2}{SINR^{0,j}} \tag{13.25}$$

Solving for d_0 yields:

$$d_0^{\alpha} = d_{ref}^{\alpha} \frac{\dfrac{P_T^{0,j} u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}} \left(\dfrac{c}{4\pi f d_{ref}}\right)^2}{SINR^{0,j}} \tag{13.26}$$

or

$$d_0 = d_{ref} \left(\frac{P_T^{0,j} u_g^0 u_l^0 G_R^0}{(kT\Delta f) I_M^{0,j}} \left(\frac{c}{4\pi f d_{ref}}\right)^2 \Big/ SINR^{0,j} \right)^{1/\alpha} \tag{13.27}$$

Figure 13.5 shows the distance as a function of the SINR according to (13.27). For illustration, the figure shows the downlink SINR as a function of the distance, and the assumptions are identical to those in the link budget in Table 13.1.

Figure 13.5 and (13.27) are useful in that they allow a system designer to quickly calculate the coverage of a cell. Given a required level of SINR, one can compute the maximum distance within which the required level of SINR can be met. All other users outside that maximum distance will experience SINRs that are less than the required level of SINR.

13.3.4 System Design Issues

Operationally, one indication that a mobile is in a difficult RF environment is the transition to a more robust burst profile. For example, the base station makes the decision on which burst profile to use on the downlink based on the downlink signal quality. However, to reduce the amount of feedback traffic on the uplink, the mobile assists in the determination of the downlink burst profile. The mobile can do so by measuring and monitoring the downlink CINR and comparing it to a range of CINR values. If the measured downlink CINR goes outside the range, then the mobile sends a RNG-REQ message (in the initial ranging interval) to request a change in the downlink burst profile [7].[2] By logging the transitions to more robust

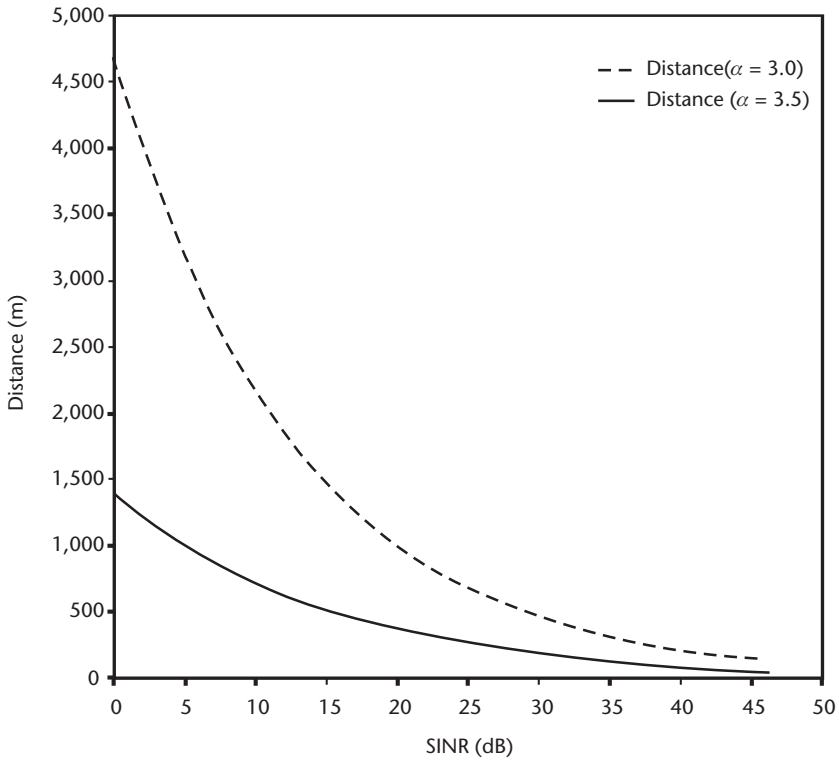


Figure 13.5 The distance as a function of the downlink SINR. The assumptions are identical to those in the link budget in Table 13.1. The operating frequency is 700 MHz, and the interference margin is kept constant as a function of the distance.

2. The mobile uses the RNG-REQ message to request a more robust downlink burst profile [7].

burst profiles and correlating them geographically, an RF engineer may determine which area has particularly difficult coverage.

On the other hand, a larger path loss (i.e., a larger α) may actually help isolate a cell from interference from other cells, especially in interference-limited systems. Because of the exponential loss coefficient, transmission rapidly loses power as a function of distance, and interference transmissions from other cells far away are more severely attenuated than the signal transmission from the home cell nearby. To see the effect, consider the expression for SIR in (13.6):

$$SIR^{0,j} = \frac{S^{0,j}}{I^{0,j}} = \frac{P_T^{0,j} L_p^0 u_g^0 u_l^0 G_R^0}{\sum_{i=1}^{M} P_T^{i,j} L_p^i u_g^i u_l^i G_R^i}$$

As α goes up, the losses L_p^i for the M interferers in the denominator become more severe (than the loss L_p^0 for user 0 in the numerator) because the interferers are presumably farther away than user 0. If the losses L_p^i in the denominator are more severe, then interference terms in the denominator become smaller, and SIR for user 0 becomes larger.

A special example of how a larger path loss helps isolate a cell from other cells' interference is a *femtocell*. A network access provider can deploy femtocells inside homes and businesses to improve indoor coverage (and to provide additional capacity). An existing broadband connection such as digital subscriber line (DSL), cable modem, or fiber can backhaul the traffic. As such, these femtocells are overlaid inside existing macrocells. For an indoor femtocell, walls incur large path loss and help reduce interference between the femtocell and other cochannel macrocells [8].

13.3.5 System Modeling Issues

As mentioned earlier, the SINR expression in (13.4) is for each subcarrier, but an OFDMA system utilizes an ensemble of subcarriers. Thus, there exists the need for a method for aggregating individual subcarriers' SINRs. Simply averaging the SINRs of individual subcarriers is not sufficient because of several reasons, including bits are spread across the subcarriers, and each subcarrier can experience a different frequency-selective fading [9].

One popular method for aggregating individual subcarriers' SINRs and calculating an overall effective SINR is the exponential effective SIR mapping (EESM) method, which is defined as:

$$SINR_{eff} = -\beta \ln\left(\frac{1}{N}\sum_{i=1}^{N} e^{-\frac{SINR_i}{\beta}}\right) \tag{13.28}$$

where $SINR_i$ is the SINR of subcarrier i and β is an adjustment factor that depends on the modulation and coding scheme used and other modem-specific characteristics. A table of β values can be found in [9]. As shown in (13.28), the EESM method is essentially a compression method that aggregates multiple SINRs into a single

effective $SINR_{eff}$. What is important is that the resulting effective $SINR_{eff}$ is the equivalent SINR for an additive white Gaussian noise (AWGN) channel, which allows the use of AWGN assumptions to assess the system performance [10].

In modeling an IEEE 802.16-based system, one typically calculates the $SINR_{eff}$ for an FEC block [5]. An FEC block consists of a number of subchannels, and a subchannel consists of multiple subcarriers depending on the subcarrier permutation mode used (see Chapter 5).

13.3.6 Concluding Remarks

Thus far, this chapter has emphasized the fact that adequate received power is important for the successful demodulation and recovery of transmitted bits. While that is certainly true, there are two more factors that merit consideration, especially in mobile broadband wireless systems. In effect, designing a mobile broadband wireless system requires a different paradigm—not only checking that the actual SINR is enough to support a desired bit rate and probability of bit error, but also ensuring that the system is robust enough to withstand the inevitable delay spread and Doppler shift. The next sections address these considerations in the design of an OFDMA system.

13.4 Designing for Temporal and Frequency Dispersions

13.4.1 Time Dispersion

Because the data symbol rates in a broadband wireless system are high, the data symbol durations are short, and short data symbols are more susceptible to a given delay spread. Thus, designers need to consider the effect of delay spread in high-speed wireless systems. One reason that OFDM is a popular technology of choice for wireless broadband is its ability to counter the effect of delay spread. By dividing one wide channel into many narrow subcarriers, OFDM lengthens the duration of each data symbol and makes each data symbol more robust against delay spread. Such robustness is achieved when the data symbol duration is much larger than the delay spread. The following example illustrates this.

Example 13.1

System performance engineers took channel-sounding measurements and determined that the propagation environment in the area of a proposed base station has an rms delay spread of 10 μs. What should be the width of the carrier (or subcarrier in the case of OFDM)?

The coherence bandwidth as a function of rms delay spread is (see Chapter 2):

$$W_c \approx \frac{1}{5\tau_{RMS}} \tag{13.29}$$

For an rms delay spread of 10 μs, the coherence bandwidth is:

$$W_c \approx \frac{1}{5(10\ \mu\text{s})} = 20\ \text{KHz}$$

So for an rms delay spread of 10 μs, the fading environment can be considered flat over a 20-kHz bandwidth [11]. In general, the higher the delay spread, the smaller the (sub)carrier width needs to be.

Two popular channel bandwidth/N_{FFT} combinations in IEEE 802.16 are 10 MHz/1,024 and 5 MHz/512. The subcarrier spacing turns out to be 10.9375 kHz in both combinations. Continuing with this example, one can compute the delay spread that can be tolerated given a subcarrier width of 10.9375 kHz: If the coherence bandwidth is 10.9375 kHz, then the rms delay spread is 1/5(10.9375 kHz) = 18.285 μs. A subcarrier width of 10.9375 kHz is sufficient to overcome the effect of an 18.285-μs rms delay spread. Incidentally, 18.285 μs corresponds to a path difference of about 5.5 km.

In a broadband wireless system, it is difficult to correct the problem of ISI by just increasing power. One indication of a possible problem with delay spread may be that for a *fixed* user, the error rate on the link remains elevated even though the SINR is sufficient (and the link budget shows plenty of margin). If delay spread is a problem and subcarrier spacing cannot be changed (due to the technology choice already made), one alternative is to reduce the size of the cell. It is likely that in a smaller cell, the path difference (and the delay spread) in the cell would be smaller.

A narrow subcarrier width reduces the effect of delay spread on adjacent data symbols (i.e., delayed versions of one data symbol interfering with a subsequent data symbol) within an OFDM symbol. In addition to the effect of delay spread on adjacent *data symbols*, a designer also needs to consider the effect of delay spread on adjacent OFDM symbols. Recall from Chapter 4 that, in a generic OFDM system, the IDFT function transforms the group of K parallel data symbols from the frequency domain into the time domain. The K parallel transformed symbols pass through a parallel-to-serial converter, and the resulting group of K transformed symbols in a series is called an OFDM symbol.

For a given delay spread, sufficient guard time must be inserted between consecutive OFDM symbols to allow for such a delay spread. The following example illustrates this.

Example 13.2

System performance engineers took channel-sounding measurements and determined that the propagation environment in the area of another proposed base station has a maximum delay spread of 4 μs. Assume that the guard time between OFDM symbols needs to be at least the same as the maximum delay spread. What is the required ratio of guard time to (useful) OFDM symbol time for a channel bandwidth/N_{FFT} combination of 10 MHz/1,024? What is the required ratio for 5 MHz/512?

For the channel bandwidth/N_{FFT} combination of 10 MHz/1,024, each transmission symbol lasts 1/10 MHz or 0.1 μs, and each OFDM symbol lasts 0.1 μs $\times$ 1,024 or 102.4 μs. Thus, the required ratio of guard time to (useful) OFDM symbol time is 4 μs/102.4 μs or 1/26.

For the channel bandwidth/N_{FFT} combination of 5 MHz/512, each transmission symbol lasts 1/5 MHz or 0.2 μs, and each OFDM symbol lasts 0.2 μs × 512 or again 102.4 μs. Thus, the required ratio of guard time to (useful) OFDM symbol time is 4 μs/102.4 μs or again 1/26.

The IEEE 802.16e-based system permits four different ratios of guard time to (useful) OFDM symbol time: 1/4, 1/8, 1/16, and 1/32 (i.e., cyclic prefix). If the required ratio is 1/26, then the designer needs to set the cyclic prefix at 1/16, as 1/32 would not be enough. In other words, one has to make sure that the guard time is at least equal to or greater than the delay spread.

To sum up, in designing a broadband wireless system, a designer would not only consider attenuation (e.g., by using tools such as the link budget), but also take into account the delay spread (e.g., by configuring the system) in the coverage area.

13.4.2 Frequency Dispersion

In addition to the time dispersion, OFDM-based systems, especially mobile broadband systems, suffer from another type of impairment. This impairment is frequency shift due to the relative velocity between the transmitter and the receiver (as well as due to any scattering objects that are moving). To see this impairment, consider Figure 13.6, which depicts the OFDM signal in the frequency domain.

The center frequency of each subcarrier is $f_c \pm jR_s$. Observe that at the peak of each subcarrier, interference contributions from neighboring subcarriers are zero. In OFDM, such alignment of subcarriers is what contributes to the orthogonality of subcarriers. Specifically, adjacent subcarriers have to be separated by $\Delta f = R_s$ to maintain orthogonality.

If there is relative motion between the transmitter and the receiver (or if there is a moving scatterer between them), then the OFDM signal will experience a Doppler shift f_D.

$$f_D = v\frac{f}{c} \tag{13.30}$$

(assuming that the shift is maximum positive) where v is the relative velocity between the transmitter and the receiver, f is the frequency, and c is the speed of light. If there is relative motion between the transmitter and the receiver, a subcarrier with the subcarrier frequency f received at the receiver is shifted to $f + f_D$.

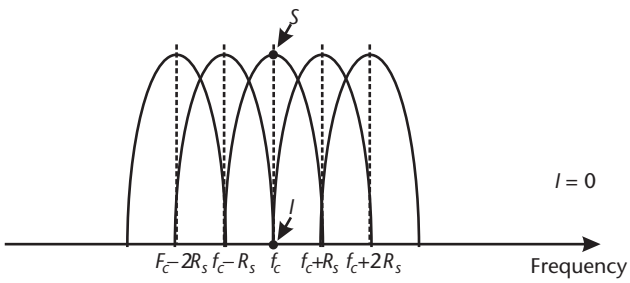


Figure 13.6 The OFDM signal. f_c is the carrier frequency, R_s is the symbol rate, and I is the interference from the adjacent subcarrier.

Without a Doppler shift, the center frequency of each subcarrier is $f = f_c \pm jR_s$ (see Figure 13.6). With a Doppler shift, the center frequency of each subcarrier is now $f' = f_c \pm jR_s + f_D$. By substituting the expression for f_D, one obtains the center frequency f' of each subcarrier (with a Doppler shift):

$$\begin{aligned} f' &= f_c \pm jR_s + f_D = f_c \pm jR_s + v\frac{f}{c} = f_c \pm jR_s + v\frac{(f_c \pm jR_s)}{c} \\ &= (f_c \pm jR_s)\left(1+\frac{v}{c}\right) = f\left(1+\frac{v}{c}\right) \end{aligned} \tag{13.31}$$

Therefore, with a Doppler shift, the frequency f of each subcarrier is multiplied by the same factor $(1 + v/c)$ [12]. Applying this factor to the subcarriers produces the OFDM signal shown in Figure 13.7.

Observe that adjacent subcarriers are no longer separated by $\Delta f = R_s$. With a Doppler shift, adjacent subcarriers are now separated by $\Delta f = R_s(1 + v/c)$, that is,

$$\begin{aligned} \Delta f &= \left[f_c + (j+1)R_s\right]\left(1+\frac{v}{c}\right) - (f_c + jR_s)\left(1+\frac{v}{c}\right) \\ &= \left[f_c + (j+1)R_s - f_c - jR_s\right]\left(1+\frac{v}{c}\right) = R_s\left(1+\frac{v}{c}\right) \end{aligned} \tag{13.32}$$

Because adjacent subcarriers are no longer separated by $\Delta f = R_s$, interference contributions from neighboring subcarriers are no longer zero at the peak of each subcarrier. In other words, the subcarriers are no longer orthogonal. As a result, each subcarrier's SINR is degraded. This effect is known as intercarrier interference (ICI).

To quantify the degradation in SINR due to ICI, one has to evaluate the expression

$$SINR = \frac{S}{N+I} \tag{13.33}$$

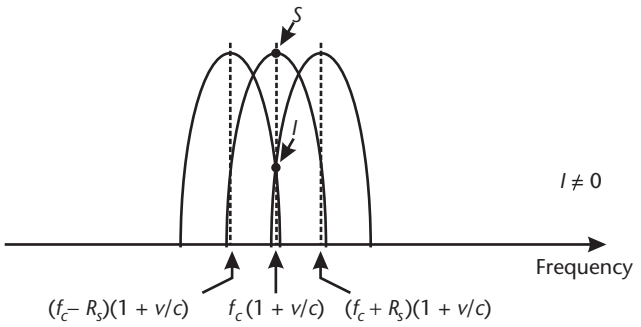


Figure 13.7 The OFDM signal experiencing Doppler shift. v is assumed to be negative. I is interference from the adjacent subcarrier.

where I is due to interference contributions from neighboring subcarriers. For a particular subcarrier j, calculating I (due to ICI) involves summing up the interference contributions from all other subcarriers m where $m \neq j$.

There have been research articles (e.g., [13–16]) that examined ICI, many of which make certain assumptions to simplify the calculation of SINR degradation. One such approximation is given by (in decibels):

$$\Delta SINR \cong -\frac{10}{\ln 10}\frac{1}{3}\left(\pi\frac{f_{shift}}{R_s}\right)^2 SNR \tag{13.34}$$

where f_{shift} is the frequency shift (e.g., caused by relative motion). The assumption used is that there are an infinite number of subcarriers; this approximation overestimates the ICI by a factor of 2 but is more accurate for subcarriers not near the edge of the band [15].

Example 13.3

A base station is located next to the freeway and has an omnidirectional transmit antenna. What is the SINR degradation experienced by a receiver in a car moving toward the base station at 90 km/hour? The receiver also has an omnidirectional receive antenna. The system is based on IEEE 802.16e and has a frequency spacing of 10.9375 kHz. Assume that the SNR is 10 dB (or 10). The system operates at 3.5 GHz.

$$v = 90\frac{\text{km}}{\text{hour}} = 90\frac{\text{km}}{\text{hour}} \times \left(\frac{1000\text{m}}{\text{km}}\right) \times \left(\frac{\text{hour}}{3600\text{s}}\right) = 25\frac{\text{m}}{\text{s}}$$

$$f_{shift} = f_D = v\frac{f}{c} = (25\ \text{m/s})\left(\frac{3.5\times 10^9\ \text{Hz}}{3\times 10^8\ \text{m/s}}\right) = 291.67\ \text{Hz}$$

$$\Delta SINR \cong -\frac{10}{\ln 10}\frac{1}{3}\left(\pi\frac{f_{shift}}{R_s}\right)^2 SNR = -\frac{10}{\ln 10}\frac{1}{3}\left(\pi\frac{291.67\ \text{Hz}}{10937.5\ \text{Hz}}\right)^2 10 = -0.1\ \text{dB}$$

Example 13.4

What is the SINR degradation experienced by the receiver in the same car if the same system operates at 700 MHz?

$$f_{shift} = f_D = v\frac{f}{c} = (25\,\text{m/s})\left(\frac{700\times 10^6\ \text{Hz}}{3\times 10^8\ \text{m/s}}\right) = 58.33\ \text{Hz}$$

$$\Delta SINR \cong -\frac{10}{\ln 10}\frac{1}{3}\left(\pi\frac{f_{shift}}{R_s}\right)^2 SNR = -\frac{10}{\ln 10}\frac{1}{3}\left(\pi\frac{58.33\ \text{Hz}}{10937.5\ \text{Hz}}\right)^2 10 = -0.004\ \text{dB}$$

Therefore, the SINR degradation due to relative motion is less severe at lower frequencies.

Note that frequency shifts of subcarriers can be caused not only by Doppler shift, but also by local oscillator (LO) drift and frequency synchronization error[3] at the receiver. It is important to assess the SINR degradation due to frequency shift by using models such as (13.34). By computing and anticipating the SINR degradation, the designer can incorporate sufficient margin into the system.

For illustration, Figure 13.8 shows the SINR degradation as a function of the ratio between f_{shift} and R_s.

Figure 13.8 uses the required SINR as the SNR in (13.34) to generate the SINR degradation. The required SINRs are shown in Table 13.2 for different modulation and coding schemes. The assumption here is that the system uses power control to maintain the link SINR right at the required SINR to minimize interference to other users in nearby cochannel cells/sectors.

In OFDM, the SINR degradation becomes more severe as the frequency shift f_{shift} grows to become a larger percentage of the subcarrier spacing R_s. In addition,

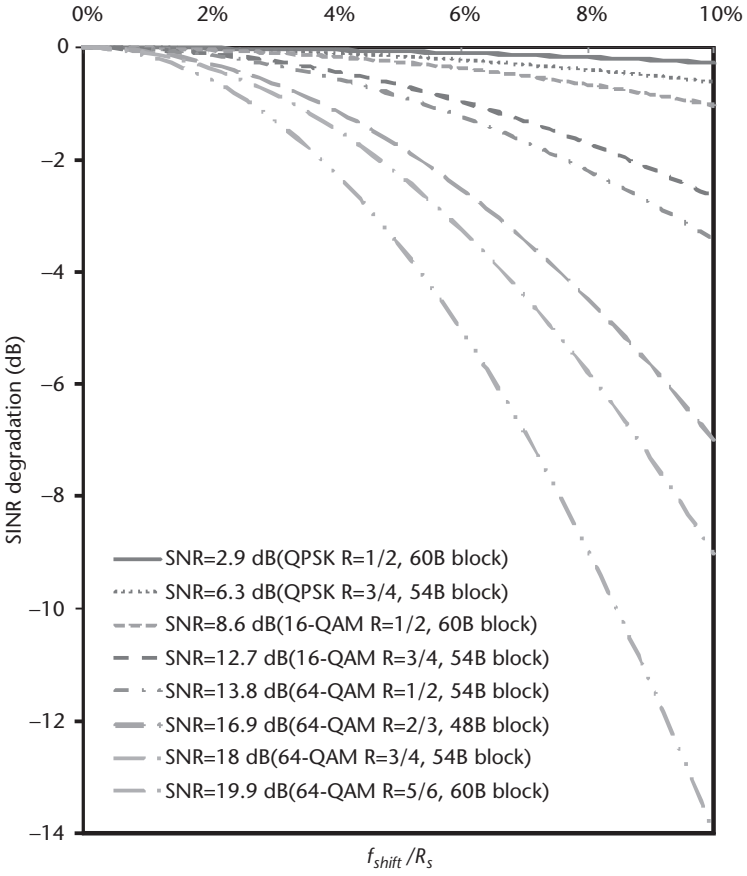


Figure 13.8 The SINR degradation as a function of the ratio between f_{shift} and R_s.

3. For example, the IEEE 802.16e standard specifies a maximum (uplink) frequency synchronization error of 2% of subcarrier spacing [7].

Table 13.2 Required SINRs for Different Modulation and Coding Schemes [6]

Modulation and Coding Schemes	*Required SINR**
QPSK, Rate 1/2 (60-byte block)	2.9 dB
QPSK, Rate 3/4 (54-byte block)	6.3 dB
16-QAM, Rate 1/2 (60-byte block)	8.6 dB
16-QAM, Rate 3/4 (54-byte block)	12.7 dB
64-QAM, Rate 1/2 (54-byte block)	13.8 dB
64-QAM, Rate 2/3 (48-byte block)	16.9 dB
64-QAM, Rate 3/4 (54-byte block)	18 dB
64-QAM, Rate 5/6 (60-byte block)	19.9 dB

* CTC for $P_b = 10^{-6}$ in AWGN channel.

higher-order modulation (e.g., QAM) suffers more from frequency shift than lower-order, more robust modulation (e.g., QPSK). This is because a higher-order modulation needs a higher required SINR.

Moreover, a time-varying channel affects the required rate of channel feedback for estimation and equalization. At 3.5 GHz, the channel coherence time T_c is

$$T_c \approx \frac{1}{2f_D} = \frac{1}{583.33\text{Hz}} = 0.00171\text{s} = 1.71\text{ ms}$$

using the parameters in Example 13.3. The minimum rate of channel feedback has to be at least once every T_c, or $1/T_c$.

$$\frac{1}{T_c} \approx 2f_D = 583.33\text{ Hz}$$

At 700 MHz, the minimum rate of channel feedback has to be at least once every T_c, or $1/T_c$.

$$\frac{1}{T_c} \approx 2f_D = 116.67\text{ Hz}$$

Therefore, we see that systems operating at lower carrier frequencies require a lower rate of channel feedback.

13.4.3 Concluding Remarks

By now, readers can see that there is a tradeoff between designing for delay spread and designing for frequency shift. A fundamental motivation behind OFDM is that a narrower subcarrier results in a longer symbol duration, and a longer symbol duration provides more resilience against delay spread. However, a narrow subcarrier

is more susceptible to a given frequency shift, f_{shift}. This is because as the subcarrier spacing R_s becomes small, the percentage f_{shift}/R_s becomes large for a given frequency shift, f_{shift}. If f_{shift}/R_s increases in (13.34), the SINR degradation worsens.

Therefore, deciding on a subcarrier spacing (hence subcarrier bandwidth) involves a tradeoff between the delay spread and the frequency shift. The subcarrier bandwidth is chosen to be small enough to counter delay spread, but at the same time, large enough to tolerate some frequency shift (due to Doppler). Because of this tradeoff, [17] has proposed adaptive subcarrier bandwidth in OFDM. The scheme calls for dynamically changing the number of subcarriers in each OFDM symbol (hence changing the subcarrier bandwidth in each time slot) based on channel conditions.

All in all, it is challenging to design a protocol that works well in all environments (local-area versus wide-area), frequencies (low versus high), and mobility (fixed versus mobile). The propagation environment affects delay spread, and frequency and mobility affect Doppler shift. It is for this reason that handover between heterogeneous networks (e.g., wireless LAN and cellular), executed by the network layer and/or layers above [18, 19] and by cognitive radios [20] that can reconfigure to latch onto the closest or the most economical base station/access point will also be important in next-generation broadband wireless networks.

References

[1] Laroia, R., S. Uppala, and J. Li, "Designing a Mobile Broadband Wireless Access Network," *IEEE Signal Processing*, Vol. 21, No. 5, 2004, pp. 20–28.

[2] Upase, B., M. Hunukumbure, and S. Vadgama, "Radio Network Dimensioning and Planning for WiMAX Networks," *Fujitsu Scientific and Technical Journal*, Vol. 43, No. 4, 2007, pp. 435–450.

[3] Yang, S. C., *CDMA RF System Engineering*, Norwood, MA: Artech House, 1998.

[4] Yang, S. C., *3G CDMA2000 Wireless System Engineering*, Norwood, MA: Artech House, 2004.

[5] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

[6] WiMAX Forum, "WiMAX Forum Mobile System Profile Release 1.0 Approved Specification," Revision 1.7.1, November 7, 2008.

[7] IEEE Standard 802.16e, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," New York: IEEE, February 28, 2006.

[8] Yeh, S. -P., et al., "WiMAX Femtocells: A Perspective on Network Architecture, Capacity, and Coverage," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 58–65.

[9] Jain, R., C. So-In, and A. -K. Al Tamimi, "System-Level Modeling of IEEE 802.16e Mobile WiMAX Networks: Key Issues," *IEEE Wireless Communications*, Vol. 15, No. 5, 2008, pp. 73–79.

[10] Yaniv, R., et al., "CINR Measurements Using the EESM Method," IEEE 802.16 Broadband Wireless Access Working Group, 2005.

[11] Yaghoobi, H., "Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN," *Intel Technology Journal*, Vol. 8, No. 3, 2004, pp. 201–212.

[12] Xiong, F., and M. Andro, "The Effect of Doppler Frequency Shift, Frequency Offset of the Local Oscillators, and Phase Noise on the Performance of Coherent OFDM Receivers," NASA/TM—2001-210595, National Aeronautics and Space Administration (NASA) Glenn Research Center, March 2001.

[13] Robertson, P., and S. Kaiser, "The Effects of Doppler Spreads in OFDM(A) Mobile Radio Systems," *Proceedings of IEEE 50th Vehicular Technology Conference*, Vol. 1, September 1999, pp. 329–333.

[14] Moose, P. H., "A Technique for Orthogonal Frequency Division Multiplexing Frequency Offset Correction," *IEEE Trans. on Communications*, Vol. 42, No. 10, 1994, pp. 2908–2914.

[15] Pollet, T., M. Van Bladel, and M. Moeneclaey, "BER Sensitivity of OFDM Systems to Carrier Frequency Offset and Wiener Phase Noise," *IEEE Trans. on Communications*, Vol. 43, No. 2/3/4, 1995, pp. 191–193.

[16] Wang, T. (R.), J. G. Proakis, and E. Masry, "Performance Degradation of OFDM Systems Due to Doppler Spreading," *IEEE Trans. on Wireless Communications*, Vol. 5, No. 6, 2006, pp. 1422–1432.

[17] Das, S. S., E. D. Carvalho, and R. Prasad, "Performance Analysis of OFDM Systems with Adaptive Sub Carrier Bandwidth," *IEEE Trans. on Wireless Communications*, Vol. 7, No. 4, 2008, pp. 1117–1122.

[18] Chiron, P., et al., "Architectures for IP-Based Network-Assisted Mobility Management Across Heterogeneous Networks," *IEEE Wireless Communications*, Vol. 15, No. 2, 2008, pp. 18–25.

[19] Eastwood, L., et al., "Mobility Using IEEE 802.21 in a Heterogeneous IEEE 802.16/802.11-Based, IMT-Advanced (4G) Network," *IEEE Wireless Communications*, Vol. 15, No. 2, 2008, pp. 26–34.

[20] Mitola, J., III, and G. Q. Maguire, Jr., "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Communications*, Vol. 6, No. 4, 1999, pp. 13–18.

CHAPTER 14

RF System Design: Capacity

14.1 Introduction

An RF system designer not only considers coverage, but also takes into account the capacity of the system. The number of mobile broadband users is expected to rise as more users become accustomed to being connected anywhere, anytime. As demand rises, capacity must also rise to meet demand. Also, wireless bandwidth is a precious commodity, and mobility makes achieving high spectral efficiency a challenge. These factors make the RF link the likely bottleneck in an end-to-end broadband mobile wireless system, and the improvement of the RF link performance can directly translate into an overall improvement of the end-to-end system [1]. Thus, RF system design for capacity is crucial for broadband mobile wireless systems. Effective design enables a system to serve the demand in a given area with fewer base stations. This way, a network access provider may meet demand with less invested capital.

In addition to discussing system capacity, this chapter presents design examples to illustrate salient system design issues. To keep the examples concise, we do not include the cases for MIMO, which would increase the raw bit rates.

14.2 Frequency Reuse

14.2.1 Fundamental Concepts

Readers are no doubt familiar with the concept of frequency reuse. In frequency reuse, a system utilizes the same blocks of frequencies more than once (i.e., reuses the frequencies). Suppose a network access provider is allocated a portion of spectrum, J, and wishes to serve an area using that spectrum. As an example, the access provider has two options. In option 1, it serves the area using the entire spectrum J, presumably using a single, tall base station at the center of the cell (see Figure 14.1). As a result, all mobiles in the area have access to the spectrum J.

In option 2, the access provider divides its spectrum J into seven blocks; each block is $J/7$ wide. At the same time, it divides the area into many cells (e.g., 49 cells) and assigns each cell a different frequency block. The access provider assigns its different blocks of frequencies in such a way that cells using identical frequency

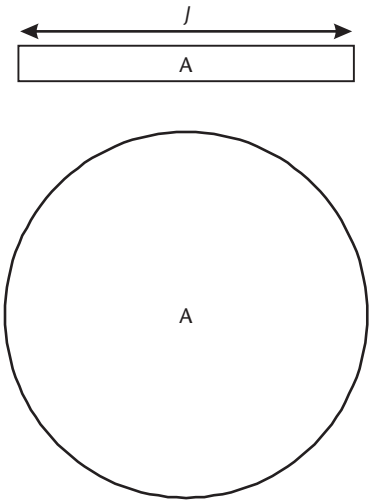


Figure 14.1 The amount of spectrum in use in the area is *J* (for option 1).

blocks (i.e., cochannel cells) are as geographically separated from each other as possible (see Figure 14.2). To implement these many smaller cells, the access provider uses base stations with antennas that are shorter. As a result, mobiles in different cells access different blocks of frequencies depending on the cells in which they operate. As Figure 14.2 shows, there are a total of $49(J/7) = 7J$ of spectrum that are used to serve the area.

Obviously, option 2 results in more frequencies available in a given area than option 1. The reason why there are more frequencies available is because

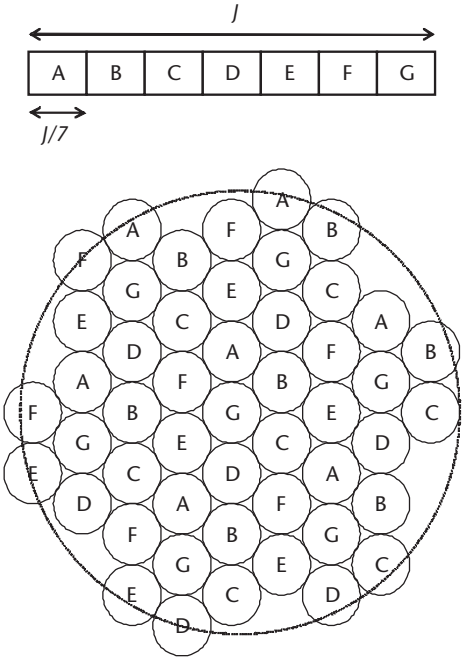


Figure 14.2 The amount of spectrum in use in the area is $7J$ (for option 2). The frequency reuse factor is $N = 7$.

frequencies are reused, and the reason why frequencies can be reused is because cochannel cells are nonadjacent and geographically separated. In fact, in frequency reuse, path loss actually helps improve the SIR of a link. If path loss is high, a cochannel transmission loses more power as it traverses a given distance; by the time the cochannel transmission reaches the nonadjacent, cochannel cell, the power is low compared to the resident transmissions used in that cochannel cell. Cochannel interfering transmissions become weak when they propagate over some distance to another cochannel cell.

14.2.2 Frequency Reuse Factors

The second option just discussed uses a (intercell) frequency reuse factor of $N = 7$ by which frequency blocks are assigned to a *cluster* of seven cells. A cluster is a group of cells that use once and only once all available frequencies of an access provider [2]. Each cell in a cluster uses a different frequency block; thus, the number of cells in a cluster corresponds to the frequency reuse factor N. In general, the assignment of frequency blocks to cells (arranged in a hexagonal layout) can be done according to the following:

$$N = i^2 + ij + j^2 \quad \text{where } i \geq 1, j \geq 0 \tag{14.1}$$

In this equation, i and j are translation distance and rotation angle, respectively. For $N = 7$, $i = 2$ and $j = 1$. What that means is that if one has a cochannel cell and would like to assign another cochannel cell, then from the first cochannel cell, one:

- Moves linearly or translates i ($i = 2$) cells over;
- Turns or rotates by $60j$ ($60j = 60$) degrees.

This is exactly what was done for the $N = 7$ reuse factor (see Figure 14.3). Incidentally, in two dimensions the hexagonal layout provides the densest packing of circles.

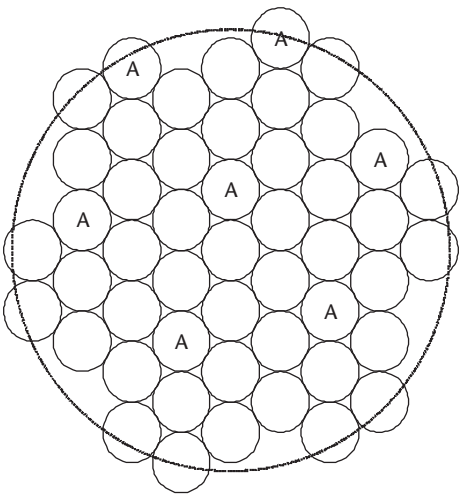


Figure 14.3 Assigning cochannel cells for the $N = 7$ reuse factor.

There are other frequency reuse factors possible besides $N = 7$. A "tighter" frequency reuse factor that is lower than $N = 7$ and that also satisfies (14.1) is $N = 4$. For $N = 4$, the access provider divides its spectrum J into four blocks; each block is $J/4$ wide. Then the access provider assigns each cell a different frequency block. In assigning the different frequency blocks to cells and separating the cochannel cells, the access provider again uses (14.1). For $N = 4$, $i = 2$ and $j = 0$. In assigning cochannel cells, the access provider moves linearly or translates two ($i = 2$) cells over, and turns or rotates by 0 ($60j = 0$) degrees. Figure 14.4 depicts the assignment of cells for the $N = 4$ reuse factor. Since there are 49 cells in the figure and each cell is assigned $J/4$ of spectrum, the total amount of spectrum used to serve the area is $49(J/4)$ or $12.25J$.

Note that a tighter reuse results in more frequencies available to serve a given area. In the above examples, one obtains a total of $12.25J$ of spectrum using the $N = 4$ reuse factor, but one obtains only $7J$ of spectrum using the $N = 7$ reuse factor. Given that a tighter frequency reuse results in more frequencies available for the system, why would we not always adopt a tighter reuse?

The reason is that the tighter the reuse, the closer the cochannel cells are from each other. Because cochannel cells use the same block of frequencies, the cochannel interference would be higher, causing the received SIR experienced to be lower. Let D be equal to the distance between cochannel cells and r be equal to the radius of a cell. In the examples shown above, cells are of an equal size, so r is a constant. However, when one compares the $N = 7$ reuse (Figure 14.2) and the $N = 4$ reuse (Figure 14.4), one can readily tell that D is smaller for the tighter reuse of $N = 4$. Figure 14.5 shows the D and the r for the two reuses $N = 7$ and $N = 4$.

There is yet another tighter reuse, $N = 3$. For $N = 3$, the access provider divides its spectrum J into three blocks; each block is $J/3$ wide. Then it assigns each cell a

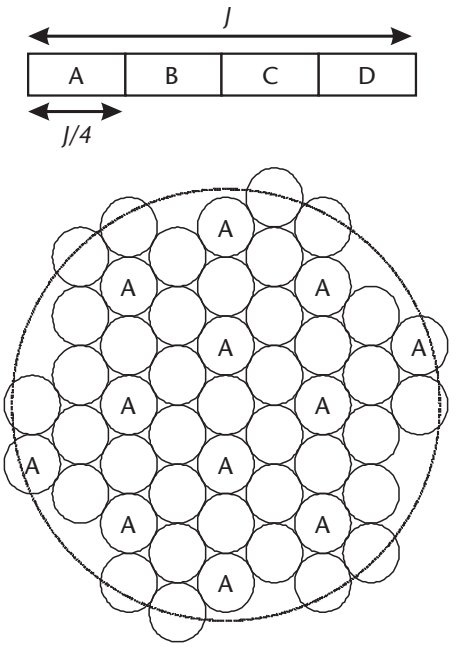


Figure 14.4 The amount of spectrum in use in the area is $12.25J$. The frequency reuse factor is $N = 4$.

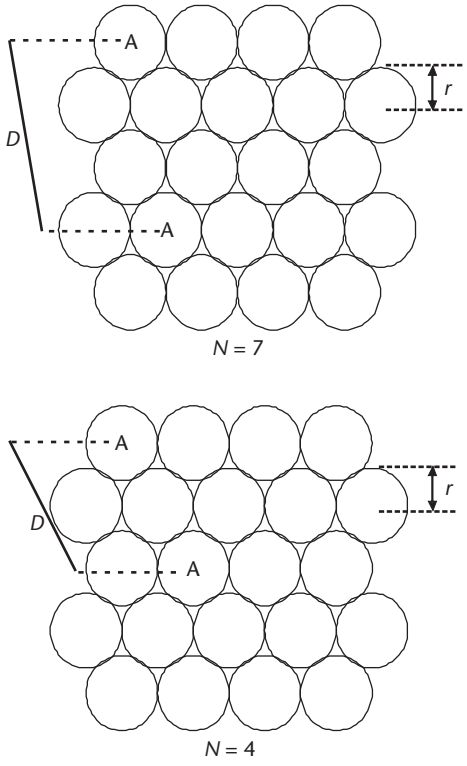


Figure 14.5 A comparison of cochannel distance *D* between *N* = 7 and *N* = 4. *r* is identical for both reuses.

different frequency block. In assigning the different frequency blocks to cells, the access provider solves (14.1) with $N = 3$. For $N = 3$, $i = 1$ and $j = 1$. To assign cochannel cells, the access provider moves linearly or translates one ($i = 1$) cell over and turns or rotates by 60 ($60j = 60$) degrees. Figure 14.6 depicts the $N = 3$ reuse. Since there are 49 cells in the figure and each cell is assigned $J/3$ of spectrum, the total amount of spectrum used to serve the area is $49(J/3)$ or $16.33J$. As expected, with a tighter $N = 3$ reuse, the total amount of spectrum is higher, but the cochannel distance D is lower and thus the SIR would be lower.

The tightest frequency reuse factor is $N = 1$—also known as the universal frequency reuse. In $N = 1$, the same frequency is reused in every cell. So for $N = 1$, the access provider assigns its entire spectrum J to every cell, and each cell uses the same frequency. Figure 14.7 shows the frequency reuse scheme in which every cell in the area is using the same frequency. Since there are 49 cells in the figure and each cell is assigned spectrum J, the total amount of spectrum used to serve the area is $49J$.

Although $N = 1$ affords the highest amount of spectrum in use, the SIR would be very low. Typically, cells that are physically adjacent (next to each other) do not use the same frequencies at the same time.[1] Doing so would create a lot of interference, especially for a mobile situated at the edge of a cell. In fact, for a mobile on the edge of a cell, the downlink SIR experienced by it is approximately 1 or 0 dB

1. $N = 1$ is routinely used in cellular CDMA.

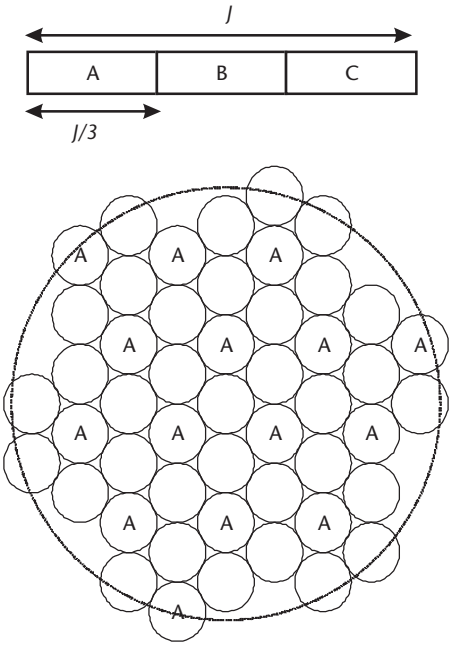


Figure 14.6 The amount of spectrum in use in the area is 16.33*J*. The frequency reuse factor is $N = 3$.

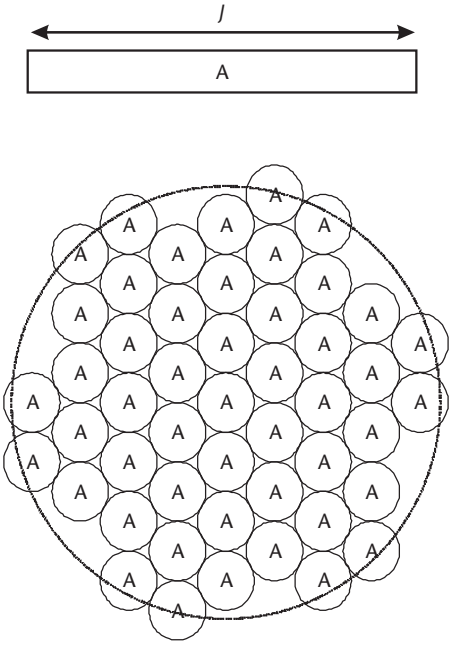


Figure 14.7 The amount of spectrum in use in the area is 49*J*. The frequency reuse factor is $N = 1$.

(assuming an equal downlink power received from each of two neighboring base stations) (see Figure 14.8).

Frequency reuse factors of 1, 3, 4, and 7 are often referred to as integer frequency reuse.

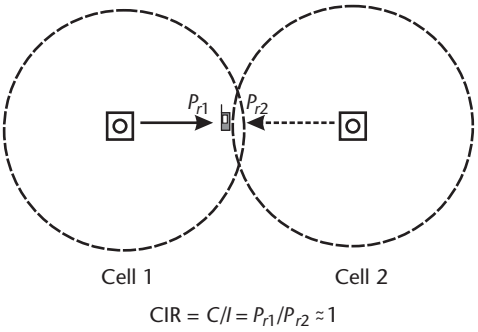


Figure 14.8 In $N = 1$, downlink SIR ≈ 1 for a mobile at the edge of a cell.

14.2.3 *D*/*r* Ratio

By using simple geometry, one can derive the parameter *D*/*r* ratio for a hexagonal arrangement of cells. Table 14.1 shows the *D*/*r* ratios for the frequency reuse factors just examined (i.e., $N = 1$, 3, 4, and 7), as well as for "looser" reuses (i.e., $N = 12$ and 19). These *D*/*r* ratios are applicable to cells of any size, as long as the following conditions are met:

1. Cells are of the same size (i.e., *r* is the same for all cells).
2. Cells are arranged in a hexagonal layout.
3. Cochannel cells are assigned per (14.1).

This relationship between reuse factor *N* and *D*/*r* is

$$\frac{D}{r} = \sqrt{3N} \tag{14.2}$$

Figure 14.9 shows the tradeoff involved with the choice of frequency reuse factor. The advantage of using a tighter reuse is that the total amount of frequencies in use increases. This is because as the frequency reuse becomes tighter, the cochannel distance decreases, and the same blocks of frequency are reused more often in the system. However, the disadvantage of using a tighter reuse is that the SIR experienced decreases (i.e., SIR decreases as *N* decreases). This is because as the reuse

Table 14.1 *D*/*r* Ratios for Different Frequency Reuse Factors

Reuse	*D*/*r*
$N = 1$	1.73
$N = 3$	3.00
$N = 4$	3.46
$N = 7$	4.58
$N = 12$	6.00
$N = 19$	7.55

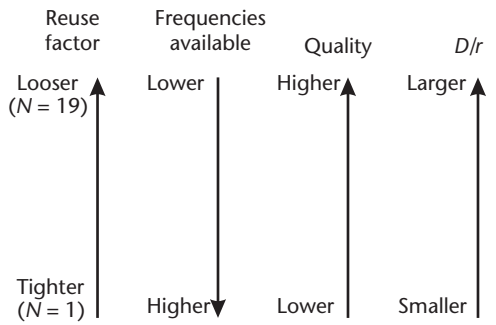


Figure 14.9 The tradeoff between quality and amount of frequencies available.

becomes tighter, the cochannel cells are closer to each other, causing cochannel interference to increase. Thus, the quality of the system goes down.

14.2.4 Frequency Reuse Patterns

Frequency planning is the task of assigning frequencies to cells such that users' SINRs meet the required SINR for a given service level and capacity is maximized. Effective frequency planning minimizes interference from nearby cochannel cells and maximizes signal from the home cell. In addition, implementing power control can also help reduce interference from nearby cochannel cells.

In performing frequency planning, system designers often choose from a set of frequency reuse patterns designated by the notation $N \times S \times K$, where

- N is the number of cells in a cluster, in which each cell uses a different frequency block. N is also known as the *intercell frequency reuse factor*.
- S is the number of sectors in a cell.
- K is the number of frequency blocks in a cell. K is also known as the *intracell frequency reuse factor* [3].

Given this notational convention, the $N = 3$ frequency reuse scheme previously shown in Figure 14.6 can be more fully described as the $3 \times 1 \times 1$ frequency reuse pattern. In the $3 \times 1 \times 1$ pattern, there are three cells in each cluster, one sector in each cell (no sectorization), and one frequency block in each cell. Figure 14.10(a) shows the $3 \times 1 \times 1$ pattern, whereas Figure 14.10(b) shows the $3 \times 3 \times 1$ pattern. Note that the $3 \times 3 \times 1$ pattern is different from the $3 \times 1 \times 1$ pattern because cells in the $3 \times 3 \times 1$ pattern are sectorized using directional antennas, which act as spatial filters to reduce cochannel interference.

The $N = 1$ universal frequency reuse scheme previously shown in Figure 14.7 is equivalent to the $1 \times 1 \times 1$ reuse pattern shown in Figure 14.11(a). In contrast, Figure 14.11(b) shows the $1 \times 3 \times 1$ reuse pattern, which uses directional antennas to realize three sectors in each cell, although all sectors use the same frequencies.

Figure 14.12(a) shows the $1 \times 3 \times 3$ reuse pattern, in which there is one cell in each cluster, three sectors in each cell, and three frequency blocks in each cell. Another reuse pattern that is looser is $3 \times 3 \times 3$ shown in Figure 14.12(b). In the $3 \times 3 \times 3$ pattern, there are three cells in each cluster, three sectors in each cell, and

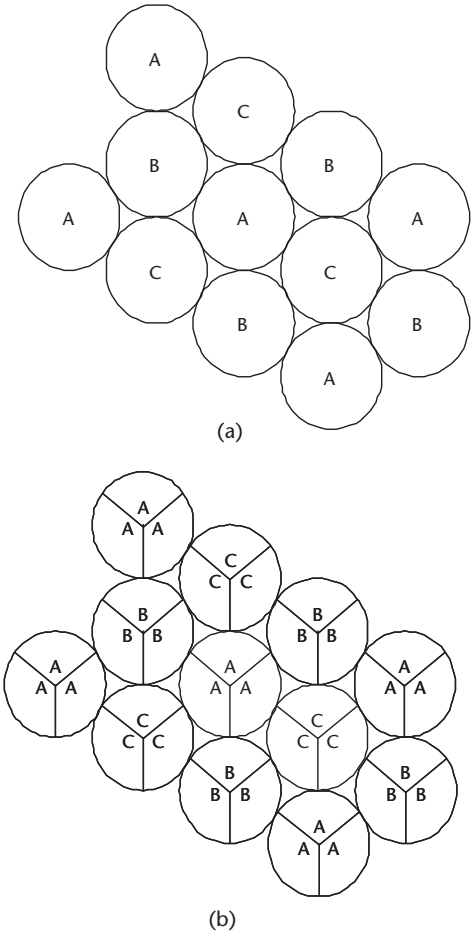


Figure 14.10 (a) $3 \times 1 \times 1$ frequency reuse pattern. (b) $3 \times 3 \times 1$ frequency reuse pattern.

three frequency blocks in each cell. As such, the $3 \times 3 \times 3$ pattern requires NK or 9 frequency blocks for assignment in each cluster.

14.2.5 Fractional Frequency Reuse

In addition to *integer* frequency reuse, an OFDMA-based system can also adopt *fractional* frequency reuse. Previously we mentioned that $N = 1$ affords the highest amount of spectrum in use, but the resulting SIR would be low, especially for mobiles at the edge of a cell. Fractional frequency reuse is a way of improving the SIRs for mobiles at cell edge. In fractional frequency reuse, mobiles near the center of the cell have access to all frequencies, while mobiles near the edge of a cell only have access to a selected frequency block (a fraction of all frequencies). As such, a cell is effectively divided into an *inner* region and an *outer* region. In the inner regions of cells, the system uses universal frequency reuse $N_{inner} = 1$. In the outer regions of cells, the system may use $N_{outer} = 3$ or greater. Figure 14.13 shows a system where $1 \times 1 \times 1$ is used in inner regions and $3 \times 1 \times 1$ is used in outer regions (a popular configuration). The three frequency blocks used in outer regions are labeled A, B, and C.

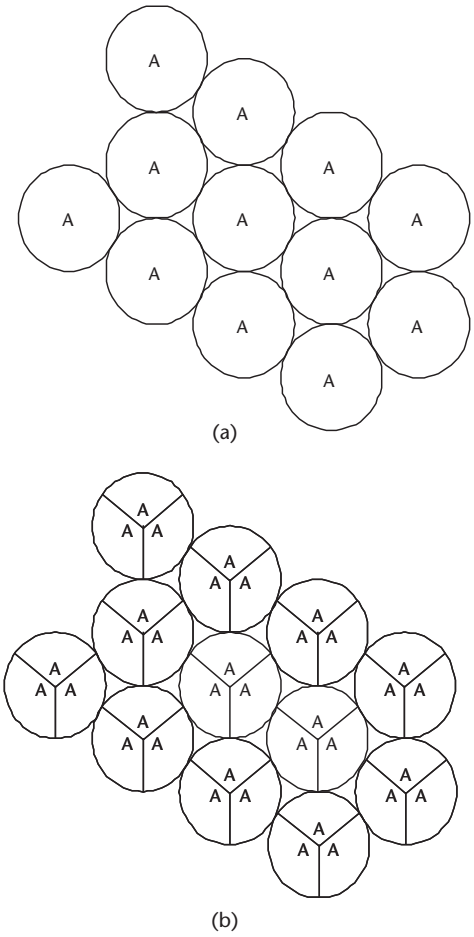


Figure 14.11 (a) $1 \times 1 \times 1$ frequency reuse pattern. (b) $1 \times 3 \times 1$ frequency reuse pattern.

Fractional frequency reuse results in a reuse factor that is a fractional number greater than 1. If φ is the ratio of the area of inner region to the total area, then the effective frequency reuse factor is

$$N = \phi N_{inner} + (1 - \phi) N_{outer} \tag{14.3}$$

In the example of fractional frequency reuse where inner regions have $N_{inner} = 1$ and outer regions have $N_{outer} = 3$, the effective N depends on φ. If $\varphi = 0$, N reverts to the original N_{outer} (= 3). As φ goes from 0 to 1, N goes down from 3 to 1 [4].

The purpose of fractional frequency reuse is to maximize the use of available frequencies for mobiles near the center of a cell while reducing interference and increasing SIR for mobiles near the edge of a cell, thus increasing bit rates for mobiles near the edge.

In fractional frequency reuse, the frequency blocks A, B, and C need not be static. In OFDMA, the number and location of subchannels constituting A, B, and C can be optimized dynamically on a frame-by-frame basis—across cells based on interference and loading; IEEE 802.16e facilitates this flexible assignment of subchannels over time through subchannel segmentation and permutation zones [2].

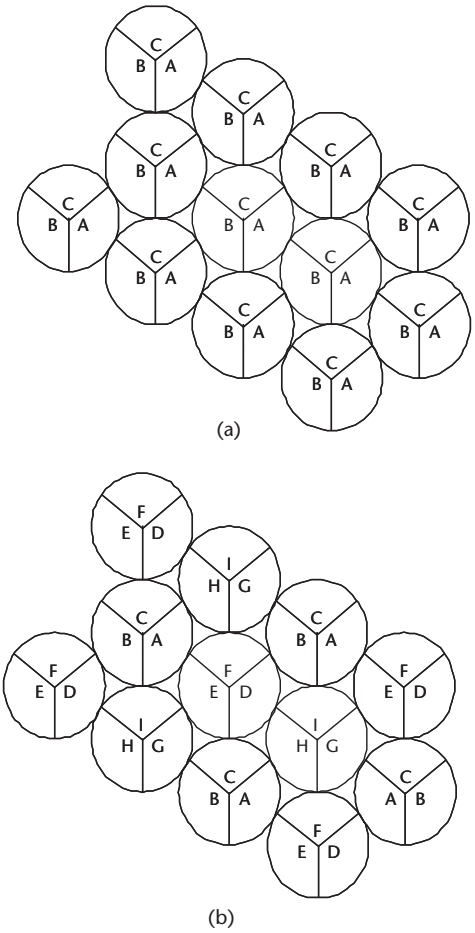


Figure 14.12 (a) $1 \times 3 \times 3$ frequency reuse pattern. (b) $3 \times 3 \times 3$ frequency reuse pattern.

For example, mobiles near the cell center can operate with $N = 1$ (i.e., all available subchannels) for part of the subframe (i.e., a zone), while mobiles near the cell edge can operate with $N = 3$ (i.e., a subset of subchannels) for another part of the subframe (i.e., another zone) [5].

Figure 14.14(a) shows a cluster of three cells (cells 1, 2, and 3) in a system that uses fractional frequency reuse. The inner region is defined by the radius r_4, and the outer region ranges from r_4 to cell radius r. Figure 14.14(b) depicts an example of a subframe used by these three cells. In first part of the subframe, transmissions are allocated to mobiles in the inner regions of the three cells. In the second part of the subframe, transmissions are allocated in frequency to mobiles in the outer regions of the three cells. Note that to avoid interference between zones, the system needs to synchronize every zone in every downlink/uplink frame in all base stations/sectors [4].

Fractional frequency reuse offers the advantage of reduced interference due to mobiles at cell edge. Mobiles at cell edge typically need higher power to communicate than mobiles near cell center. Thus, cochannel interference from mobiles at cell edge is high and reduces system capacity. By assigning mobiles near the cell center and mobiles near the cell edge to different parts of the subframe, fractional

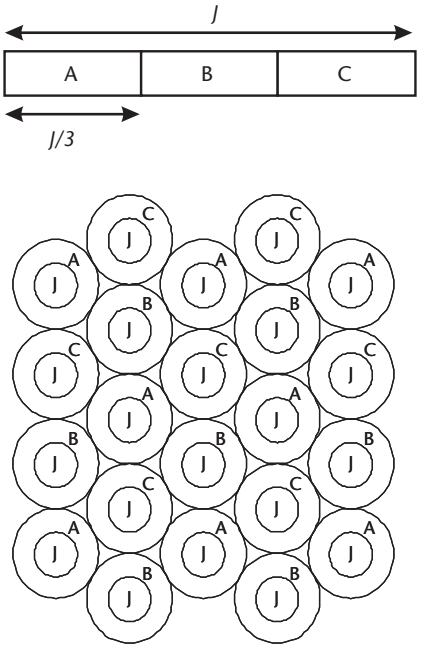


Figure 14.13 An example of fractional frequency reuse where inner regions use $1 \times 1 \times 1$ and outer regions use $3 \times 1 \times 1$.

frequency reuse effectively implements a time division of the available capacity between these two groups of mobiles, reduces cochannel interference from faraway mobiles, and increases system capacity [6].

For a sectorized configuration, Figure 14.15 shows an example of fractional frequency reuse where inner regions use $1 \times 3 \times 1$ and outer regions use $1 \times 3 \times 3$.

One study used simulation to assess the capacity and outage probability of a 19-cell system utilizing 10 MHz of bandwidth [4]; each cell has three sectors, and the path loss exponent is 3.5. It defines outage probability as the probability that the SINR falls below a minimum $SINR_{req}$, which is set at 6 dB. The study found that a frequency reuse factor of 1 is unacceptable because the SINR is too low, resulting in an outage probability that is greater than 10%. For fractional frequency reuse, as φ goes up, the outage probability increases. The optimal φ found is about 0.2, which results in the highest sector capacity and an outage probability of less than 10% [4].

14.3 Allocation of Capacity

4G systems have significantly shifted the way they allocate resources to users. Such a shift began to occur with 3G systems. As a mobile moves away from the base station, path loss increases. Figure 14.16 shows that in a classical 2G system employing power control (e.g., CDMA), the base station responds to this increase in path loss by increasing its transmit power (through downlink power control). This way, power received at the mobile is kept constant. As a result, the base station maintains a constant bit rate and QoS delivered to the mobile. Constant bit rate and QoS

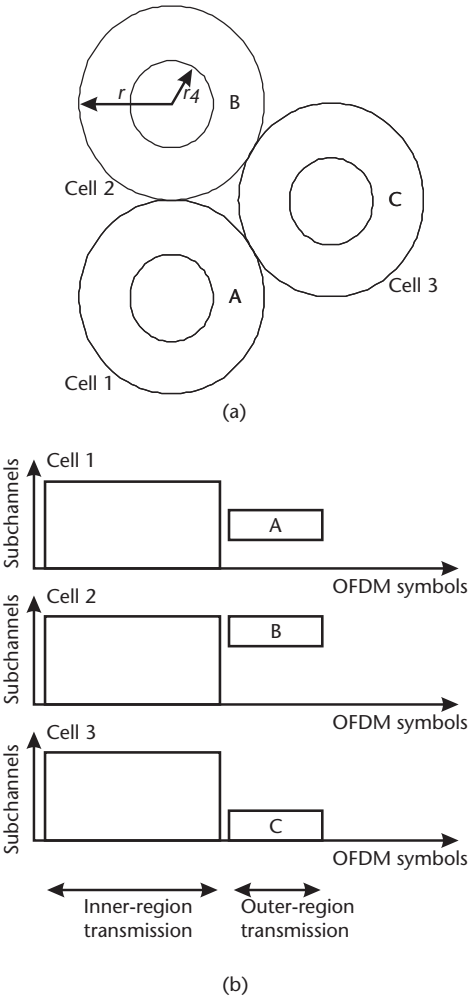


Figure 14.14 (a) Three cells in a cluster that uses fractional frequency reuse. (b) An example of their transmissions of subframes. (*After:* [6].)

regardless of a mobile's location are important in 2G because 2G systems primarily supported circuit-switched applications such as voice [7].

Maintaining constant bit rate and QoS regardless of the mobile's distance from the base station has two disadvantages, especially for broadband mobile systems. First, increasing the transmit power to a mobile far away means less downlink power resources for other mobiles in the same cell. Second, it is not necessary to maintain constant bit rate and QoS if data transmission is bursty and can tolerate latency. Therefore, in 4G broadband mobile systems, the base station can use its power resources to deliver the highest possible bit rate (on the downlink) to those mobiles that are closest to the base station. Figure 14.17 shows a base station that does not attempt to deliver a constant bit rate and QoS. As a mobile moves away from the base station, the mobile's received power decreases, but the base station does not increase the transmit power to compensate. Rather, the base station decreases the bit rate delivered to the mobile. As a result, the base station controls the

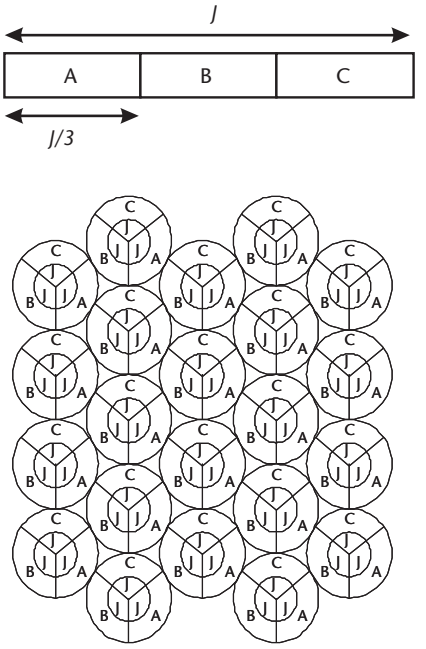


Figure 14.15 An example of fractional frequency reuse where inner regions use $1 \times 3 \times 1$ and outer regions use $1 \times 3 \times 3$.

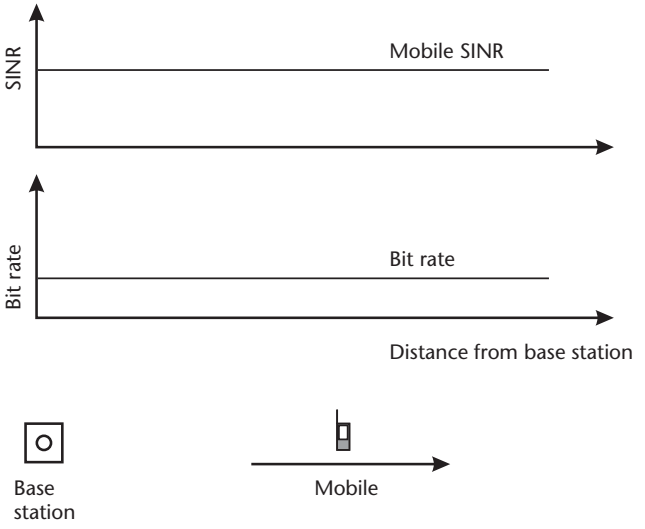


Figure 14.16 In 2G systems, the base station controls the power to maintain a constant bit rate and quality of service [7].

bit rates using a (relatively) constant transmit power [7]. In other words, instead of striving to give users a fixed bit rate and QoS, the base station now changes the bit rate and QoS depending on coverage. In OFDMA-based systems, this can be done using adaptive modulation and coding (AMC).

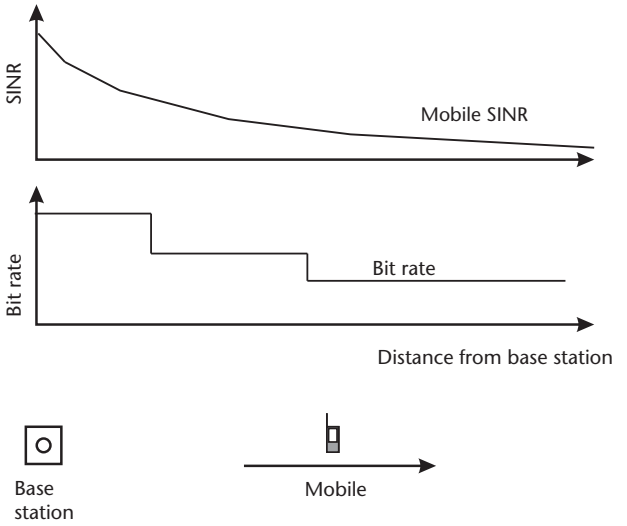


Figure 14.17 In 4G systems, the base station controls the bit rates using a constant transmit power [7].

14.4 Capacity

14.4.1 Instantaneous Bit Rate

System designers need to ascertain how many bits per second of capacity can be supplied by a base station (or sector) utilizing a given amount of spectrum. This information is needed to assess the supply of capacity and to estimate how many base stations (or sectors) are needed to accommodate the offered traffic in a given area. In this section, we illustrate how an instantaneous bit rate can be calculated under different conditions.

At the fundamental level, the bit rate R_b in bits per second corresponds to the amount of bits (in an OFDM symbol) that can be pumped through the channel in the duration of an OFDM symbol (over which an OFDM symbol lasts):

$$R_b = \frac{N_{FFT} \log_2 M}{T_s} \tag{14.4}$$

where

- N_{FFT} is the total number of subcarriers;
- M is the number of data symbols in the constellation (e.g., $M = 64$ for 64-QAM);
- T_s is the total OFDM symbol duration.

Readers may recognize that, in (14.4), N_{FFT} (the total number of subcarriers) is the same as the number of transformed symbols in an OFDM symbol. This is because, at the transmitter, all N_{FFT} data symbols are transformed by the IDFT function in parallel, and the resulting N_{FFT} transformed symbols (plus the additional cyclic prefix symbols) in parallel are converted to serial in the time domain.

Therefore, there are effectively N_{FFT} transformed symbols in the duration of an OFDM symbol (see Figure 4.3 in Chapter 4 for the OFDM transmitter). Because $\log_2 M$ is the number of bits carried by each data symbol, $N_{FFT}\log_2 M$ equates to the number of bits transmitted within the duration of an OFDM symbol.

N_{FFT} include all types of subcarriers, including guard subcarriers, the DC subcarrier, pilot subcarriers, and data subcarriers. Thus, one should use

$$R_b = \frac{(N_{used} - 1)\log_2 M}{T_s} \tag{14.5}$$

where N_{used} is the number of used subcarriers, including data subcarriers, pilot subcarriers, and the DC subcarrier (see Chapter 4), so $(N_{used} - 1)$ is the number of pilot subcarriers and data subcarriers (excluding the DC subcarrier).

R_b can be rewritten as follows by breaking up the total OFDM symbol time, T_s, into two components, T_b and T_g:

$$R_b = \frac{(N_{used} - 1)\log_2 M}{T_b + T_g} = \frac{(N_{used} - 1)\log_2 M}{T_b + GT_b} = \frac{(N_{used} - 1)\log_2 M}{T_b(1 + G)} \tag{14.6}$$

where:

- T_b is the useful OFDM symbol time;
- T_g is the guard time (i.e., cyclic prefix time) for an OFDM symbol;
- G is the ratio of guard time for an OFDM symbol to useful OFDM symbol time (i.e., $G = T_g/T_b$).

Because the (useful) OFDM symbol time T_b is the inverse of the OFDM subcarrier frequency spacing Δf (i.e., $T_b = 1/\Delta f$) due to Fourier transform, (14.6) can be rewritten as

$$R_b = \frac{(N_{used} - 1)\log_2 M}{\frac{1}{\Delta f}(1 + G)} = \Delta f \frac{(N_{used} - 1)\log_2 M}{1 + G} \tag{14.7}$$

Since the frequency spacing Δf between adjacent subcarriers can be approximated by the total bandwidth W divided by the total number of subcarriers N_{FFT} (i.e., $\Delta f \approx W/N_{FFT}$), the equation becomes:

$$R_b = \frac{W(N_{used} - 1)\log_2 M}{N_{FFT}(1 + G)} \tag{14.8}$$

Equation (14.8) can be examined from a different perspective: The bit rate in bits per second is always equal to the transmission symbol rate in symbols per second multiplied by the number of bits per symbol ($\log_2 M$). Under the best condition, the transmission symbol rate is the same as the bandwidth W. In other words, the transmission symbols are running at a transmission symbol rate of W symbols per second. Since each data symbol carries $\log_2 M$ bits of data, the bit rate $R_b = W\log_2 M$. However, R_b has to be reduced due to overheads. Specifically, R_b is scaled down by $1/(1 + G)$ due to the guard time overhead and by $(N_{used} - 1)/N_{FFT}$ due to the number of guard subcarriers and the DC subcarrier.

To complete the derivation, we need to incorporate a few more factors. First, the system uses oversampling (implementation specific to IEEE 802.16), so an oversampling factor n needs to be included. With oversampling, the frequency spacing Δf between adjacent subcarriers becomes $\Delta f = n(W/N_{FFT})$. Oversampling has the effect of *decreasing* the time between samples in the time domain and *increasing* the frequency spacing between subcarriers in the frequency domain.

Second, the quantity $(N_{used} - 1)$ still includes the overhead of pilot subcarriers, so a factor p is needed to exclude the pilot subcarriers. p is the ratio of the number of data subcarriers to the number of pilot subcarriers and data subcarriers. p depends on the specific permutation modes used. For example, the PUSC permutation modes results in four pilot subcarriers and 24 data subcarriers in each subchannel (see Chapter 5); thus, for PUSC, $p = 24/(4 + 24) = 24/28$.

Third, the error-correcting code rate R is needed to exclude the overhead bits for forward error correction. Incorporating these factors yields[2]

$$R_b = \frac{nW(N_{used} - 1)\log_2 M}{N_{FFT}(1 + G)} Rp \tag{14.9}$$

Note that the repetition factor is not included.

This bit rate is the instantaneous bit rate over an OFDM symbol. For a frame, this equation assumes that all OFDM symbols in the frame are used for transmission. Thus the bit rate specified by (14.9) represents the theoretical maximum in any one direction (downlink or uplink) for a corresponding modulation and coding scheme (M and R). This is the raw bit rate that the physical layer supports.

The corresponding raw spectral efficiency e (in bits per second per hertz) is

$$e = \frac{R_b}{W} = \frac{n(N_{used} - 1)\log_2 M}{N_{FFT}(1 + G)} Rp \tag{14.10}$$

which represents the bit rate deliverable using 1 Hz of spectrum.

2. In actuality, nW is really floor(nW/8,000)8,000 to produce an integer sampling frequency [8].

14.4.2 Instantaneous Bit Rate: Examples

To get an idea of the kinds of bit rates that can be delivered by the physical layer, we perform some sample calculations. We first examine the more efficient set of parameters to see the kind of bit rate attainable at the high end.

Example 14.1

The first scenario is 1,024-FFT OFDMA downlink using PUSC and 64-QAM modulation with rate 3/4 code in a 10-MHz bandwidth. The following are the parameters:

- $N_{FFT} = 1{,}024$;
- $M = 64$;
- $R = 3/4$;
- $W = 10$ MHz;
- $n = 28/25$;
- $N_{used} = 841$;
- $p = 24/28$;
- $G = 1/8$.

Substituting these parameters into equation (14.9) produces:

$$R_b = \frac{(28/25)(10 \times 10^6 \text{ Hz})(841-1)\log_2 64}{1024(1+1/8)}(3/4)(24/28) = 31.5 \times 10^6 \text{ bps or } 31.5 \text{ Mbps}$$

This bit rate assumes that all OFDM symbols in the frame are used for the downlink. Of course, this bit rate is partly due to the efficient modulation of 64-QAM with rate 3/4 code. Using such a modulation and coding scheme requires that the received SINR to be high. This means that the mobile has to be very close to the base station and in a favorable propagation environment. If there is more than one mobile, then those mobiles would have to share this bit rate at the MAC layer. In TDD, it means that those mobiles would divvy up and share OFDM symbols in time and subchannels in frequency.

For reference, the raw spectral efficiency for this example is

$$e = \frac{(28/25)(841-1)\log_2 64}{1024(1+1/8)}(3/4)(24/28) = 3.15 \text{ bps/Hz}$$

For a mobile that is far away from the base station or in a less favorable propagation environment, the system should adaptively throttle down to a more robust burst profile. The second example examines the more robust set of parameters to see the kinds of bit rate attainable at the low end.

Example 14.2

The second scenario is 1,024-FFT OFDMA downlink using PUSC and QPSK modulation with rate 1/2 code in a 10-MHz bandwidth. The following are the parameters:

- $N_{FFT} = 1{,}024$;
- $M = 4$;
- $R = 1/2$;
- $W = 10$ MHz;
- $n = 28/25$;
- $N_{used} = 841$;
- $p = 24/28$;
- $G = 1/8$.

Substituting these parameters into the equation yields:

$$R_b = \frac{(28/25)(10 \times 10^6 \text{ Hz})(841-1)\log_2 4}{1024(1+1/8)}(1/2)(24/28) = 7.0 \times 10^6 \text{ bps or } 7.0 \text{ Mbps}$$

Again, this bit rate assumes that all OFDM symbols in the frame are used for the downlink and is shared among the mobiles. If a mobile (or mobiles) is far away from the base station or in a less favorable propagation environment, the received SINR would be low, and the system changes to a more robust burst profile (QPSK with $R = 1/2$ instead of 64-QAM with $R = 3/4$) to ensure reliability of the link. The tradeoff, of course, is a lower bit rate and lower spectral efficiency. Incidentally, the raw spectral efficiency is:

$$e = \frac{(28/25)(841-1)\log_2 4}{1024(1+1/8)}(1/2)(24/28) = 0.7 \text{ bps/Hz}$$

Example 14.3

The third scenario is the downlink case of the link budget (Table 13.1) shown in Chapter 13: 1,024-FFT OFDMA downlink using PUSC and 16-QAM modulation with rate 1/2 code in a 10-MHz bandwidth. The following are the parameters:

- $N_{FFT} = 1{,}024$;
- $M = 16$;
- $R = 1/2$;
- $W = 10$ MHz;
- $n = 28/25$;
- $N_{used} = 841$;

- $p = 24/28$;
- $G = 1/8$.

Substituting these parameters into the equation yields:

$$R_b = \frac{(28/25)(10 \times 10^6 \text{ Hz})(841-1)\log_2 16}{1024(1+1/8)}(1/2)(24/28) = 14 \times 10^6 \text{ bps or 14 Mbps}$$

This bit rate again assumes that all OFDM symbols in the frame are used for the downlink and is shared among the mobiles. The raw spectral efficiency is:

$$e = \frac{(28/25)(841-1)\log_2 16}{1024(1+1/8)}(1/2)(24/28) = 1.4 \text{ bps/Hz}$$

14.4.3 Effective Data Rate

Equation (14.9) is the instantaneous bit rate over one OFDM symbol. This is the raw bit rate that the physical layer supports. As such, it assumes that all OFDM symbols in the frame are available for data transport. However, in actuality, there are various overheads in a frame. In addition, some OFDM symbols are for the downlink, and some OFDM symbols are for the uplink. Therefore, to calculate the effective data rate one needs to consider the following factors.

First, the symbols for the various overheads need to be taken into account. The overheads in an OFDMA frame include the preamble, FCH, DL-MAP, UL-MAP, TTG, RTG, and so forth (see Chapter 5). In OFDMA (IEEE 802.16e), each OFDM symbol lasts T_s which is

$$T_s = T_b + T_g = T_b(1+G) = \frac{1}{\Delta f}(1+G) \tag{14.11}$$

For $\Delta f = 10.9375$ kHz and $G = 1/8$, $T_s = 102.9$ μs. If a frame lasts 5 ms, then there are (5 ms/102.9 μs) = 48.6 OFDM symbols in a frame. Assuming 1.6 symbols are used for TTG and RTG means that 47 OFDM symbols are left for overheads and data for both the downlink and the uplink [9].

The amount of overheads depends on the number of active connections and the traffic types transported by these connections [10]. For example, for FTP traffic running over 10 active connections, there are 7 OFDM symbols for overheads in the downlink subframe and 3 OFDM symbols for overheads in the uplink subframe; these overheads leave (47 – 7 – 3) = 37 OFDM symbols in a frame for data [11]. For VoIP traffic, much of the overhead occurs in the MAP messages [12], and the MAP overhead increases linearly with the number of connections [11]. Thus, (14.9) needs to be adjusted based on the amount of overheads to arrive at the effective data rate [10].

Second, in a TDD frame, some OFDM symbols are dedicated to the downlink (in the downlink subframe) and some OFDM symbols are dedicated to the uplink (in the uplink subframe). So in TDD, the effective data rate that the MAC layer obtains has to be apportioned between the downlink and the uplink. A common assumption of the ratio between the downlink and the uplink is 3:1 for Web-oriented applications. For example, applying the 3:1 ratio to the available 37 OFDM symbols yields 28 OFDM symbols for data in the downlink subframe and 9 OFDM symbols for data in the uplink subframe [11].

14.5 Capacity and Coverage

It is clear that, based on the above examples, the capacity supplied by a single base station depends on how far is the mobile and how good is the propagation environment. Continuing with the above examples, consider two scenarios shown in Figures 14.18 and 14.19.

Figure 14.18 shows the first scenario (Example 14.1) that has a base station and a ring of coverage area closest around the base station. It is assumed that the ring of coverage area is close enough to the base station so that all mobiles in this ring can support 64QAM modulation with rate 3/4 code. In other words, all mobiles in this ring can meet the $SINR_{req}$ for 64QAM with rate 3/4 code. If all parameters used by the mobiles (in this ring of coverage area) and the base station are the same as Example 14.1, then the base station can supply, at most, a raw bit rate of 31.5 Mbps to the mobiles in this ring of coverage area, assuming there are no other mobiles outside this ring. Of course, the mobiles would have to share this capacity.

Figure 14.19 shows the second scenario that has the same base station but a ring of coverage area some distance away from the base station. In this figure, the ring of coverage area is so far from the base station that all mobiles in this ring can only support QPSK modulation with rate 1/2 code (i.e., meet the $SINR_{req}$ for QPSK with a rate 1/2 code). If all other parameters used by the mobiles (in this ring of coverage area) and the base station are the same as Example 14.2, then the base station can only supply a raw bit rate of 7 Mbps to the mobiles in this ring of coverage area (assuming there are no other mobiles outside this ring). Again, the mobiles would have to share this capacity.

In general, the mobiles would be geographically distributed from being close to the base station to being far from the base station. Thus, the actual bit rate that can be supplied by a base station will vary. In the scenarios shown above, the base station can supply a maximum instantaneous bit rate of 31.5 Mbps in 10 MHz of spectrum.

Figure 14.20 shows a base station with concentric rings of coverage areas around it. Each concentric ring can support a specific modulation and coding scheme, ranging from 64QAM (closest to the base station) to QPSK (farthest from the base station). To keep the figure concise, Figure 14.20 does not show the code rates. The radius of each concentric ring r_M can be calculated by using:

$$r_M = d_{ref}\left(\frac{P_T u_g u_l G_R}{(kT\Delta f)I_M}\left(\frac{c}{4\pi d_{ref} f}\right)^2 \Big/ SINR_{req,M}\right)^{1/\alpha} \tag{14.12}$$

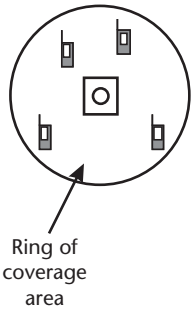


Figure 14.18 The best scenario: the mobiles in a ring of coverage area immediately around the base station.

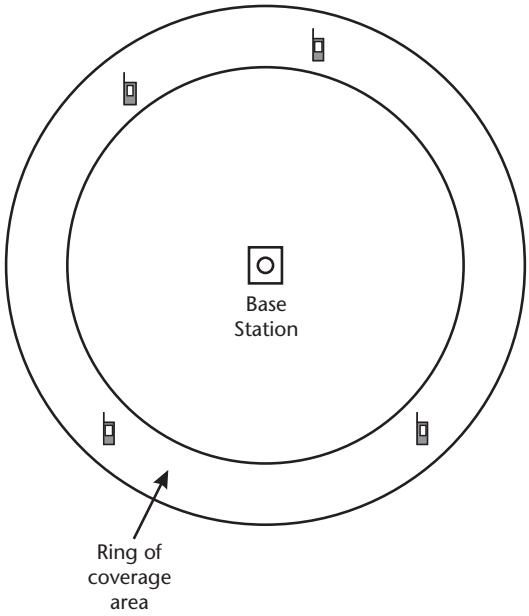


Figure 14.19 The worst scenario: the mobiles in a ring of coverage area far away from the base station (at the cell boundary).

This equation is adapted from (13.27) in Chapter 13, which is the closed-form model of distance as a function of the SINR.[3] The $SINR_{req,M}$ in this equation can be treated as the required SINR for a particular modulation M and coding scheme. Given the required SINR, one can calculate the radius (distance) within which that required SINR can be met. A sample set of required SINRs are shown in Table 13.2.

In Figure 14.20, r_M is the radius within which mobiles may operate using modulation M, and $SINR_{req,M}$ is the required SINR for modulation M. In particular:

- The 64-QAM ring ranges from 0 to r_{64}.
- The 16-QAM ring ranges from r_{64} to r_{16}.

3. For brevity, notation for subcarrier j is omitted.

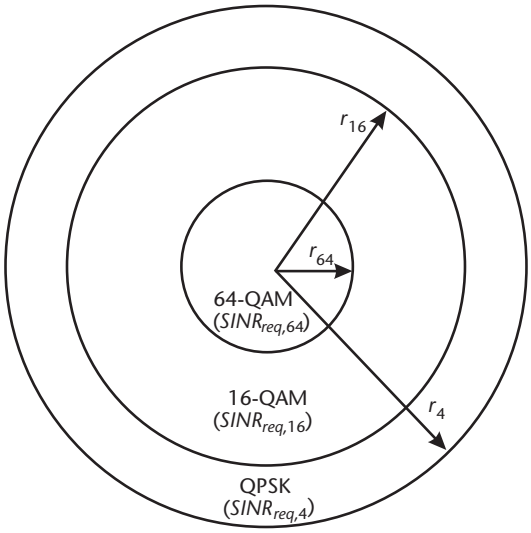


Figure 14.20 Concentric rings of coverage areas as a function of $SINR_{req}$.

- The QPSK ring ranges from r_{16} to r_4.

As the required SINR increases (e.g., from 16-QAM to 64-QAM), the radius of the corresponding modulation decreases.

In addition to the effect of $SINR_{req}$, there is also the effect of interference, which is shown in Figure 14.21. This figure shows that:

- From the perspective of *coverage*, as the interference rise I_M (e.g., from other cells) increases, the radii of all concentric rings shrink.
- From the perspective of *capacity*, for a fixed user in a cell, as the interference rise I_M increases, that user's received SINR decreases. Thus, the raw bit rate deliverable to the user may decrease too.

There have been work in the literature about both capacity and coverage in OFDMA-based systems. For example, one study [3] performed Monte Carlo simulations of an IEEE 802.16e system for six different frequency reuse patterns and assessed their capacity and SINR. The study assumes that the network access provider has 15 MHz (three blocks of 5 MHz) of spectrum available. Table 14.2 summarizes the results.

Some observations can be made on the results of this study:

- The reuse patterns $1 \times 1 \times 1$ and $3 \times 1 \times 1$ have the lowest average capacity.
- The reuse pattern $1 \times 3 \times 1$ has the highest average capacity, but at a cost of a very low average SINR.
- The reuse pattern $3 \times 3 \times 1$ affords the best compromise between average capacity and average SINR, followed by the reuse pattern $1 \times 3 \times 3$.

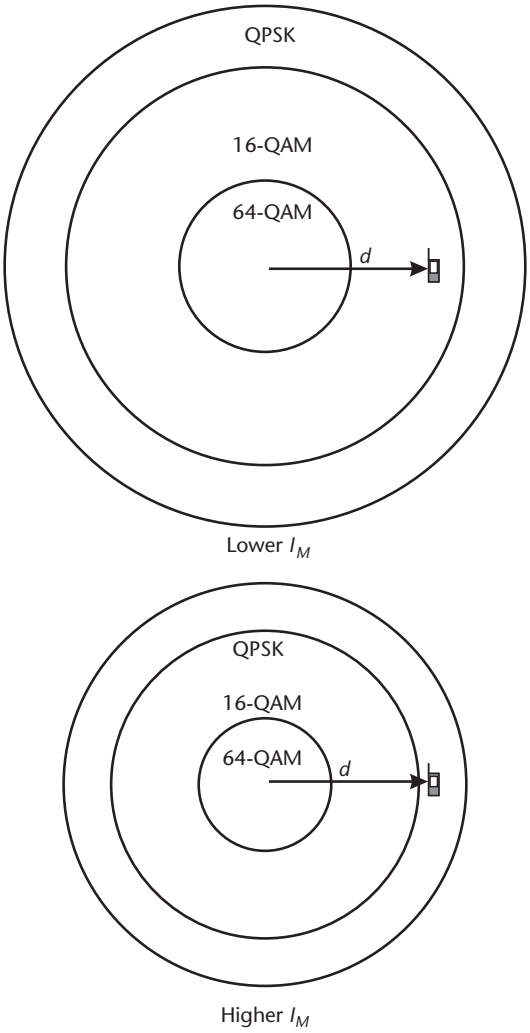


Figure 14.21 Concentric rings of coverage areas as a function of I_M.

Table 14.2 Simulated Capacity and Coverage [3]

Reuse Pattern	*Average Capacity (Mbps)*	*Average SINR (dB)*
$3 \times 1 \times 1$	5.2	12.4
$1 \times 1 \times 1$	5.3	3.8
$3 \times 3 \times 3$	8.3	20.1
$1 \times 3 \times 3$	15.5	12.1
$3 \times 3 \times 1$	16.8	13.3
$1 \times 3 \times 1$	18.8	4.7

- Increasing intercell frequency reuse factor N and/or intracell frequency reuse factor K increases the average SINR.

It is now well known that a system needs a (intercell or intracell) frequency reuse factor of about 3 [13]. Note that the above results do not include the effects of adaptive beamforming. Adaptive beamforming would increase both average SINR and average capacity.

14.6 Conclusions

For 4G systems built on IEEE 802.16 or other OFDMA-based schemes that adaptively supply capacity to a mobile (depending on distance and propagation condition), the smaller the cell, the higher the capacity. This is because the smaller the cell, the closer mobiles are to the base station, and the more efficient the modulation can be supported. However, the smaller the cells, the more cells are needed to cover a given area. Therefore, the discussion of *capacity* is tightly coupled with *coverage*, and a system designer must address both goals at the same time.

Typically, the coverage requirement first determines the cell radius (size). If the resulting cell capacity can handle the demand within that radius, then the total number of cells required for network-wide coverage is computed. After the computation, if the number of cells required for coverage is less than the number of cells required for capacity, then the system designer performs an iterative process to check if the cell radius can shrink to meet the required capacity [10].

References

[1] Laroia, R., S. Uppala, and J. Li, "Designing a Mobile Broadband Wireless Access Network," *IEEE Signal Processing,* Vol. 21, No. 5, 2004, pp. 20–28.

[2] Nuaymi, L., *WiMAX: Technology for Broadband Wireless Access*, New York: John Wiley & Sons, 2007.

[3] Maqbool, M., M. Couperchoux, and P. Godlewski, "Comparison of Various Frequency Reuse Patterns for WiMAX Networks with Adaptive Beamforming," *Proc. IEEE Vehicular Technology Conference*, May 11–14, 2008, pp. 2583–2586.

[4] Jia, H. et al., "On the Performance of IEEE 802.16 OFDMA System under different Frequency Reuse and Subcarrier Permutation Patterns," *Proc. IEEE International Conference on Communications*, June 24–28, 2007, pp. 5720–5725.

[5] Wang, F., et al., "Mobile WiMAX Systems: Performance and Evolution," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 41–49.

[6] Giuliano, R., C. Monti, and P. Loreti, "WiMAX Fractional Frequency Reuse for Rural Environments," *IEEE Wireless Communications*, Vol. 15, No. 3, 2008, pp. 60–65.

[7] Yang, S. C., *3G CDMA2000 Wireless System Engineering*, Norwood, MA: Artech House, 2004.

[8] IEEE Standard 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York: IEEE, October 1, 2004.

[9] WiMAX Forum, "WiMAX™ System Evaluation Methodology," 2008.

[10] Upase, B., M. Hunukumbure, and S. Vadgama, "Radio Network Dimensioning and Planning for WiMAX Networks," *Fujitsu Scientific and Technical Journal*, Vol. 43, No. 4, 2007, pp. 435–450.

[11] "Mobile WiMAX—Part I: A Technical Overview and Performance Evaluation," WiMAX Forum, August 2006.

[12] Fong, M.-H., et al., "Improved VoIP Capacity in Mobile WiMAX Systems using Persistent Resource Allocation," *IEEE Communications*, Vol. 46, No. 10, 2008, pp. 50–57.

[13] Belghith, A., and L. Nuaymi, "WiMAX Capacity Estimations and Simulation Results," *Proc. IEEE Vehicular Technology Conference*, May 11–14, 2008, pp. 1741–1745.

About the Author

Samuel C. Yang holds an undergraduate degree from Cornell University and two graduate degrees from Stanford University, all in electrical engineering. He also holds a Ph.D. in information science from Claremont Graduate University. Dr. Yang is currently a professor at California State University, Fullerton, where he teaches, conducts research, and consults in the areas of networks and wireless systems. He was named the Outstanding Faculty of the Year by his college for research, teaching, and service.

Before entering academia, Dr. Yang worked in the telecommunications industry in both managerial and professional capacities. He was a manager of RF planning at Verizon Wireless, where he managed the planning and design of its 2G and 3G wireless networks in western United States. In 1995, he played a key role in the design and commercialization of the first large-scale CDMA network in North America. Prior to Verizon Wireless, he worked at Hughes Space and Communications (now Boeing Satellite Systems), where he was a technical lead on several international communication satellite programs for Japan, China, and Thailand. While at Hughes, he also conducted research in multiple-access techniques and channel simulation and was a systems engineer on NASA's Magellan radar-mapping mission to Venus.

Dr. Yang has published articles in refereed journals and conference proceedings and is the author of two books in wireless communications, one of which was on the publisher's best sellers list. His current interests are the planning, analysis, design, and management of wireless networks and how organizations can best benefit from their use. Dr. Yang is also a registered professional engineer in the state of California.

Index

Recent Titles in the Artech House Mobile Communications Series

John Walker, Series Editor

3G CDMA2000 Wireless System Engineering, Samuel C. Yang

3G Multimedia Network Services, Accounting, and User Profiles, Freddy Ghys, Marcel Mampaey, Michel Smouts, and Arto Vaaraniemi

802.11 WLANs and IP Networking: Security, QoS, and Mobility, Anand R. Prasad, Neeli R. Prasad

Achieving Interoperability in Critical IT and Communications Systems, Robert I. Desourdis, Peter J. Rosamilia,

Christopher P. Jacobson, James E. Sinclair, and James R. McClure

Advances in 3G Enhanced Technologies for Wireless Communications, Jiangzhou Wang and Tung-Sang Ng, editors

Advances in Mobile Information Systems, John Walker, editor

Advances in Mobile Radio Access Networks, Y. Jay Guo

Applied Satellite Navigation Using GPS, GALILEO, and Augmentation Systems, Ramjee Prasad and Marina Ruggieri

Artificial Intelligence in Wireless Communications, Thomas W. Rondeau and Charles W. Bostian

Broadband Wireless Access and Local Network: Mobile WiMax and WiFi, Byeong Gi Lee and Sunghyun Choi

CDMA for Wireless Personal Communications, Ramjee Prasad

CDMA Mobile Radio Design, John B. Groe and Lawrence E. Larson

CDMA RF System Engineering, Samuel C. Yang

CDMA Systems Capacity Engineering, Kiseon Kim and Insoo Koo

CDMA Systems Engineering Handbook, Jhong S. Lee and Leonard E. Miller

Cell Planning for Wireless Communications, Manuel F. Cátedra and Jesús Pérez-Arriaga

Cellular Communications: Worldwide Market Development, Garry A. Garrard

Cellular Mobile Systems Engineering, Saleh Faruque

The Complete Wireless Communications Professional: A Guide for Engineers and Managers, William Webb

EDGE for Mobile Internet, Emmanuel Seurre, Patrick Savelli, and Pierre-Jean Pietri

Emerging Public Safety Wireless Communication Systems,
Robert I. Desourdis, Jr., et al.

The Future of Wireless Communications, William Webb

Geographic Information Systems Demystified, Stephen R. Galati

GPRS for Mobile Internet, Emmanuel Seurre, Patrick Savelli, and Pierre-Jean Pietri

GPRS: Gateway to Third Generation Mobile Networks, Gunnar Heine and Holger Sagkob

GSM and Personal Communications Handbook, Siegmund M. Redl, Matthias K. Weber, and Malcolm W. Oliphant

GSM Networks: Protocols, Terminology, and Implementation, Gunnar Heine

GSM System Engineering, Asha Mehrotra

Handbook of Land-Mobile Radio System Coverage, Garry C. Hess

Handbook of Mobile Radio Networks, Sami Tabbane

High-Speed Wireless ATM and LANs, Benny Bing

Interference Analysis and Reduction for Wireless Systems, Peter Stavroulakis

Introduction to 3G Mobile Communications, Second Edition, Juha Korhonen

Introduction to Communication Systems Simulation, Maurice Schiff

Introduction to Digital Professional Mobile Radio, Hans-Peter A. Ketterling

Introduction to GPS: The Global Positioning System, Ahmed El-Rabbany

An Introduction to GSM, Siegmund M. Redl, Matthias K. Weber, and Malcolm W. Oliphant

Introduction to Mobile Communications Engineering, José M. Hernando and F. Pérez-Fontán

Introduction to Radio Propagation for Fixed and Mobile Communications, John Doble

Introduction to Wireless Local Loop, Second Edition: Broadband and Narrowband Systems, William Webb

IS-136 TDMA Technology, Economics, and Services, Lawrence Harte, Adrian Smith, and Charles A. Jacobs

Location Management and Routing in Mobile Wireless Networks, Amitava Mukherjee, Somprakash Bandyopadhyay, and Debashis Saha

Mobile Data Communications Systems, Peter Wong and David Britland

Mobile IP Technology for M-Business, Mark Norris

Mobile Satellite Communications, Shingo Ohmori, Hiromitsu Wakana, and Seiichiro Kawase

Mobile Telecommunications Standards: GSM, UMTS, TETRA, and ERMES, Rudi Bekkers

Mobile Telecommunications: Standards, Regulation, and Applications, Rudi Bekkers and Jan Smits

Multiantenna Digital Radio Transmission, Massimiliano Martone

Multiantenna Wireless Communications Systems, Sergio Barbarossa

Multipath Phenomena in Cellular Networks, Nathan Blaunstein and Jørgen Bach Andersen

Multiuser Detection in CDMA Mobile Terminals, Piero Castoldi

OFDMA for Broadband Wireless Access, Slawomir Pietrzyk

OFDMA System Analysis and Design, Samuel C. Yang

Personal Wireless Communication with DECT and PWT, John Phillips and Gerard Mac Namee

Practical Wireless Data Modem Design, Jonathon Y. C. Cheah

Prime Codes with Applications to CDMA Optical and Wireless Networks, Guu-Chang Yang and Wing C. Kwong

Quantitative Analysis of Cognitive Radio and Network Performance, Preston Marshall

QoS in Integrated 3G Networks, Robert Lloyd-Evans

Radio Engineering for Wireless Communication and Sensor Applications, Antti V. Räisänen and Arto Lehto

Radio Propagation in Cellular Networks, Nathan Blaunstein

Radio Resource Management for Wireless Networks, Jens Zander and Seong-Lyun Kim

Radiowave Propagation and Antennas for Personal Communications, Third Edition, Kazimierz Siwiak and Yasaman Bahreini

RDS: The Radio Data System, Dietmar Kopitz and Bev Marks

Resource Allocation in Hierarchical Cellular Systems, Lauro Ortigoza-Guerrero and A. Hamid Aghvami

RF and Baseband Techniques for Software-Defined Radio, Peter B. Kenington

RF and Microwave Circuit Design for Wireless Communications, Lawrence E. Larson, editor

Sample Rate Conversion in Software Configurable Radios, Tim Hentschel

Signal Processing Applications in CDMA Communications, Hui Liu

Smart Antenna Engineering, Ahmed El Zooghby

Software Defined Radio for 3G, Paul Burns

Spread Spectrum CDMA Systems for Wireless Communications, Savo G. Glisic and Branka Vucetic

Technologies and Systems for Access and Transport Networks, Jan A. Audestad

Third Generation Wireless Systems, Volume 1: Post-Shannon Signal Architectures, George M. Calhoun

Traffic Analysis and Design of Wireless IP Networks, Toni Janevski

Transmission Systems Design Handbook for Wireless Networks, Harvey Lehpamer

UMTS and Mobile Computing, Alexander Joseph Huber and Josef Franz Huber

Understanding Cellular Radio, William Webb

Understanding Digital PCS: The TDMA Standard, Cameron Kelly Coursey

Understanding GPS: Principles and Applications, Second Edtion, Elliott D. Kaplan and Christopher J. Hegarty, editors

Understanding WAP: Wireless Applications, Devices, and Services, Marcel van der Heijden and Marcus Taylor, editors

Universal Wireless Personal Communications, Ramjee Prasad

WCDMA: Towards IP Mobility and Mobile Internet, Tero Ojanperä and Ramjee Prasad, editors

Wireless Communications in Developing Countries: Cellular and Satellite Systems, Rachael E. Schwartz

Wireless Communications Evolution to 3G and Beyond, Saad Z. Asif

Wireless Intelligent Networking, Gerry Christensen, Paul G. Florack, and Robert Duncan

Wireless LAN Standards and Applications, Asunción Santamaría and Francisco J. López-Hernández, editors

Wireless Technician's Handbook, Second Edition, Andrew Miceli

For further information on these and other Artech House titles,including previously considered out-of-print books now available through our In-Print-Forever® (IPF®) program, contact:

Artech House
685 Canton Street
Norwood, MA 02062
Phone: 781-769-9750
Fax: 781-769-6334
e-mail: artech@artechhouse.com

Artech House
16 Sussex Street
London SW1V 4RW UK
Phone: +44 (0)20 7596-8750
Fax: +44 (0)20 7630-0166
e-mail: artech-uk@artechhouse.com

Find us on the World Wide Web at: www.artechhouse.com