CHAPTER 4

WIRELESS MEDIUM ACCESS ALTERNATIVES

4.1 INTRODUCTION

This chapter presents an overview of the access methods commonly used in wireless networks. Access methods form a part of Layer 2 of the OSI protocol stack and Layer 3 of the IEEE 802 standard for LANs that is responsible for interacting with the medium to coordinate the successful operation of multiple terminals over the wireless channel. Most multiple access methods were originally developed for wired networks and later on adopted to the wireless medium. However, requirements on the wired and wireless media are different, thereby demanding modifications in the original protocols to make them suitable for the wireless medium. Today the main differences between wireless and wired channels are availability of bandwidth and reliability of transmissions. The wired medium is moving toward optical media with enormous bandwidth and very reliable transmission. Bandwidth in wireless systems is always limited because the medium (air) cannot be duplicated, and the medium is shared between all wireless systems, including multichannel broadcast television and a number of other bandwidth demanding applications and services. In the case of wired operation, we can always lay additional cable to increase the capacity as needed even if it is an expensive proposition. In a wireless environment, we can reduce the size of *cells* to increase capacity as discussed in Chapter 5. With the reduction of the size of the cells, the number of cells increases, and the need for improvements in the wired infrastructure to connect these cells increases. Also the complexity of the network for handling additional handoffs and mobility management increases posing a practical limitation upon the maximum capacity of the network. As far as transmission reliability is concerned, as we saw in Chapter 2, the wireless medium always suffers from multipath and fading, which causes a serious threat to reliable data transmission over the communication link. Because the wireless channel is so unreliable, as discussed in Chapter 3, people have developed a number of signal processing techniques to improve transmission reliability over the wireless channel. In spite of these techniques, the reliability of the wireless medium is far below that of the wired medium used as the backbone of the wireless networks.

Although in practice we prefer to have the same access method and the same frame structure for wired backbone and the wireless access, wireless networks often use different packet sizes and a modified access method to optimize the performance to the specifics of the unreliable wireless medium.

Example 4.1: IEEE 802.3 and IEEE 802.11

The IEEE 802.3 standard (based on Ethernet) is the successful and dominant standard for wired LANs. Consequently, the IEEE 802.11 WLAN standard, in ideal situations, desired using the same access method as previously established with IEEE 802.3. Carrier sense multiple-access with collision detection (CSMA/CD) is the protocol used in the Ethernet. However, collision detection in wireless channels is extremely challenging, and IEEE 802.11 had to resort to carrier sense multiple-access with collision avoidance (CSMA/CA) that can be viewed as a wireless adaptation of IEEE 802.3.

Example 4.2: ATM and Wireless ATM

In the 1990s, ATM was perceived to be the transmission scheme for all future networking. In the mid 1990s when wireless solutions were considered, a wireless ATM working group was formed to extend the ATM short packet solution with QoS for wireless access. The group had to make significant compromises as discussed in Chapter 12 because ATM was designed for broadband and reliable transmission over optical channels.

To avoid substantial overlap with existing literature, we describe access methods used in wireless networks with justification of why and how they are employed in different wireless networks.

As we explained in previous chapters, wireless networks have evolved around voice or data applications, and as a result we can divide them into voice- and data-oriented networks. The access methods adopted by voice- and data-oriented networks are quite different. Voice-oriented networks are designed for relatively long telephone conversations as the main application. Each communication session for this application exchanges several megabytes of information in both directions. A signaling channel that exchanges short messages between two calling components sets up the call by obtaining resources (such as the link, switches, etc.) in the telephone network at the beginning of the conversation and terminates these arrangements by releasing the resources at the end of the call. The wireless access methods evolved for interaction with these networks assigns a slot of time, a portion of frequency, or a specific code to a user preferably for the entire length of the conversation. We refer to these techniques as fixed-assignment channel access methods or channel partitioning techniques. Data networks were originally designed for bursts of data for which the supporting network does not have a separate signaling channel. In packet communications each packet carries some "signaling information" related to the address of the destination and the source. We refer to the access methods used in these networks as random-access methods accommodating randomly arriving packets of data. Certain local area data networks also *take turns* in accessing the medium as in the case of token passing and polling schemes. In some other cases, the random access mechanisms are used to temporarily *reserve* the medium for transmitting the packet. In the next two sections of this chapter, we provide a short description of the fixed-assignment and random access methods used in voice- and data-oriented wireless networks, respectively.

4.2 FIXED-ASSIGNMENT ACCESS FOR VOICE-ORIENTED NETWORKS

All existing voice-oriented wireless networks such as cellular telephony or PCS services use fixed-assignment channel access or channel partitioning techniques. In the fixed-assignment access method, a fixed allocation of channel resources, frequency, time, or a spread spectrum code are made available on a predetermined

basis to a single user for the duration of the communication session. The three basic fixed-assignment multiple-access methods are FDMA, TDMA, and CDMA. The choice of an access method will have a great impact on the capacity and QoS provided by a network. The impact of multiple access schemes is so important that we commonly refer to various voice-oriented wireless systems by their channel access method, which is only a part of the layer two specification of the air interface of the network.

Example 4.3: Common Terminology for Digital Cellular Systems

The GSM and the North American IS-136 digital cellular standards are commonly referred to as digital TDMA cellular systems and the IS-95/IMT-2000 standards are called digital CDMA cellular systems.

In reality, different systems use different modulation techniques as well. However, as we will see in the rest of this book, the impact of the choice of access method on the capacity and overall performance of the network is much more profound. Consequently, the system is really distinguished by its access method. As we will see in our examples of cellular networks, a network that is identified with an access technique often uses other random or fixed assignment techniques as a part of its overall operation. However, it is identified by the access techniques employed for transferring the main information source for which the network is designed to carry.

Example 4.4: Random Access Techniques in Cellular Networks

GSM uses slotted ALOHA (a random access method) to establish a link between the mobile terminal and the base station. It also has an optional frequency-hopping pattern that improves the system performance when there is fading of the radio signal. However, the GSM network is built for voice communications, and each session uses TDMA as the access method.

Another important design parameter related to the access method is the differentiation between the carrier frequencies of the forward (downlink—communication between the base station and mobile terminals) and reverse (uplink—communication between the mobile terminal and the base station) channels. If both forward and reverse channels use the same frequency band for communications, but the forward and reverse channels employ alternating time slots, the system is referred to as employing TDD. If the forward and reverse channels use different carrier frequencies that are sufficiently separated, the duplexing scheme is referred to as FDD. With TDD, because only one frequency carrier is needed for a duplex operation, we can share more of the RF circuitry between the forward and the reverse channels. The reciprocity of the channel in TDD allows for exact open-loop power control and simultaneous synchronization of the forward and reverse channels. TDD techniques are used in systems intended for low-power local area communications where interference must be carefully controlled and where low complexity and low-power consumption are very important. Thus TDD systems are often used in local area pico- or microcellular systems deployed by PCS networks. FDD is

mostly used in macrocellular systems designed for coverage of several tens of kilometers where implementation of TDD is more challenging (see Fig. 4.1).

4.2.1 Frequency Division Multiple Access (FDMA)

In an FDMA environment, all users can transmit signals simultaneously, and they are separated from one another by their frequency of operation. The FDMA technique is built upon FDM. FDM is the oldest and still a commonly used multiplexing technique in the trunks connecting switches in the PSTN. It is also the choice of radio and TV broadcast, as well as cable TV distribution. FDM is more suitable for analog technology because it is easier to implement. When FDM is used for channel access, it is referred to as FDMA.

Example 4.5: FDMA in AMPS with FDD

Figure 4.1(a) shows the FDMA/FDD system commonly used in 1G analog cellular systems such as AMPS and a number of early cordless telephones. In FDMA/FDD systems, forward and reverse channels use different carrier frequencies, and a fixed subchannel pair is assigned to a user terminal during the communication session. At the receiving end, the mobile terminal filters the designated channel out of the composite signal. As shown in Figure 4.2(a), the AMPS system allocates 30 kHz of bandwidth for each forward and reverse channel. There are a total of 421 channels in 25 MHz of spectrum assigned to each direction; 395 of these channels are used for the voice traffic and the rest for signaling.

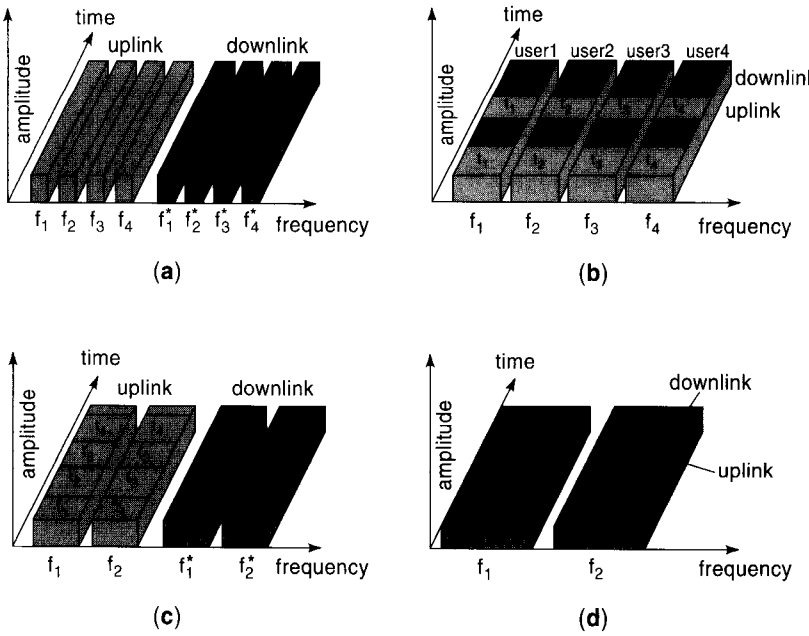

Figure 4.1 (a) FDMA/FDD, (b) FDMA/TDD, (c) TDMA/FDD with multiple carriers, (d) TDMA/TDD with multiple carriers.

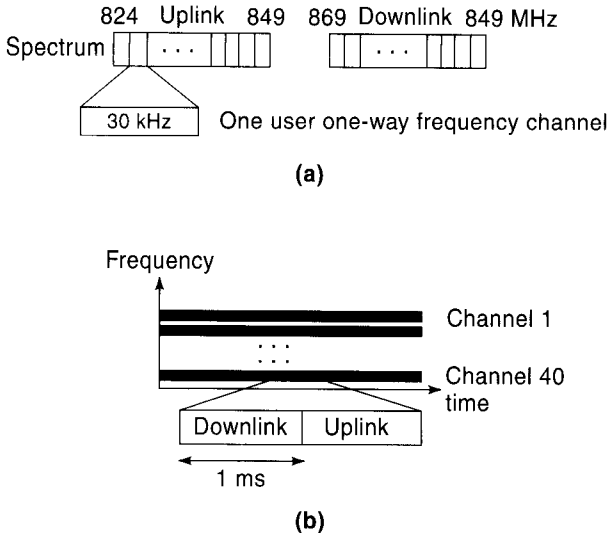

Figure 4.2 (a) FDMA/FDD in AMPS and (b) FDMA/TDD in CT-2.

Example 4.6: FDMA in CT-2 with TDD

Figure 4.1(b) shows an FDMA/TDD system used in the CT-2 digital cordless telephony standard. Each user employs a single carrier frequency for all communications. The forward and reverse transmissions take turns via alternating time slots. This system was designed for distances of up to 100 meters, and a voice conversation is based on 32 kbps ADPCM voice coding. As shown in Figure 4.2(b) the total allocated bandwidth for CT-2 is 4 MHz, supporting 40 carriers each using 100 kHz of bandwidth.

The designer of an FDMA system must pay special attention to adjacent channel interference, in particular in the reverse channel. In both forward and reverse channels, the signal transmitted must be kept confined within its assigned band, at least to the extent that the out-of-band energy causes negligible interference to the users employing adjacent channels. Operation of the forward channel in wireless FDMA networks is very similar to wired FDM networks. In forward wireless channels, in a manner similar to that of wired FDM systems, the signal received by all mobile terminals has the same received power, and interference is controlled by adjusting the sharpness of the transmitter and receiver filters for the separate carrier frequencies. The problem of adjacent channel interference is much more challenging on the reverse channel. On the reverse channel, mobile terminals will be operating at different distances from the BS. The RSS at the BS, of a signal transmitted by a mobile terminal close to the BS, and the RSS at the BS of a transmission by a mobile terminal at edges of the cell are often substantially different, causing problems in detecting the weaker signal. This problem is usually referred to as the near-far problem. If the out-of-band emissions are large, they may swamp the actual information-carrying signal.

Problem 1: Near-Far Problem

a) What is the difference between the received signal strength of two terminals located in 10 m and 1 km from a base station in an open area?

b) Explain the effects of shadow fading on the difference in the RSS.

c) What would be the impact if the two terminals were operating in two adjacent channels? Assume out-of-band radiation that is 40 dB below the main lobe.

Solution:

a) As we saw in Chapter 2, the received signal strength falls by around 40 dB per decade of distance in open areas. Therefore, the received powers from a mobile terminal that is 10 meters from a BS and another, that is at a distance of 1 km, are 80 dB apart.

b) In addition to the fall of the RSS with distance, we also discussed the issues of multipath and shadow fading in radio channels that cause power fluctuations on the order of several tens of dBs. Therefore, the difference in the received powers due to the near-far problem may exceed even 100 dB.

c) If the out-of-band emission is only 40 dB below that of the transmitted power, it may exceed the strength of the information-bearing signal by almost 60 dB.

To handle the near-far problem, FDMA cellular systems adopt two different measures. First, when frequencies are assigned to a cell, they are grouped such that the frequencies in each cell are as far apart as possible. The second measure employed is power control that is discussed in Chapter 6. In addition, whenever FDMA is employed, *guard bands* are included in the frequency channel to further reduce adjacent channel interference. This, however, has the effect of reducing the overall spectrum efficiency.

4.2.2 Time Division Multiple Access (TDMA)

In TDMA systems, a number of users share the same frequency band by taking assigned turns in using the channel. The TDMA technique is built upon the TDM scheme commonly used in the trunks for the telephones systems. The major advantage of the TDMA over FDMA is its format flexibility. Because of the fully digital format and the flexibility of buffering and multiplexing functions, time-slot assignments among multiple users are readily adjustable to provide different access rates for different users. This feature is particularly adopted in the PSTN, and the TDM scheme forms the backbone of all digital connections in the heart of the PSTN. The hierarchy of digital transmission trunks used in North America is the so-called T-carrier system that has an equivalent European system (the E-carriers) approved by the ITU. In the hierarchy of digital transmission rates standardized throughout North America, the basic building block is the 1.544 Mbps link known as T-1 carrier. A T-1 transmission frame is formed by TDD 24 PCM-encoded voice channels, each carrying 64 kbps of users data. Service providers often lease T-carriers to interconnect their own switches and routers and for forming their own networks.

Example 4.7: The Use of T-carriers in Cellular Networks

Cellular networks often lease T-carriers from the long-haul telephone companies to interconnect their own switches referred to as mobile switching centers

(MSCs). The difference between the MSC and a regular switch in the PSTN is that the MSC can support mobility of the terminal. The details of these differences are discussed in later chapters when we provide examples of cellular networks. The end-user subscribes to the cellular service provider.

Example 4.8: The Use of T-1 Lines in the Internet

The routers in the Internet are sometimes connected through leased T-carrier telephone lines to form part of the Internet. The difference between a router and a PSTN switch is that the router can handle packet switching whereas the PSTN switch uses circuit switching. The end-user subscribes to an Internet service provider (ISP) in this case.

With TDMA, a transmit controller assigns time slots to users, and an assigned time slot is held by a user until the user releases it. At the receiving end, a receiver station synchronizes to the TDMA signal frame, and extracts the time slot designated for that user. The heart of this operation is synchronization that was not needed for FDMA systems. The TDMA concept was developed in the 1960s for use in digital satellite communication systems and first became operational commercially in telephone networks in the mid-1970s [PAH95].

In cellular and cordless systems, the migration to TDMA from FDMA took place in the 2G systems. The first cellular standard adopting TDMA was GSM. The GSM standard was initiated to support international roaming among Scandinavian countries in particular and the rest of Europe in general. The digital voice adoption in TDMA format facilitated the network implementation, resulted in improvements in the quality of the voice, and provided a flexible format to integrate data services in the cellular network. The FDMA systems in the United States very quickly observed a capacity crunch in major cities, and among the options for increasing capacity, TDMA was adopted initially through the IS-54 system that was later on replaced by IS-136. TDMA was adopted in 2G cordless telephones such as DECT to provide format flexibility and to allow more compact and low-power terminals.

Example 4.9: TDMA in GSM

Figure 4.1(c) shows an FDMA/TDMA/FDD channel used in 2G digital cellular networks. Figure 4.3 shows a particular example of the 8-slot TDMA scheme used in the GSM system. Forward and reverse channels use separate carrier frequencies (FDD). Each carrier can support up to eight simultaneous users via TDMA, each using a 13 kbps encoded digital speech, within a 200 MHz carrier bandwidth. A total of 124 frequency carriers (FDMA) are available in the 25 MHz allocated band in each direction. One hundred kHz of band is allocated as a guard band at each edge of the overall allocated band.

Example 4.10: TDMA in DECT

Figure 4.1(d) shows an FDMA/TDMA/TDD system used in the Pan-European digital PCS standard DECT. Because distances are short, a TDD format allows

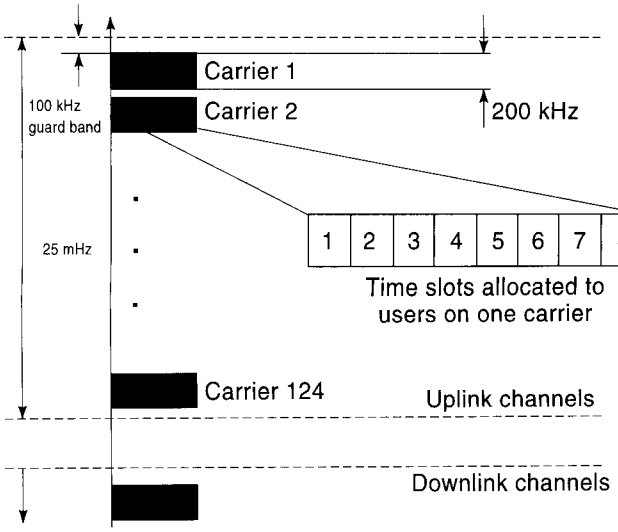

Figure 4.3 FDMA/TDMA/FDD in GSM.

using the same frequency for forward and reverse operations. The bandwidth per carrier is 1.728 MHz which can support up to 12 ADPCM coded speech channels via TDMA. The total allocated band in the EC is 10 MHz that can support five carriers (FDMA). Figure 4.4 shows the details of the TDMA/TDD time slots use in the DECT system. The frame duration is 10 ms, with 5 ms for portable-to-fixed station and 5 ms for fixed-to-portable. The transmitter transfers information in signal bursts which it transmits in slots of duration 10/24 = 0.417 ms. With 480 bits per slot (including a 64-bit guard time), the total bit rate is 1.152 Mbps. Each slot contains 64 bits for system control (C, P, Q, and M channels) and 320 bits for user information (I channel).

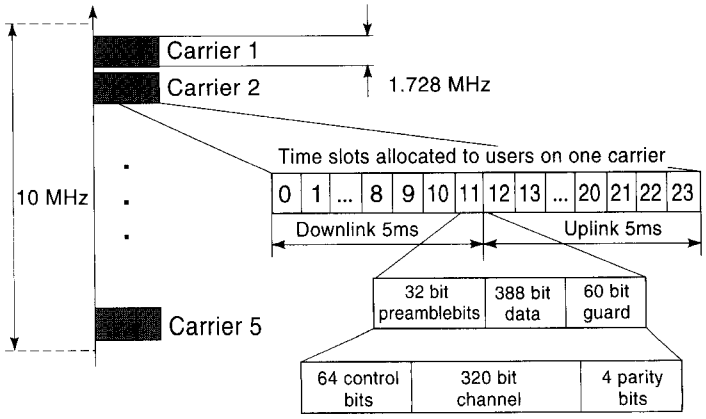

Figure 4.4 FDMA/TDMA/TDD in DECT.

Example 4.11: TDMA in IS-136

Figure 4.5 shows the frame format for the TDMA/FDD with six slots considered for IS-136 both for the forward (base to mobile) and reverse (mobile to base) channels. In IS-136 each 30 kHz digital channel has a channel transmission rate of 48.6 kbps. The 48.6 kbps stream is divided into six TDMA channels of 8.1 kbps each. The IS-136 slot and frame format, shown in Figure 4.5, is much simpler than that of the GSM standard. The 40-ms frame is composed of six 6.67-ms time slots. Each slot contains 324 bits, including 260 bits of user data, and 12 bits of system control information in a slow associated control channel (SACCH). There is also a 28-bit synchronization sequence, and a 12-bit digital verification color code (DVCC) used to identify the frequency channel to which the mobile terminal is tuned. In the mobile-to-base direction, the slot also contains a guard time interval of a six-bit duration when no signal is transmitted, and a six-bit ramp interval to allow the transmitter to reach its full output power level.

Due to the near-far problem, the received signal on the reverse channel from a user occupying a time slot can be much larger than the received power from the terminal using the adjacent time slot. In such a case, the receiver will have difficulties in distinguishing the weaker signal from the background noise. In a manner similar to FDMA systems, TDMA systems also use power control to handle this near-far problem.

4.2.3 Code-Division Multiple Access (CDMA)

With the growing interest in the integration of voice, data, and video traffic in telecommunication networks, CDMA appears increasingly attractive as the wireless access method of choice. Fundamentally, integration of various types of traffic is readily accomplished in a CDMA environment as coexistence in such an

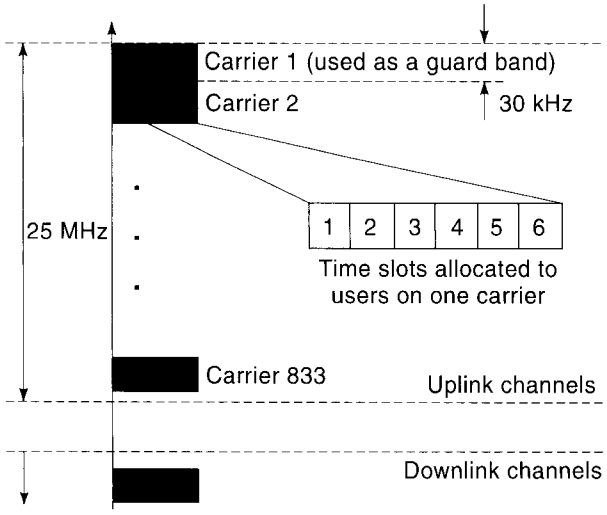

Figure 4.5 FDMA/TDMA/FDD in IS-136 standard.

environment does not require any specific coordination among user terminals. In principle, CDMA can accommodate various wireless users with different bandwidth requirements, switching methods and technical characteristics without any need for coordination. Of course, because each user signal contributes to the interference seen by other users, power control techniques are essential in the efficient operation of a CDMA system.

To illustrate CDMA and how it is related to FDMA and TDMA, it is useful to think of the available band and time as resources we use to share among multiple users. In FDMA, the frequency band is divided into slots, and each user occupies that frequency throughout the communication session. In TDMA, a larger frequency band is shared among the terminals, and each user uses a slot of time during the communication session. As shown in Figure 4.6, in a CDMA environment multiple users use the same band at the same time, and the user is differentiated by a code that acts as the key to identify that user. These codes are selected so that when they are used at the same time in the same band a receiver knowing the code of a particular user can detect that user among all the received signals. In the CDMA/FDD [Figure 4.7(a)] that is used in IS-95 and IMT-2000, the forward and reverse channels use different carrier frequencies. If both transmitter and receiver use the same carrier frequency [Figure 4.7(b)], the system is CDMA/TDD.

In CDMA, each user is a source of noise to the receiver of other users, and if we increase the number of users beyond a certain value, the entire system collapses because the signal received in each specific receiver will be buried under the noise caused by many other users. An important question is, how many users can simultaneously use a CDMA system before the system collapses?

4.2.3.1 Capacity of CDMA

CDMA systems are implemented based on the spread spectrum technology that is presented in Chapter 3. In its most simplified form, a spread spectrum transmitter spreads the signal power over a spectrum N times wider than the spectrum

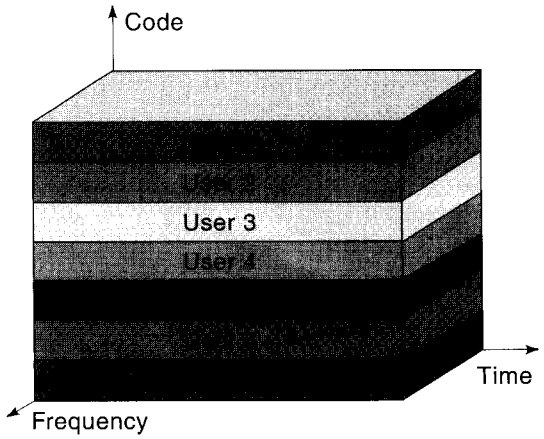

Figure 4.6 Simple illustration of CDMA.

of the message signal. In other words, an information bandwidth of R occupies a transmission bandwidth of W, where:

$$W = NR \tag{4.1}$$

The spread spectrum receiver processes the received signal with a *processing gain* of N. This means that during the processing at the receiver, the power of the received signal having the code of that particular receiver will be increased N times beyond the value before processing.

Let us consider the situation of a single cell in a cellular system employing CDMA. Assume that we have M simultaneous users on the reverse channel of a CDMA network. Further let us assume that we have an ideal power control enforced on the channel so that the received power of signals from all terminals has the same value P. Then, the received power from the target user after processing at the receiver is NP, and the received interference from $M - 1$ other terminals is $(M - 1)P$. If we also assume that a cellular system is interference limited and the background noise is dominated by the interference noise from other users, the received signal-to-interference ratio for the target receiver will be:

$$S_r = \frac{NP}{(M-1)P} = \frac{N}{M-1} \tag{4.2}$$

All users always have a requirement for the acceptable error rate of the received data stream. For a given modulation and coding specification of the system, that error rate requirement will be supported by a minimum S_r requirement that can be used in Eq. (4.2) to solve for the number of simultaneous users. Then, solving Eqs. (4.1) and (4.2) for M, we will have:

$$M = \frac{W}{R}\frac{1}{S_r} + 1 \cong \frac{W}{R}\frac{1}{S_r} \tag{4.3}$$

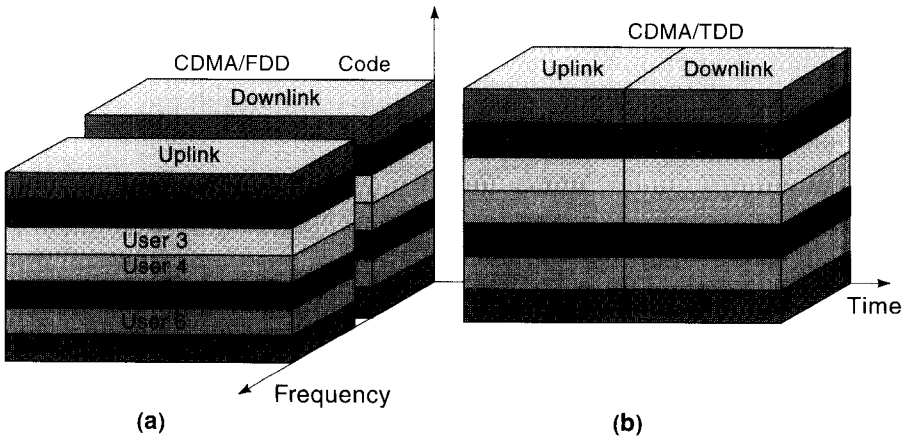

Figure 4.7 (a) CDMA/FDD and (b) CDMA/TDD.

Problem 2: Capacity of One Carrier in a Single-Cell CDMA System

Using QPSK modulation and convolutional coding, the IS-95 digital cellular systems require 3 dB < S_r < 9 dB. The bandwidth of the channel is 1.25 MHz, and the transmission rate is R = 9600 bps. Find the capacity of a single IS-95 cell.

Solution:

Using Equation (4.3) we can support from up to

$$M = \frac{1.25 \text{ MHz}}{9600 \text{ bps}} \frac{1}{8} \approx 16 \text{ to } M = \frac{1.25 \text{ MHz}}{9600 \text{ bps}} \frac{1}{2} \approx 65 \text{ users.}$$

4.2.3.2 Practical Considerations

In the practical design of digital cellular systems, three other parameters affect the number of users that can be supported by the system, as well as the bandwidth efficiency of the system. These are the number of sectors in each base station antenna, the voice activity factor, and the interference increase factor. These parameters are quantified as factors used in the calculation of the number of simultaneous users that the CDMA system can support. The use of sectored antennas is an important factor in maximizing bandwidth efficiency. Cell sectorization using directional antennas reduces the overall interference, increasing the allowable number of simultaneous users by a *sectorization gain factor,* which we denote by G_A. With ideal sectorization the users in one sector of a base station antenna do not interfere with the users operating in other sectors, and $G_A = N_{sec}$ where N_{sec} is the number of sectors in the cell. In practice antenna patterns cannot be designed to have ideal characteristics, and due to multipath reflections, users in general communicate with more than one sector. Three-sector base station antennas are commonly used in cellular systems, and a typical value of the sectorization gain factor is assumed to be G_A = 2.5 (4 dB). The voice activity interference reduction factor G_v is the ratio of the total connection time to the active talkspurt time. On the average, in a two-way conversation, each user talks roughly 50 percent of the time. The short pauses in the flow of natural speech reduce the activity factor further to about 40 percent of the connection time in each direction. As a result, the typical number used for G_v is 2.5 (4 dB). The interference increase factor H_0 accounts for users in other cells in the CDMA system. Because all neighboring cells in a CDMA cellular network operate at the same frequency, they will cause additional interference. This interference is relatively small due to the processing gain of the system and the distances involved; a value of H_0 = 1.6 (2 dB) is commonly used in the industry.

Incorporating these three factors as a correction to Equation (4.3), the number of simultaneous users that can be supported in a CDMA cell can be approximated by

$$M = \frac{W}{R}\frac{1}{S_r} + 1 \cong \frac{W}{R}\frac{1}{S_r}\frac{G_A G_v}{H_0} \tag{4.4}$$

If we define the *performance improvement factor* in a digital cellular system as

$$K = \frac{G_A G_v}{H_0} \tag{4.5}$$

assuming the typical parameter values given earlier, the performance improvement factor is $K = 4$ (6 dB).

Problem 3: Capacity of One Carrier in a Multi-Cell CDMA System with Correction Factors

Determine the multicell IS-95 CDMA capacity with correction for sectorization and voice activity. Use the numbers from Problem 2.

Solution:

If we continue the previous example with the new correction factor included, the range for the number of simultaneous users becomes $64 < M < 260$.

4.2.4 Comparison of CDMA, TDMA, and FDMA

With the success of IS-95 CDMA systems in its challenge to conventional IS-136 TDMA systems in the United States and the adoption of W-CDMA as the primary choice for the 3G cellular networks, one wonders why CDMA has become the favorite choice for wireless access in voice-oriented networks. Spread spectrum technology became the favorite technology for military applications because of its capability to provide a low probability of interception and strong resistance to interference from jamming. In the cellular industry, CDMA was introduced as an alternative to TDMA to improve the capacity of 2G cellular systems in the United States. As a result, much of the early debates in this area were focused on calculation of the capacity of CDMA as it is compared with TDMA. However, capacity is not the only reason for the success of the CDMA technology. As a matter of fact, calculation of the capacity of CDMA using the simple approach provided earlier is *not* very conclusive and is subject to a number of assumptions such as perfect power control that cannot be practically met. The first CDMA service providers in the United States were using slogans such as "you cannot believe your ears!" to address the superior quality of voice for the CDMA. However, the superiority of voice is partially dependent on the speech coder, and it is not a CDMA versus TDMA issue. In order to provide a good explanation for the success of a complex and multidisciplinary technology, such as a cellular network, addressing consumer market issues has always been very important. Those of us involved in this debate for the past decade have seen the discussion of the ups and downs of CDMA in variety of forums. One of the most interesting events that the principal author remembers was in 1997 in a major wireless conference in Taipei where one the most famous figures in this debate in his keynote speech at the opening of the conference declared that "we have seen in the past that the VHS which was not a better technology defeated BETA." In his perception, at that time, CDMA was similar to BETA. In less than a year or so after that, CDMA was selected by a number of different communities around the world as the technology of choice for 3G and IMT-2000.

In the rest of this section, we bring out a number of issues that may enlighten the reader toward a deeper understanding of the technical aspects of CDMA systems as they are compared with TDMA and FDMA networks. We hope that this may lead the reader to her/his own conclusion about the success of CDMA.

Format Flexibility. Telephone voice was the dominant source of income for the telecommunication industry up to the end of the past century. In the new millennium, the strong emergence of Internet and cable TV industries has created a case for other popular multimedia applications. The cellular phones that were designed for telephony applications are now being used for other applications and need support for multimedia applications. To support a variety of data rates with different requirements, a network needs format flexibility. As we discussed earlier, one of the reasons for migrating from analog FDMA to digital TDMA was that TDMA provides a more flexible environment for integration of voice and data. The time slots of a TDMA network designed for voice transmission can be used individually or in a group format to transmit data from users and to support different data rates. However, all these users should be time synchronized and the quality of the transmission channel is the same for all of them. The chief advantage of CDMA relative to TDMA is its flexibility in timing and the quality of transmission. In CDMA users are separated by their codes, unaffected by the transmission time relative to other users. The power of the user can also be adjusted with respect to others to support a certain quality of transmission. In CDMA each user is far more liberated from the other users, allowing a fertile setting to accommodate different service requirements to support a variety of transmission rates with different qualities of transmission to support multimedia or any other emerging application.

Performance in Multipath Fading. As we saw in Chapter 2, multipath in wireless channels causes frequency selective fading. In frequency selective fading, when the transmission band of a narrowband system coincides with the location of the fade, no useful signal is received. As we increase the transmission bandwidth, fading will occupy only a portion of the transmission band, providing an opportunity for a wideband receiver to take advantage of the portion of the transmission band not under fade and a more reliable communication link. In Chapter 3 we introduced DFE, OFDM, sectored antennas, and spread spectrum as technologies that can be employed in wideband systems to handle frequency selective fading. The wider the bandwidth, the better is the opportunity for averaging out the faded frequency.

These technologies are not used in the 1G analog cellular FDMA systems because they were analog systems and these techniques are digital. The Pan European GSM digital cellular system uses 200 kHz of band, and the standard recommends using DFE. The North American digital cellular system, IS-136, uses digital transmission over the same analog band of 30 kHz of the North American AMPS system and does not recommend equalization because the bandwidth is not very large. An equalizer needs additional circuitry, and some power budget at the receiver that was one of the drawbacks considered in IS-136. The bandwidth of the IS-95 CDMA system is 1.25MHz and W-CDMA systems for 3G networks use bandwidths that are as high 10 MHz. RAKE receivers are used to increase the benefits of wideband transmission by taking advantage of the so-called in-band or time

diversity of the wideband signal. This is one of the reasons for having a better quality of voice in CDMA systems. As we mentioned earlier, quality of voice is also effected by the robustness of the speech-coding algorithm, coverage of service, methods to handle interference, handoffs, and power control as well.

System Capacity. Comparison of the capacity depends on a number of issues, including the frequency reuse factor, speech coding rate, and the type of antenna. Therefore a fair comparison would be difficult unless we go to practical systems. The following simple example compares the capacity of FDMA (AMPS), TDMA (IS-136), and CDMA (IS-95) used in debates to evaluate alternatives for the 2G North American digital cellular systems to replace the 1G analog.

Problem 4: Comparison of the Capacity of Different 2G Systems

Compare the capacity of IS-95 CDMA with AMPS FDMA and IS-136 TDMA systems. For the CDMA system, assume an acceptable signal to interference ratio of 6 dB, data rate of 9600 bps, voice duty cycle of 50 percent, effective antenna separation factor of 2.75 (close to ideal 3-sector antenna), and neighboring cell interference factor of 1.67.

Solution:

For the IS-95 CDMA using Equation (4.4) for each carrier with W = 1.25 MHz, R = 9600 bps, S_r = 4 (6dB), G_v = 2 (50 percent voice activity), G_A = 2.75, and H_0 = 1.67 we have:

$$M = \frac{W}{R}\frac{1}{S_r}\frac{G_A G_v}{H_0} = 108 \text{ users per cell}$$

For the IS-136 with a carrier bandwidth of W_c = 30 kHz, the number of users per carrier of N_u = 3, and frequency reuse factor of K = 4 (commonly used in these systems), each W = 1.25 MHz of bandwidth provides for

$$M = \frac{W}{W_c}\frac{N_u}{K} = 31.25 \text{ users per cell}$$

For the AMPS analog system with carrier bandwidth of W_c = 30 kHz, and frequency reuse factor of K = 7 (commonly used in these systems), each W = 1.25 MHz of bandwidth provides for

$$M = \frac{W}{W_c}\frac{1}{K} = 6 \text{ users per channel}$$

Another example of this form is instructive to compare these systems with the GSM.

Problem 5: Comparison of NA Systems with GSM

Determine the capacity of GSM for K = 3.

Solution:

For the GSM system with a carrier bandwidth of W_c = 200 kHz, the number of users per carrier of N_u = 8, and frequency reuse factor of K = 3 (commonly used in these systems), each W = 1.25MHz of bandwidth provides for

$$M = \frac{W}{W_c}\frac{N_u}{K} = 16.7 \text{ users per cell}$$

Handoff. As we discuss in Chapter 6, handoff occurs when a received signal in an MS becomes weak and another BS can provide a stronger signal to the MS. The 1G FDMA cellular systems often used the so-called hard-decision handoff in which the base station controller monitors the received signal from the BS and at the appropriate time switches the connection from one BS to another. TDMA systems use the so-called *mobile-assisted handoff* in which the mobile station monitors the received signal from available BSs and reports it to the base station controller which then makes a decision on the handoff. Because adjacent cells in both FDMA and TDMA use different frequencies, the MS has to disconnect from and reconnect to the network that will appear as a click to the user. Handoffs occur at the edge of the cells when the received signals from both BSs are weak. The signals also fluctuate anyway because they are arriving over radio channels. As a result, decision making for the handoff time is often complex, and the user experiences a period of poor signal quality and possibly several clicks during the completion of the handoff process. Because adjacent cells in a CDMA network use the same frequency, a mobile moving from one cell to another can make "seamless" handoff by the use of signal combining. When the mobile station approaches the boundary between cells, it communicates with both cells. A controller combines the signals from both links to form a better communication link. When a reliable link has been established with the new base station, the mobile stops communicating with the previous base station, and communication is fully established with the new base station. This technique is referred to as soft handoff. Soft handoff provides a dual diversity for the received signal from two links which improves the quality of reception and eliminates clicking as well as the ping-pong problem. Handoff is an important issue that has many more details and we will discuss these details in Chapter 6.

Power Control. As we discussed earlier in this chapter, power control is necessary for FDMA and TDMA systems to control adjacent channel interference and mitigate the unexpected interference caused by the near-far problem. In FDMA and TDMA systems, some sort of power control is needed to improve the quality of the voice delivered to the user. In CDMA, however, the capacity of the system depends *directly* on the power control, and an accurate power control mechanism is needed for proper operation of the network. With CDMA, power control is the key ingredient in maximizing the number of users that can operate simultaneously in the system. As a result, CDMA systems adjust the transmitted power more often and with smaller adjustment steps to support a more refined control of power. Better power control also saves on the transmission power of the MS, which increases the life of the battery. The more refined power control in CDMA systems also helps in power management of the MS, which is an extremely important practical issue for users of the mobile terminals. These issues are further discussed in Chapters 6 and 8.

Implementation Complexity. Spread spectrum is a two-layer modulation technique requiring greater circuit complexity than conventional modulation schemes.

This in turn will lead to higher electronic power consumption and larger weight and cost for mobile terminals. Gradual improvements in battery and integrated circuit technologies, however, have made this issue transparent to the user.

4.2.5 Performance of Fixed-Assignment Access Methods

Fixed assignment access methods are used with circuit switched cellular and PCS telephone networks. In these networks, in a manner similar to the wired multichannel environments, the performance of the network is measured by the blockage rate of an initiated call. A call does not go through for two reasons: (1) when the calling number is not available, and (2) when the telephone company is out of resources to provide a line for the communication session. In POTS, for both cases the user hears a busy tone signal and cannot distinguish between the two types of blockage. In most cellular systems, however, type (1) blockage results in a response that is a busy tone and type (2) with a message such as "All the circuits are busy at this time please try your call later." In the rest of this book, we refer to blockage rate only as a type (2) blockage rate. The statistical properties of the traffic offered to the network are also a function of time. The telephone service providers often design their networks so that the blockage rate at peak traffic is always below a certain percentage. Cellular operators often try to keep this average blockage rate below 2 percent.

The blockage rate is a function of the number of subscribers, number of initiated calls, and the length of the conversations. In telephone networks, the Erlang equations are used to relate the probability of blockage to the average rate of the arriving calls and the average length of a call. In wired networks, the number of lines or subscribers that can connect to a multichannel switch is a fixed number. The telephone company monitors the statistics of the calls over a long period of time and upgrades the switches with the growth of subscribers so that the blockage rate during peak traffic times remains below the objective value. In cellular telephony and PCS networks, the number of subscribers operating in a cell is also a function of time. In the downtown areas, everyone uses their cellular telephones during the day, and in the evenings they use them in their residential area which is covered by a different cell. Therefore, traffic fluctuations in cellular telephone networks are much more than the traffic fluctuations in POTS. In addition, telephone companies can easily increase the capacity of their networks by increasing their investment on the number of transmission lines and quality of switches supporting network connections. In wireless networks, the overall number of available channels for communications is ultimately limited by the availability of the frequency bands assigned for network operation. To respond to the fluctuations of the traffic and cope with the bandwidth limitations, cellular operators use complex frequency assignment strategies to share the available resources in an optimal manner. Some of these issues are discussed in Chapter 5.

4.2.5.1 Traffic Engineering Using the Erlang Equations

The Erlang equations are the core of the traffic engineering for telephony applications. The two basic equations used for traffic engineering are Erlang B and Erlang C equations. The Erlang B equation relates the probability of blockage

$B(N,\rho)$ to the number of channels N and the normalized call density in units of channels ρ. The Erlang B formula is:

$$B(N, \rho) = \frac{\rho^N / N!}{\sum_{i}^{N} (\rho^i / i!)} \tag{4.6}$$

where $\rho = \lambda / \mu$, λ is the call arrival rate and μ is the service rate of the calls.[1]

Problem 6: Call Blocking Using Erlang B Formula

We want to provide a wireless public phone service with five lines to a ferry crossing between Helsinki and Stockholm carrying 100 passengers where on the average each passenger makes a three-minute telephone call every two hours. What is the probability of a passenger approaching the telephones and none of the ~~four~~ five lines are available?

Solution:

In practice, often the probability of call blockage is given, and we need to calculate the number of subscribers. Here we need an inverse function for the Erlang equation that is not available. As a result, a number of tables and graphs are available for this inverse mapping. Figure 4.8 shows a graph relating the probability of blockage $B(N,\rho)$ to the number of channels N and the normalized traffic per available channels ρ. From this graph, we can estimate the blocking probability. The traffic load is 100 users $\times$ 1 call/user $\times$ 3 minutes/call per 120 minutes = 2.5 Erlangs. Because there are five lines available and the traffic is 2.5 Erlangs, the blocking probability is roughly 0.07.

Problem 7: Capacity Using Erlang B Formula

An IS-136 cellular phone provider owns 50 cell sites and 19 traffic carriers per cell each with a bandwidth of 30 kHz. Assuming each user makes three calls per hour and the average holding time per call of five minutes, determine the total number of subscribers that the service provider can support with a blocking rate of less than 2 percent.

Solution:

The total number of channels is $N = 19 \times 3 = 57$ per cell. For $B(N,\rho) = 0.02$ and $N = 57$ Figure 4.8 shows that the $\rho = 45$ Erlangs. With an average of five calls per minute, the service rate is $\mu = 1/5$ minutes, and the acceptable arrival rate of the calls is $\lambda = \rho \times \mu = 1/5\ (\text{min}^{-1}) \times 45\ (\text{Erlang}) = 9\ (\text{Erlang/min})$. With an average of 3 calls per hour, the system can accept 9 (Erlang/min) / 3 (Erlang) / 60 (min) = 180 subscribers per cell. Therefore the total number of subscribers are 180 (subscribers/cell) $\times$ 50 (cells) = 8,000 subscribers.

The Erlang C formula relates the waiting time in a queue if a call does not go through, but it is buffered until a channel is available. These equations start with

[1]The equation assumes that the arrivals are Poisson, and the service rate is exponential. For details, see [BER87].

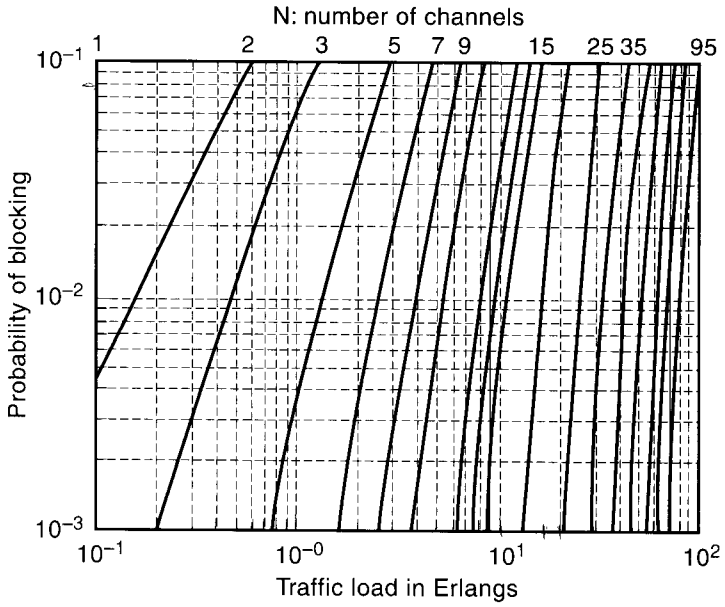

Figure 4.8 Erlang B chart showing the blocking probability as a function of offered traffic and number of channels.

the probability that a call does not get processed immediately and gets delayed. The probability that a call gets delayed is given by:

$$P(\text{delay} > 0) = \frac{\rho^N}{\rho^N + N!\left(1 - \frac{\rho}{N}\right)\sum_{k=0}^{N-1}\frac{\rho^k}{k!}} \tag{4.7}$$

Because of the complexity of the calculation, tables or graphs are again used to provide values for this probability based on normalized values of ρ. Figure 4.9 illustrates the relationship between probability of delay, number of channels N, and the normalized traffic per available channel ρ. The probability of having a delay that is more than a time t is given by:

$$P[delay > t] = P[delay > 0]e^{-(N-\rho)\mu t} \tag{4.8}$$

This indicates the exponential distribution of the delay time. The average delay is then given by the average of the exponential distribution:

$$D = P[delay > 0]\frac{1}{\mu(N - \rho)} \tag{4.9}$$

Problem 8: Call Delay Using Erlang C Formula

For the ferry described in Problem 6 answer the following questions:

a) What is the average delay for a passenger to get access to the telephone?

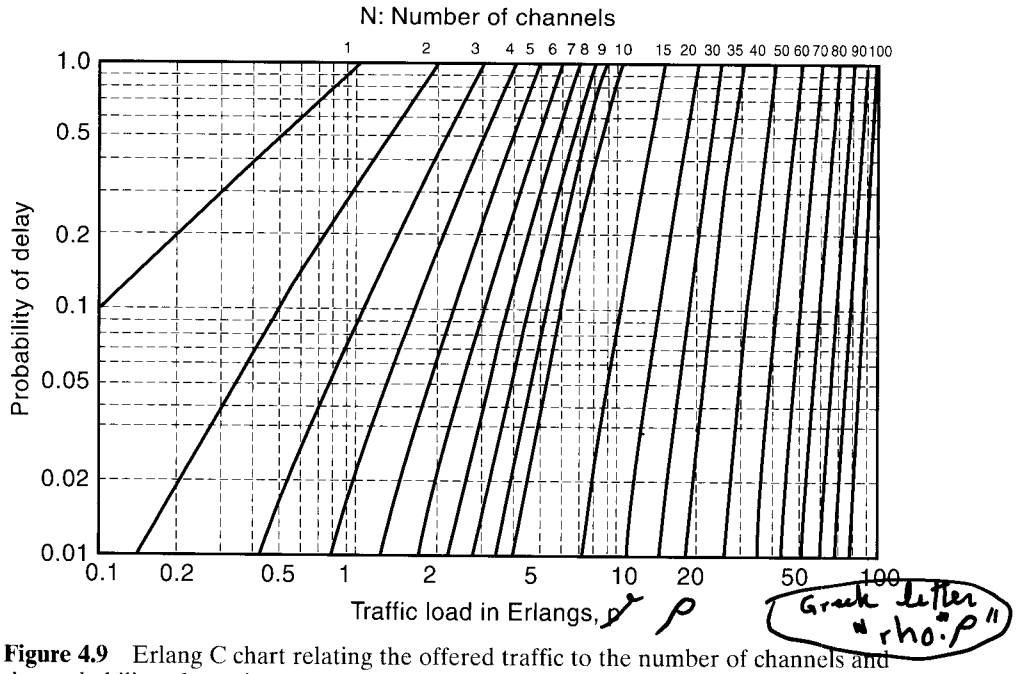

Figure 4.9 Erlang C chart relating the offered traffic to the number of channels and the probability of queuing.

b) What is the probability of having a passenger waiting more than a minute for the access to the telephone?

Solution:

a) Using Equation (4.7) for $N = 5$ and $\rho = 2.5$ we have $P[\text{Delay} > 0] = 0.13$. Using Equation (4.9), the average delay is $0.13/(5 - 2.5)/3 = 0.17$ minutes.

b) Using Equation (4.8) $P[\text{delay} > 1\text{min}] = 0.13 \exp[-(5 - 2.5)1/3] = 0.13 \exp(-0.83) = 0.0565$.

4.3 RANDOM ACCESS FOR DATA-ORIENTED NETWORKS

Random access methods have evolved around bursty data applications for computer communications. In our discussion of fixed-assignment access methods, we noted that such methods make relatively efficient use of communications resources when each user has a steady flow of information to be transmitted. This would be the case, for example, with digitized voiced traffic, data file transfer, or facsimile transmission. However, if the information to be transmitted is intermittent or bursty in nature, fixed-assignment access methods can result in communication resources being wasted for much of the duration of the connection. Furthermore, in wireless networks, where subscribers pay for service as a function of channel connection time, fixed-assignment access can be an expensive means of transmitting short messages and will also involve large call setup times. *Random access* methods

provide a more flexible and efficient way of managing channel access for communicating short bursty messages. In contrast to fixed-assignment access schemes, random access schemes provide each user station with varying degrees of freedom in gaining access to the network whenever information is to be sent. A natural consequence of randomness of user access is that there is contention among the users of the network for access to a channel, and this is manifested in collisions of contending transmissions. Therefore these access schemes are sometimes called contention-based schemes or simply *contention schemes*.

Random access techniques are widely used in wired LANs, and the literature in computer networking provides adequate description of these techniques. When applied to wireless applications, these techniques often are modified from their original wired version [CHA00]. The objective of the rest of this section is to describe the evolution of random access techniques that are used in wireless networks. We first discuss the random access methods used in wireless data networks, and then we provide some details of the access methods used in WLAN applications.

4.3.1 Random Access Methods for Mobile Data Services

The random access methods used in mobile data networks can be divided into two groups. The first group consists of ALOHA-based access methods for which the mobile terminals transmit their contention packet without any coordination between them. The second group is the carrier-sense based random access techniques for which the terminal senses the availability of the channel before it transmits its packets.

4.3.1.1 ALOHA-Based Wireless Random Access Techniques

The original *ALOHA protocol* is sometimes called *pure ALOHA* to distinguish it from subsequent enhancements of the original protocol. This protocol derives its name from the ALOHA system, a communications network developed by Norman Abramson and his colleagues at the University of Hawaii and first put into operation in 1971 [ABR70]. The initial system used ground-based UHF radios to connect computers on several of the island campuses with the university's main computer center on Oahu, by use of a random access protocol which has since been known as the ALOHA protocol. The word ALOHA means "hello" in Hawaiian.

The basic concept of ALOHA protocol is very simple. A mobile terminal transmits an information packet when the packet arrives from the upper layers of the protocol stack. Simply put, mobile terminals say "hello" to the air interface as the packet arrives. Each packet is encoded with an error-detection code. The BS checks the parity of the received packet. If the parity checks properly, the BS sends a short acknowledgment packet to the MS. Of course, because the MS packets are transmitted at arbitrary times, there will be collisions between packets whenever packet transmissions overlap by any amount of time, as indicated in Figure 4.10(a). Thus, after sending a packet, the user waits a length of time more than the round-trip delay for an acknowledgment from the receiver. If no acknowledgment is received, the packet is assumed lost in a collision, and it is transmitted again with a randomly selected delay to avoid repeated collisions.

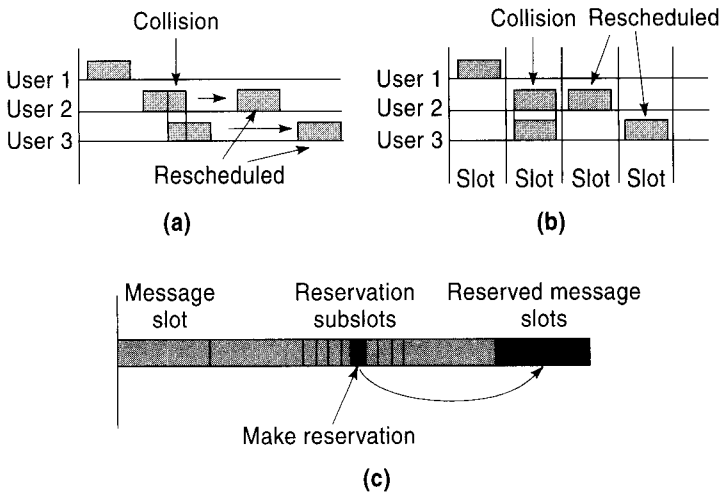

Figure 4.10 (a) Pure ALOHA protocol, (b) slotted ALOHA protocol, and (c) reservation ALOHA.

The advantage of ALOHA protocol is that it is very simple, and it does not impose any synchronization between mobile terminals. The terminals transmit their packets as they become ready for transmission, and if there is a collision, they simply retransmit. The disadvantage of the ALOHA protocol is its low throughput under heavy load conditions. If we assume that packets arrive randomly, they have the same length, and are generated from a large population of terminals. The maximum throughput of the pure ALOHA is 18 percent.

Problem 9: Throughput of Pure ALOHA

a) What is the maximum throughput of a pure ALOHA network with a large number of users and a transmission rate of 1 Mbps?

b) What is the throughput of a TDMA network with the same transmission rate?

c) What is the throughput of the ALOHA network if only one user was effective?

Solution:

a) For a large number of mobile terminals, each using a transmission rate of 1 Mbps to access a BS using ALOHA protocol, the maximum data rate that successfully passes through to the BS is 180 kbps.

b) If we have a TDMA system with negligible overhead (long packets), the throughput defined this way is nearly 100 percent and the BS receives data at a maximum rate of 1 Mbps.

c) The 1 Mbps can be attained in an ALOHA system only if we have one user (no-collision) who transmits all the time.

In wireless channels where bandwidth limitations often impose serious concerns for data communications applications, this technique is often changed to its synchronized version referred to as slotted ALOHA. The maximum throughput of

a slotted ALOHA system under the conditions mentioned earlier is 36 percent which is double the throughput of pure ALOHA.

In slotted ALOHA protocol, shown in Figure 4.10(b), the transmission time is divided into time slots. The BS transmits a beacon signal for timing, and all MSs synchronize their time slots to this beacon signal. When a user terminal generates a packet of data, the packet is buffered and transmitted at the start of the next time slot. With this scheme we eliminate partial packet collision. Assuming equal length packets, either we have a complete collision or we have no collisions. This doubles the throughput of the network. The report on collision and retransmission mechanisms remains the same as in pure ALOHA. Because of its simplicity, the slotted ALOHA protocol is commonly used in the early stages of registration of an MS to initiate a communication link with the BS.

Example 4.12: Slotted ALOHA in GSM

In the GSM system, the initial contact between the MS and the BS to establish a traffic channel for TDMA voice communications is performed through a random access channel using slotted ALOHA protocol. Other voice-oriented cellular systems adopt similar approaches as the first step in the registration process of an MS.

Throughput of slotted ALOHA protocol is still very low for wireless data applications. This technique is sometimes combined with TDMA systems to form the so-called reservation-ALOHA (R-ALOHA) protocol, shown in Figure 4.10(c) In R-ALOHA, time slots are divided into contention periods and contention free periods. During the contention interval, an MS uses very short packets to contend for the upcoming contention free intervals that will be used for transmission of long information packets. The R-ALOHA protocol was used in the Altair WLANs that were developed in the early 1990s to operate in licensed frequency bands around 18–19 GHz. The detailed implementation of R-ALOHA can take a variety of forms, and for that reason sometimes it is used under different names. The follow-

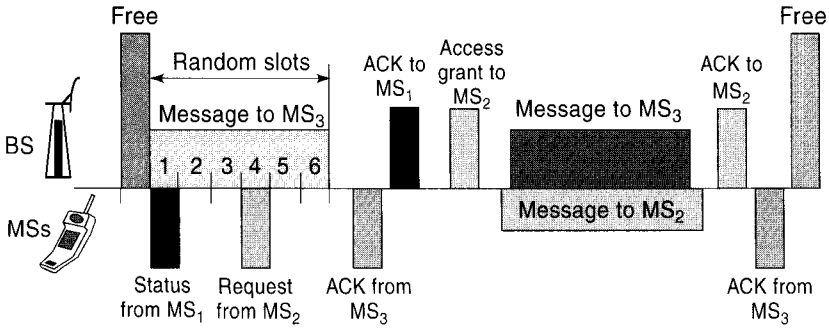

Figure 4.11 Dynamic slotted ALOHA used in Mobitex.

ing example provides some details of the so-called dynamic slotted ALOHA protocol that is used in the Mobitex mobile data networks.

Example 4.13: Dynamic Slotted ALOHA

Mobitex has a full-duplex communication capability (simultaneous transmissions on the uplink and downlink) and employs a *dynamic* slotted ALOHA protocol. Suppose that there are three mobile stations MS_1, MS_2, and MS_3 in a cell. The situation is such that the BS has two messages to send to MS_3, MS_1 has a short status update that requires one slot, MS_2 has a long message to send, and MS_3 has nothing to transmit. An MS can transmit only during certain "free" cycles consisting of several slots of equal length that are periodically initiated by the BS using a FREE frame on the downlink. In this example, shown in Figure 4.11, the BS indicates that there are six free slots for contention, each of a certain length. This can change depending on the traffic; hence the term "dynamic." Also note that MSs cannot transmit whenever they want as in slotted ALOHA. The MSs with traffic to send, such as, MS_1 and MS_2, select one of the six slots at random. In this case, MS_1 selects slot 1 and MS_2 selects slot 4. Hence there is no collision. MS_1 is able to transmit its short status update in slot 1 after which it ceases transmission. MS_2 transmits in slot 4 requesting access to the channel. Simultaneously, the BS would have transmitted its message to MS_3. Upon receipt of the message, MS_3 acknowledges it. The free slots are designed to be of the duration of the downlink message to MS_3 so that the acknowledgment from MS_3 can be received without contention. The BS also acknowledges the status report from MS_1 and sends an access grant to MS_2. As MS_2 transmits its long message on the uplink, the BS can simultaneously send the second message to MS_3. After proper ACKs are transmitted and received, a new FREE cycle is started.

Example 4.14: Packet Reservation Multiple Access (PRMA)

An example of a system that uses reservation for integrating voice and data services is the work done by David Goodman and his colleagues in developing the concept of packet reservation multiple access (PRMA) [GOO89], [GOO91]. PRMA is a method for transmitting, in a wireless environment, a variable mixture of voice packets and data packets. The PRMA system is closely related to R-ALOHA, in that it merges characteristics of slotted ALOHA and TDMA protocols. PRMA has been developed for use in centralized networks operating over short-range radio channels. Short propagation times are an important ingredient in providing acceptable delay characteristics for voice service.

The transmission format in PRMA is organized into frames, each containing a fixed number of time slots. The frame rate is identical to the arrival rate of speech packets. The terminals identify each slot as either "reserved" or "available" in accordance with a feedback message received from the base station at the end of the slot. In the next frame, only the user terminal that reserved the slot can use a reserved slot. Any terminal not holding a reservation that has information to transmit can use an available slot.

Terminals can send two types of information, referred to as periodic and random. Speech packets are always periodic. Data packets can be random, if they are

isolated, or periodic if they are contained in a long unbroken stream of information. One bit in the packet header specifies the type of information in the packet. A terminal having periodic information to send starts transmitting in contention for the next available time slot. Upon successfully detecting the first packet in the information burst, the base station grants the sending terminal a reservation for exclusive use of the same time slot in the next frame. The terminal in effect "owns" that time slot in all succeeding frames as long as it has an unbroken stream of packets to send. After the end of the information burst, the terminal sends nothing in its reserved slot. This in turn causes the base station to transmit a negative acknowledgment feedback message indicating that the slot is once again available.

To transmit a packet, a terminal must verify two conditions. The current time slot must be available, and the terminal must have permission to transmit. Permission is granted according to the state of a pseudorandom number generator, permissions at different terminals being statistically independent. The terminal attempts to transmit the initial packet of a burst until the base station acknowledges successful reception of the packet or until the terminal discards the packet, because it has been held too long. The maximum holding time, D_{max} seconds, is determined by delay constraints on speech communication, and is a design parameter of the PRMA system. If the terminal drops the first packet of a speech burst, it continues to contend for a reservation to send subsequent packets. It drops additional speech packets as their holding times exceed the limit D_{max}. Terminals with data packets store packets indefinitely while they contend for slot reservations (equivalent to setting D_{max} to infinity). Thus as a PRMA system becomes congested, both the speech packet dropping rate and the data packet delay increase.

In [GOO91] Goodman and Wei analyze PRMA efficiency, which they quantify as the maximum number of conversations per channel that the system can support within a chosen constraint on packet-dropping probability. In their work they adopted a constraint of $P_{drop} < 0.01$. They used a speech source rate of 32 kbps and a header length of 64 bits in each packet. Using computer simulation methods, they investigated the effects of six system variables on PRMA efficiency: (1) channel rate, (2) frame duration, (3) speech activity detector, (4) maximum delay, (5) permission probability, and (6) number of conversations. Over the range of conditions examined, they found many PRMA configurations capable of supporting about 1.6 conversations per channel and found that this level of efficiency could be maintained over a wide range of conditions.

Example 4.15: Reservation in GPRS

A single 200 kHz carrier in GSM has eight time slots, each capable of carrying data at 9.6 kbps (standard), 14.4 kbps (enhanced), or 21.4 kbps (if forward error correction is completely omitted). The raw data rate can thus be as high as $8 \times 21.4 = 171.2$ kbps. The same time slots can be reserved for data access using slotted ALOHA. Medium access is based on a slotted ALOHA reservation protocol. In the contention phase, a slotted ALOHA random access technique is used to transmit reservation requests, the BS then transmits a notification to the MS indicating the channel allocation for an uplink transmission, and finally the MS can transfer data on the allocated slots without contention. On the downlink,

the BS transmits a notification to the MS, indicating the channel allocation for downlink transmission of data to the MS. The MS will monitor the indicated channels, and the transfer occurs without contention.

4.3.1.2 CSMA-Based Wireless Random Access Techniques

The main drawback of ALOHA based contention protocols is the lack of efficiency caused by the collision and retransmission process. In ALOHA, users do not take into account what other users are doing when they attempt to transmit data packets and there are no mechanisms to avoid collisions. A simple method to avoid collisions is to sense the channel before transmission of a packet. If there is another user transmitting on the channel, it is obvious that a terminal should delay the transmission of the packet. Protocols employing this concept are referred to as CSMA or listen-before-talk (LBT) protocols. Figure 4.12 shows the basic concept of the CSMA protocol. Terminal "1" will sense the channel first and then sends a packet. This is followed by a sensing and packet transmission by terminal "1" again. During the second transmission time of terminal "1," terminal "2" senses the channel and discovers that another terminal is using the medium. It then delays its transmission for a later time using a back-off algorithm. The CSMA protocol reduces the packet collision probability significantly compared with ALOHA protocol. However, it cannot eliminate the collisions entirely. Sometimes, as shown in Figure 4.12, two terminals sense the channel busy and reschedule their packets for a later time but their transmission time overlap with each other causing a collision. Such situations do not cause a significant operational problem because the collisions can be handled in the same way as they were handled in ALOHA. However, if the propagation time between the terminals is very long, such situations happen more frequently, thereby reducing the effectiveness of carrier sensing in preventing collisions. As a result, several variations of CSMA have been employed in local area applications, whereas ALOHA protocols are preferred in wide area applications.

Example 4.16: Examples of Wireless Networks that Employ ALOHA and CSMA

As described earlier, Mobitex uses a variation of ALOHA protocol, and the IEEE 802.11 standard for WLANs employs a version of CSMA protocol. ALOHA protocol is also used in the random access logical channels in cellular telephone and satellite communication applications.

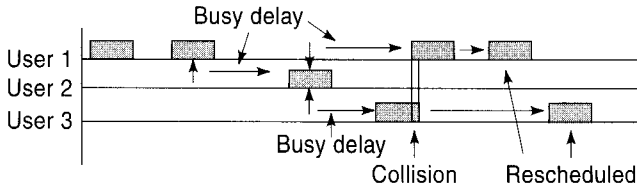

Figure 4.12 Basic operation of CSMA protocol.

A number of strategies are used for the sensing procedure and retransmission mechanisms that have resulted in a number of variations of the CSMA protocol for a variety of wired and wireless data networks. Figure 4.13 depicts the key elements of distinction among these protocols. If after sensing the channel, the terminal attempts another sensing only after a random waiting period the carrier-sensing mechanism is called "nonpersistent." After sensing a busy channel, if the terminal continues sensing the channel until the channel becomes free, the protocol is referred to as a "persistent." In persistent operation, after the channel becomes free, if the terminal transmits its packet immediately, it is referred to as "1-persistent" CSMA, and if it runs a random number generator and based on the outcome transmit its packet with a probability "p," the protocol is called p-persistent CSMA.

In a wireless network, due to multipath and shadow fading as well as the mobility of terminals, sensing the availability of the channel is not as simple as in the case of wired channels. Typically in a wireless network, two terminals can each be within range of some intended third terminal but out of range of each other, because they are separated by excessive distance or by some physical obstacle that makes direct communication between the two terminals impossible. This situation, where the two terminals cannot sense the transmission of each other, but a third terminal can sense both of them, is referred to as the *hidden terminal problem.* This is a more likely situation in cases of radio networks covering wider geographic areas in which hilly terrain blocks some groups of user terminals from sensing other groups. In this situation the CSMA protocol will successfully prevent collisions among the users of one group but will fail to prevent collisions between users in groups hidden from one another.

To resolve the hidden terminal problem, we need to facilitate the sensing procedure. In multihop ad hoc networks, where there is no centralized station or infrastructure, a protocol called *busy-tone multiple-access* (BTMA) has been used in packet radio for military applications. A brief summary of BTMA is given in [TOB80], where a number of packet communication protocols are discussed and compared. In the BTMA scheme, the system bandwidth is divided into two channels, a *message channel* and a *busy-tone channel.* Whenever a station sends signal

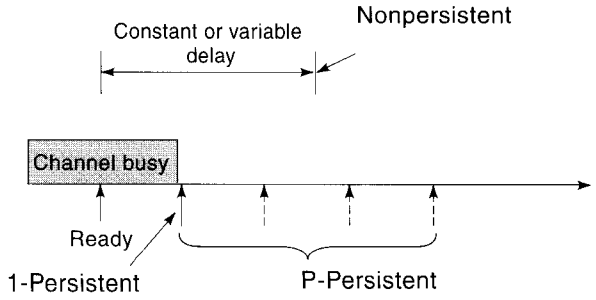

Figure 4.13 Retransmission alternatives for CSMA [STA00].

energy on the message channel, it transmits a simple busy-tone signal (e.g., a sinusoid) on its busy-tone channel. When any other terminal senses a busy-tone signal, it turns on its own busy tone. In other words, as a terminal detects that some user is on the message channel, it sounds the alarm on the busy tone channel in an attempt to inform every user, including those hidden to the transmitting terminal. A user station with a packet ready to send first senses the busy-tone channel to determine if the network is occupied.

Most cellular mobile data networks use different frequencies for forward (downlink) and reverse (uplink channels). The messages in the forward channel are transmitted from the mobile data base station that is designed and deployed to provide a comprehensive and reliable coverage. In another words, the base stations are not hidden to the mobile terminals, whereas the mobile terminals may be hidden from one another. In this situation one may use the forward channel to announce the availability of the channel for the mobile terminals. This concept is used in a protocol referred to as *digital* or *data sense multiple-access* (DSMA). DSMA is very popular in mobile data networks, and it is used in CDPD, ARDIS, and TETRA. In DSMA, the forward channel broadcasts a periodic busy-idle bit announcing availability of reverse channel for data transmission. A mobile terminal checks the busy-idle bit prior to transmission of its packet. As soon as the mobile station starts its transmission, the base station will change the busy-idle bit to the busy state to prevent other mobile terminals from transmission. Because the sensing process is performed after demodulation of data from the digital information, it is referred to digital or data sense, rather than carrier sense, multiple access.

4.3.2 Access Methods for Wireless LANs

Compared with a WAN, a LAN operates over shorter distances with smaller propagation delays and consequently a transmission medium that is well suited for variations of the CSMA protocol. Low-speed WANs are developed for communicating shorter messages while local area networks are designed to facilitate large file transfers at high data rates. As a result, the length of the packets in LANs is much larger than the length of the packets in low-speed mobile data networks. When the length of the packets is long, it would be very useful to pay further attention to packet collisions. LANs often employ variations of the CSMA protocol that either stop transmission as soon as a packet collision is detected or add additional features to the avoid collision.

Problem 10: Packet Sizes in Wireless Data Networks

a) Determine the transfer time of a 20 kB file with a mobile data network with a transmission rate of 10 kbps.

b) Repeat for an 802.11 WLAN operating at 2 Mbps.

c) What is the length of the file that the WLAN of part (b) can carry in the time that mobile data service of part (a) carries its 20 kB file?

Solution:

a) The early mobile data networks, such as ARDIS and Mobitex, limited the length of a file to around 20 kB. For a data rate of around 10 kbps it would take 20 (kB) × 8 (B/b) × 10 Mbps = 16 seconds to transfer such a file.

b) An IEEE 802.11 network operating at 2 Mbps would transfer this file in 80 ms.

c) In a 16-second time interval, the same WLAN transfers a 4 MB file.

The most popular version of the CSMA for wired LANs is CSMA with collision detection (CSMA/CD) adopted in the IEEE 802.3 (Ethernet) standard, the dominant standard for wired LANs supporting data rates that can be up to several gigabits per second. The basic operation of CSMA/CD is the same as CSMA implementations discussed earlier. The defining feature of CSMA/CD is that it provides for detection of a collision shortly after its onset, and each transmitter involved in the collision stops transmission as soon as it senses a collision. In this way colliding packets can be aborted promptly, minimizing the wastage of channel occupancy time by transmissions destined to be unsuccessful. Unlike CSMA, which requires an acknowledgment (or lack of an acknowledgment) to learn the status of a packet collision, CSMA/CD requires no such feedback information, because the collision-detection mechanism is built into the transmitter. When a collision is detected, the transmission is immediately aborted, a jamming signal is transmitted, and a retransmission back-off procedure is initiated, just as in CSMA [PAH95]. As is the case with any random access scheme, proper design of the back-off algorithm is an important element in assuring stable operation of the network.

Example 4.17: Binary Exponential Back-off

The back-off algorithm recommended by IEEE 802.3 Ethernet is referred to as the binary exponential back-off algorithm that is combined with 1-persistence CSMA protocol with collision detection. When a terminal senses a transmission, it continues sensing (persistent) until the transmission is completed. After the channel becomes free, the terminal sends its own packet. If another terminal was also waiting, a collision occurs because of the 1-persistence, and the two terminals reattempt transmission with a probability of ½ after a time slot that spans twice the maximum propagation delay allowed between the two terminals. A time slot that spans twice the maximum propagation delay is selected to ensure that in the worse case scenario the terminal will be able to detect the collision. If a second collision occurs, the terminals reattempt with the probability of ¼ that is half of the previous retransmission probability. If collision persists, the terminal continues reducing its retransmission probability by half up to 10 times, and after that it continues with the same probability six more times. If no transmission is possible after 16 attempts, the MAC layer reports to the higher layers that the network is congested and transmission shall be stopped. This procedure exponentially increases the back-off time and gives the back-off strategy its name. The disadvantage of this procedure is that the packets arriving later have a higher chance to survive the collision that results in an unfair first-come, last-serve environment. It can be shown that the average waiting time for the exponential back-off algorithm is $5.4T$ where T is the time slot used for waiting [TAN97], [STA00].

The CSMA/CD scheme is used in many IR-based LANs, where both transmission and reception are inherently directional. In such an environment, a transmitting station can always compare the received signal from other terminals with its own transmitted signal to detect a collision. Radio propagation is not directional, posing a serious problem in determining other transmissions during your own transmission. As a result, collision detection mechanisms are not well suited for radio LANs. However, compatibility is very important for WLANs, and, therefore, designers of these networks have had to consider CSMA/CD for compatibility with the Ethernet backbone LANs that dominate the wired-LAN industry.

Although collision detection is easily performed on a wired network, simply by sensing voltage levels against a threshold, such a simple scheme is not readily applicable to radio channels because of fading and other radio channel characteristics. The one approach that can be adopted for detecting collisions is to have the transmitting station demodulate the channel signal and compare the resulting information with its own transmitted information. Disagreements can be taken as an indicator of collisions, and the packet can be immediately aborted. However, on a wireless channel the transmitting terminal's own signal dominates all other signals received in its vicinity, and thus the receiver may fail to recognize the collision and simply retrieve its own signal. To avoid this situation, the station's transmitting antenna pattern should be different from its receiving pattern. Arranging this situation is not convenient in radio terminals because it requires directional antennas and expensive front-end amplifiers for both transmitters and receivers.

The approach called CSMA/CA, shown in Figure 4.14, is actually adopted by the IEEE 802.11 wireless LAN standard. The elements of CSMA/CA used in the IEEE 802.11 are interframe spacing (IFS), contention window (CW), and a back-off counter. The CW intervals are used for contention and transmission of the packet frames. The IFS is used as an interval between two CW intervals. The back-off counter is used to organize the back-off procedure for transmission of packets. The method of operation is best described by an example.

Example 4.18: Operation of Collision Avoidance in IEEE 802.11

Figure 4.15 provides an example for the operation of the CSMA/CA mechanism used in the IEEE 802.11 standard. Stations A, B, C, D, and E are engaged in contention for transmission of their packet frames. Station A has a frame in the air when stations B, C, and D sense the channel and find it busy. Each of the three stations will run its random number generator to get a back-off time by random. Station C followed by D and B draws the smallest number. All three terminals

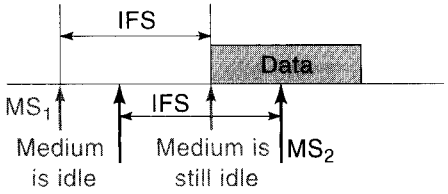

Figure 4.14 CSMA/CA adopted by the IEEE 802.11.

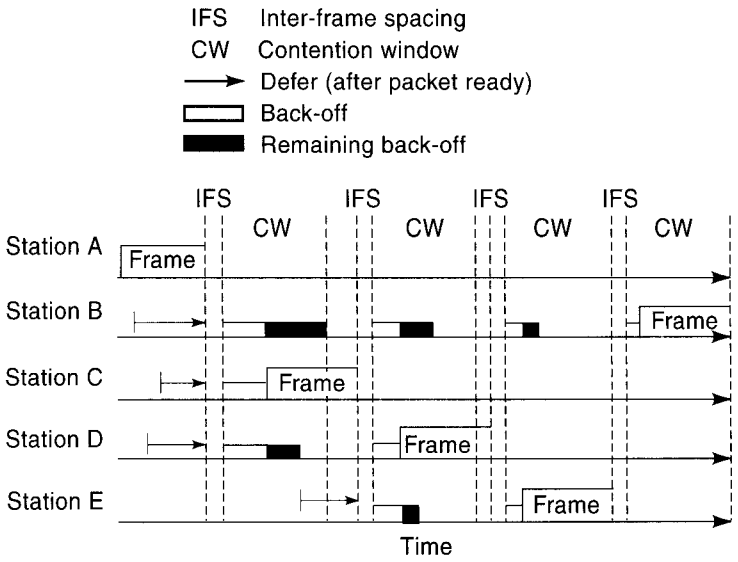

Figure 4.15 Illustration of CSMA/CA.

persist on sensing the channel and defer their transmission until the transmission of the frame from terminal A is completed. After completion, all three terminals wait for the IFS period and start their counters immediately after completion of this period. As soon as the first terminal, station C in this example, finishes counting its waiting time, it starts transmission of its frame. The other two terminals, B and D, freeze their counter to the value that they have reached at the start of transmission for terminal C. During transmission of the frame from station C, station E senses the channel, runs its own random number generator that in this case ends up with a number larger than the remainder of D and smaller than the remainder of B, and defers its transmission for after the completion of station C's frame. In the same manner as the previous instance, all terminals wait for IFS and start their counter. Station D runs out of its random waiting time earlier and transmits its own packet. Stations B and E freeze their counters and wait for the completion of the frame transmission from terminal D and the IFS period after that before they start running down their counters. The counter for terminal E runs down to zero earlier, and this terminal sends its frame while B freezes its counter. After the IFS period following completion of the frame from station E, the counter in station B counts down to zero before it sends its own frame. The advantage of this back-off strategy over the exponential back-off used in IEEE 802.3 is that the collision detection procedure is eliminated and the waiting time is fairly distributed in a way that on average enforces a first-come, first-serve policy.

[WIL95b]

Another related technique considered for collision avoidance in wireless LANs is the *combing* method [WIL95]. As shown in Figure 4.16, the time is divided into comb and data transmission intervals. During the comb period, each station alternates between transmission and listening periods according to a code assigned to the station. All stations will continue advancing in their code until they sense a carrier during their listening period. If they do not sense a carrier at the end of the

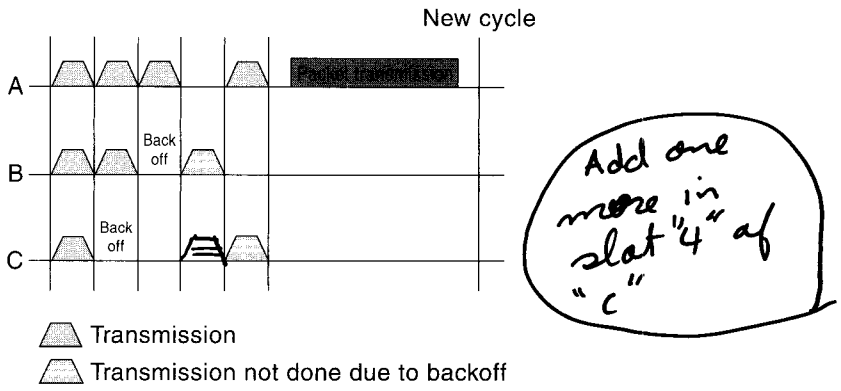

Figure 4.16 Illustration of combing.

code, they transmit their packet. If they sense a carrier, they postpone their transmission until the next comb interval. A simple example will further clarify this method.

Example 4.19: Combing for Collision Avoidance

Figure 4.16 shows three stations with five digit codes 11101 (terminal A), 11010 (terminal B), and 10011 (terminal C). All three terminals transmit their carrier during the first slot, because all codes have 1 in that slot. In the second slot, terminal C will listen and after sensing the carriers of other two terminals withdraws from contention and waits for the next combing. In the third period station, B goes to listening state, and after sensing the carrier of terminal A defers its transmission until the next cycle. Terminal A continues its sequence of alternating transmissions and listening until the end of the comb period when it transmits its data packet (as it heard no other terminal). After completion of the data transmission from station A, the other two terminals will wait for an interpacket spacing for a new contention after which station B transmits its packet. Station C will transmit after the second transmission cycle.

In CSMA/CA, as we will see later, priority is assigned by dividing the IFS into several different sized intervals associated with different priority levels. In combing, priority can be arranged by assigning different classes of numbers to the codes. The lower priority packets will receive earlier zero codes, and higher priority packets will have a zero in their codes in the later intervals.

Another access method used in wireless LANs is the request-to-send/ clear-to-send (RTS/CTS) mechanism shown in Figure 4.17. A terminal ready for transmission sends a short RTS packet identifying the source address, destination address, and the length of the data to be transmitted. The destination station will respond with a CTS packet. The source terminal will send its packet with no contention. After acknowledgment from the destination terminal, the channel will be available for other usage. IEEE 802.11 supports this feature as well as CSMA/CA. This method provides a unique access right to a terminal to transmit without any contention.

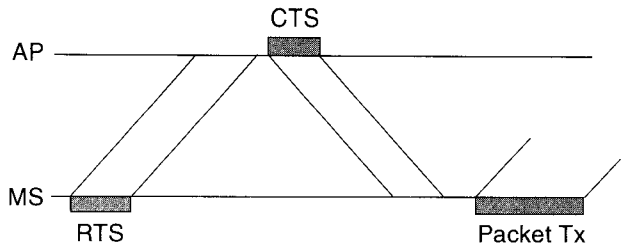

Figure 4.17 RTS/CTS in IEEE 802.11.

4.3.3 Performance of Random Access Methods

In voice-oriented circuit switched networks, performance is measured by the probability of blockage (blockage rate) of initiating a call. If the call is not blocked, a fixed rate full-duplex channel is allocated to the user for the entire communication session. In other words, interaction between the user and the network takes place in two steps. First, during the call establishment procedure, the user negotiates the availability of a line with the network, and if successful (not blocked), the network guarantees a connection with a certain QoS (data rate, delay, error rates) to the user. For real-time interactive applications such as telephone conversations or video conferencing, if the user does not talk, the resource allocated to the user is wasted. If these facilities, originally designed for two-way voice application, are used for data application then (1) for bursty data file transfers during the idle times between transmission of two packet bursts allocated resources are wasted, and (2) large file transfers suffer a long delay or waiting time for the transfer because resources allocated to each user is more restricted.

Users of packet switched networks are always connected, and there is neither an initiation (negotiation) procedure to be blocked nor a fixed QoS to be allocated. In this situation, analysis of the performance for real-time interactive applications such as telephone conversations is complicated and will be addressed later. Performance of packet switched networks for data applications is often measured by the average throughput, S, and average delay, D, versus the total offered traffic, G. The *channel throughput, S,* is the average number of successful packet transmissions per time interval T_p. The offered traffic, G, is the number of packet transmission attempts per packet time slot T_p that includes new arriving packets, as well as retransmissions of old packets. The average delay D is the average waiting time before successful transmission, normalized to the packet duration T_p. The standard unit of traffic flow is Erlang, which can be thought of as the number of the packets per packet duration time T_p. The throughput is always between zero and one Erlang, whereas the offered traffic, G, may exceed one Erlang.

The analyses of the relationships between S, G, and D for a variety of medium access protocols have been a subject of research for a few decades. This analysis depends on the assumptions on the statistical behavior of the traffic, number of terminals, relative duration of the packets, and the details of the implementation. Assuming a large number of terminals generating fixed length packets with a Poisson distribution,[2] Table 4.1 summarizes the throughput expressions for

[2]Poisson distribution assumes packets are generated independent from one another, and the interarrival time between the packets forms an exponentially distributed random variable.

Table 4.1 Throughput of Various Random Access Protocols

Protocol	Throughput
Pure ALOHA	$S = Ge^{-2G}$
Slotted ALOHA	$S = Ge^{-G}$
Unslotted 1-persistent CSMA	$S = \frac{G[1 + G + aG(1 + G + aG/2)]e^{-G(1+2a)}}{G(1 + 2a) - (1 - e^{-aG}) + (1 + aG)e^{-G(1+a)}}$
Slotted 1-persistent CSMA	$S = \frac{G[1 + a - e^{-aG}]e^{-G(1+a)}}{(1 + a)(1 - e^{-ag}) + ae^{-G(1+a)}}$
Unslotted nonpersistent CSMA	$S = \frac{Ge^{-aG}}{G(1 + 2a) + e^{-aG}}$
Slotted nonpersistent CSMA	$S = \frac{aGe^{-aG}}{1 - e^{-aG} + a}$

ALOHA, and 1-persistent and nonpersistent CSMA protocols, including the slotted and unslotted versions of each. The expressions for p-persistent protocols are very involved and are not included here. The interested reader should refer to [KLE75b], [TOB75], and [TAK85], where the derivations of the other CSMA expressions can also be found. The expressions in the table are also derived in [HAM86] and [KEI89]. The parameter a in this table corresponds to the normalized propagation delay defined as $a = \tau/T_p$, where τ is the maximum propagation delay for the signal to go from one end of the network to the other end.

Problem 11: Calculation of the Normalized Propagation Delay

Determine the parameter a in IEEE 802.3 (Ethernet) 10 Mbps LANs and IEEE 802.11 2 Mbps LANs.

Solution:

The IEEE 802.3 standard for star LANs allows a maximum length of 200 m between two terminals. The propagation speed in the cables is usually approximated by 200,000 km/s, resulting in $\tau = 1\mu s$. The IEEE 802.11 allows maximum distance of 100 meters between the AP and the MS. The radio propagation is at the rate of 300,000Km/s, resulting in $\tau = 0.33$ μs. For a star LAN operating at 10 Mbps with 1,000 bit packets, the value is $a = 0.01$. For an IEEE 802.11 operating at 2 Mbps with the same packet size $a = 0.00066$.

Figure 4.18 shows plots of throughput S versus offered traffic load G for the six protocols listed in Table 4.1, with a normalized propagation delay of $a = 0.01$. All curves follow the same pattern. Initially as the offered traffic G increases the throughput, S also increases up to a point where it reaches a maximum S_{max}. After the throughput reaches its maximum value, an increase in the offered traffic actually reduces the throughput. The first region depicts the stable operation of the network in which an increase in aggregating traffic G, which includes arriving traffic as well as retransmissions due to collisions, increases the total successful transmissions and thus S. The second region represents unstable operation where an increase in G actually reduces the throughput S because of congestion and the eventually halting of the operation. In

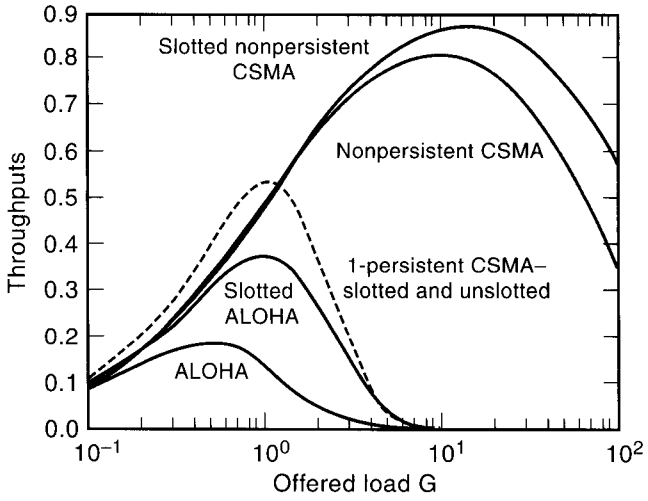

Figure 4.18 Throughput S versus offered traffic load G for various random access protocols.

practice, as we saw in the last section, retransmission techniques adopted for the real implementation include back-off mechanisms to prevent operation in unstable regions.

The throughput curves for the slotted and unslotted versions of 1-persistent CSMA are essentially indistinguishable. It can be seen from the figure that for low levels of offered traffic, the 1-persistent protocols provide the best throughput, but at higher load levels the nonpersistent protocols are by far the best. It can also be seen that the slotted nonpersistent CSMA protocol has a peak throughput almost twice that of persistent CSMA schemes.

The equations in Table 4.1 can also be used to calculate capacity, which is defined as the peak value S_{max} of throughput over the entire range of offered traffic load G [HAM86]. An example is helpful to show how to relate the curve to a particular system.

Problem 12: Relating Throughput and Offered Traffic to Data Rates

To relate throughput and offered traffic to data rates, assume that we have a centralized network that supports a maximum data rate of 10 Mbps and serves a large set of user terminals with the pure ALOHA protocol.

a) What is the maximum throughput of the network?

b) What is the offered traffic in the medium and how is it composed?

Solution:

a) Because the peak value of the throughput is $S = 18\%$ the terminals contending for access to the central module can altogether succeed in getting at most 1.8 Mbps of information through the network.

b) At that peak the total traffic from the terminals is 5 Mbps (because the peak occurs at $G = 0.5$), which is composed of 1.8 Mbps of successfully delivered packets (some mixture of new and old packets) and 3.2 Mbps of packets doomed to collide with one another.

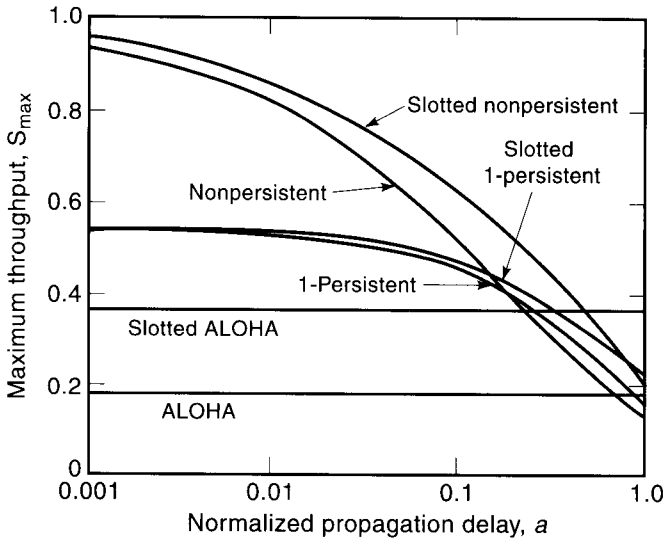

Figure 4.19 Capacity versus normalized propagation delay for various random access protocols.

Plots of capacity versus normalized propagation delay are plotted in Figure 4.19 for the same set of ALOHA and CSMA schemes. The curves show that for each type of protocol the capacity has a distinctive behavior as a function of normalized propagation delay a. For the ALOHA protocols, capacity is independent of a, and is the largest of all the protocols (compared when a is large). As we discussed earlier, this is the case where the area of coverage is large and propagation delays are comparable to the length of packets. The plots in Figure 4.19 also show that the capacity of 1-persistent CSMA is less sensitive to the normalized propagation delay for small a, than is nonpersistent CSMA. However, for small a, nonpersistent CSMA yields a larger capacity than does 1-persistent CSMA, though the situation reverses as a approaches the range of 0.3 to 0.5 [HAM86].

Another important performance measure for packet data communications is the delay characteristics of the transmitted packets. For real-time applications and voice conversations, if the delay is more than a certain value (several hundred milliseconds), the packet is not useful, and it is dropped. Therefore, we need to analyze the delay characteristics of the channel to determine the capacity of the access method. In the data transfer applications, the delay characteristic is usually related to the throughput of the medium, and it usually follows a hockey-stick shape. At low traffic when a small fraction of the maximum throughput is utilized, the delay often remains the same as the transmission delay. As the throughput increases, the number of retransmitted packets increases, resulting in higher average delay for the packets. Around the maximum throughput, the delay retransmissions grow rapidly, pushing the network toward unstable condition where the channel is dominated with retransmissions, and the packet delays grow extremely large. Figure 4.20 shows the delay-throughput behavior of the ALOHA, slotted ALOHA, and CSMA protocols [TAN00].

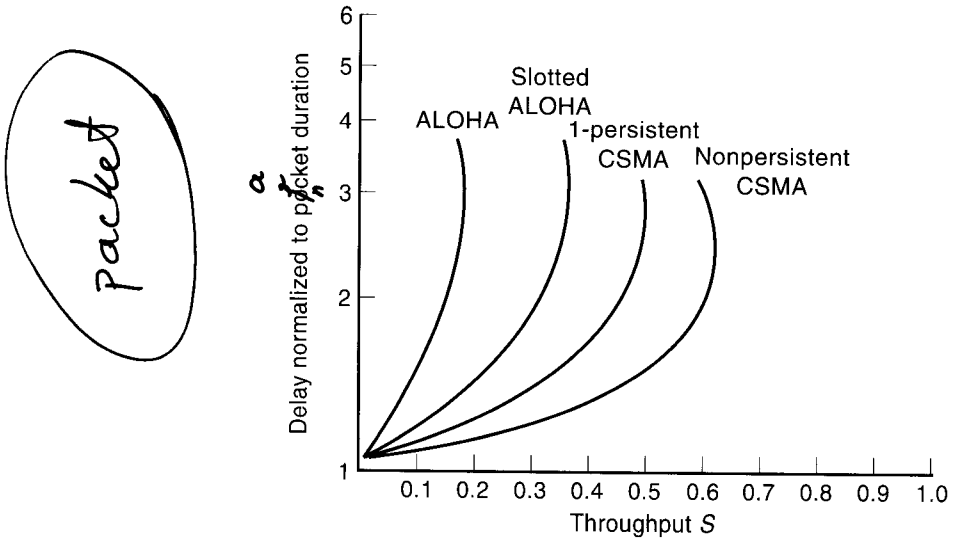

Figure 4.20 Delay-throughout behavior of random access protocols.

4.3.3.1 Practical Considerations

The analysis provided in the previous chapter is abstract and is used to provide an intuitive framework for the operation of different classes of access methods. In practice implementations deviate considerably from the abstract, and the performance is evaluated by analysis or simulation of case-by-case situations. Examples of this type of analysis for the CSMA/CD with exponential back-off algorithm used in the IEEE 802.3 Ethernet and performance of token ring access method used in IEEE 802.5 are available in the last chapter of [STA00].

4.3.3.2 Complications Caused by Wireless Channel

Three factors that are effective in throughput analysis in wired environment are propagation delay, users' idle period (not transmitting), and packet collisions. In a wireless environment, analysis of the real throughput of a protocol is much more complicated because it involves hidden terminal and capture effects. To analyze these effects, let's assume we have a centralized AP with a number of terminals connected to it, which communicate via a random access method.

Figure 4.21 demonstrates the basic concept behind the hidden terminal problem. The two terminals contending to communicate with the AP are both in the coverage area of the AP, but they are out of the coverage area of each other. Limited antenna range and shadowing are two major causes for hidden terminal degradation. The hidden terminal problem does not effect the performance of the ALOHA type protocols, but it degrades the performance of CSMA protocols. In a CSMA environment effected by the hidden terminal problem, some terminals cannot sense the carrier of the transmitting terminal and their transmitted packets have a higher probability of colliding and degrading the overall throughput.

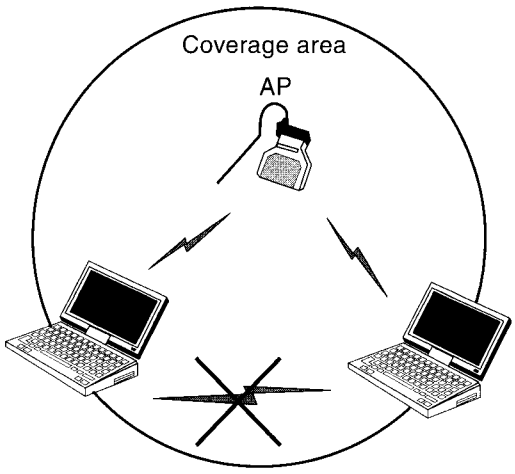

Figure 4.21 Hidden terminal problem.

In real installations, the coverage area of the AP is usually larger than that the mobile terminals, because the AP is installed in a selected location to optimize the coverage (high on the walls or on the ceiling), which will increase the negative impacts of the hidden terminal problem. Assuming that the coverage area of the AP and the mobile terminals are the same, there is still no guarantee that all the terminals in the coverage area of the AP can hear one another. This is because two terminals at the maximum distance L from the AP could be as far as $2L$ apart. Therefore, the hidden terminal problem is unavoidable and natural to the operation of the centralized access systems using CSMA protocol that are common in WLAN operations.

Another phenomenon impacting the throughput of a radio network is capture. In radio channels, sometimes collision of two packets may not destroy both packets. Because of signal fading or the near-far effect, packets from different transmitting stations can arrive with different power levels, and the strongest packet may survive a collision. Figure 4.22 shows the basic concept of the capture phenomenon. The received power from the terminal closer to the access point is much larger than the received power from the terminal located at a distance. If two packets collide in time, the packet with the weaker signal will appear as a background noise, and the AP captures (detects) the packet from the closer terminal successfully. The capture effect increases the throughput of the radio network because in calculating the throughput, we always assume that the colliding packets are destroyed (not detected).

The hidden terminal effects were first analyzed for different types of CSMA protocols used in rapidly moving packet radio networks for military applications, and busy tone signaling was suggested for eliminating the hidden terminal problem. More recently, there have been efforts to analyze the effects of capture and hidden terminals in a WLAN environment using various assumptions [ZHA92], [ZAH97].

In reality, the capture of a packet is a random process, which is a function of the modulation technique used for transmission, received signal-to-noise ratio, and the length of the packet.

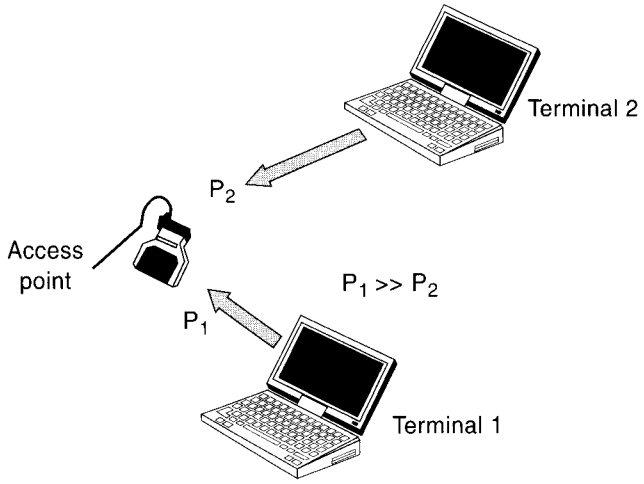

Figure 4.22 The capture effect.

Example 4.20: Capture Effect and Throughput

Figure 4.23 [ZHA92] shows the effects of capture on the throughput of the conventional slotted ALOHA and the CSMA systems for a variety of packet lengths of 16, 64, and 640 bits. Also shown for comparison are the curves for conventional nonpersistent CSMA and slotted ALOHA without capture. With capture, the maximum throughput of CSMA with packet length 16 bits is 0.88 Erlang, which is 0.065 Erlang more than the case without capture. The maximum throughput for

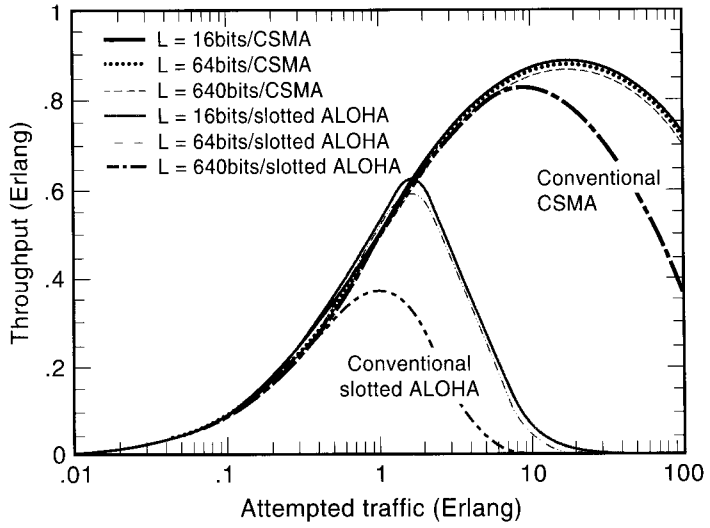

Figure 4.23 Effects of packet length on throughput for CSMA and slotted ALOHA with capture. The modulation is BPSK, and the SNR = 20 dB.

slotted ALOHA with the same packet length is 0.591 Erlang, which is 0.231 Erlang higher than the case without capture.

In slow fading channels, if the terminal generating the test packet is in a "good" location, the interference from other packets is small, and all the bits of the test packet survive the collision. In contrast, for a test packet originating from a terminal in a "bad" location, all the bits are subject to high probability of error, and the packet does not survive the collision. As a result, the system shows minimal sensitivity to the choice of packet length, which is consistent with our assumption of slow fading. Figure 4.24 [ZHA92] shows the delay-throughput for the CSMA protocol with and without capture for a 640-bit packet network. The packet delay is normalized to the length of the packet. For both cases as the throughput reaches around its maximum value system becomes rapidly unstable, causing unacceptable delays for the delivery of the packets. When the capture effects are included, maximum throughput is increased, and the instability occurs at a slightly higher value of the throughput.

Example 4.21: Capture and Hidden Terminals in WLANs

Figure 4.25 [ZAH97] shows the throughput versus offered traffic curves for a WLAN access point using CSMA protocol and surround by a large number of terminals uniformly distributed within the AP's coverage area. In this scenario, as shown in Figure 4.26 each terminal senses a group of terminals within its coverage area (area I in the figure) and cannot sense those that are out of its coverage area but are still within the coverage area of the AP (area II in the figure). The

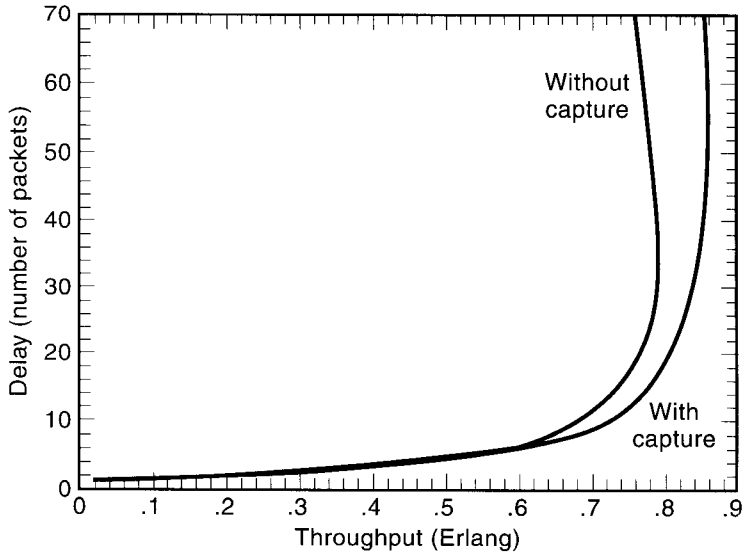

Figure 4.24 Delay versus throughput of CSMA for BPSK modulation and SNR = 20 dB, with and without capture.

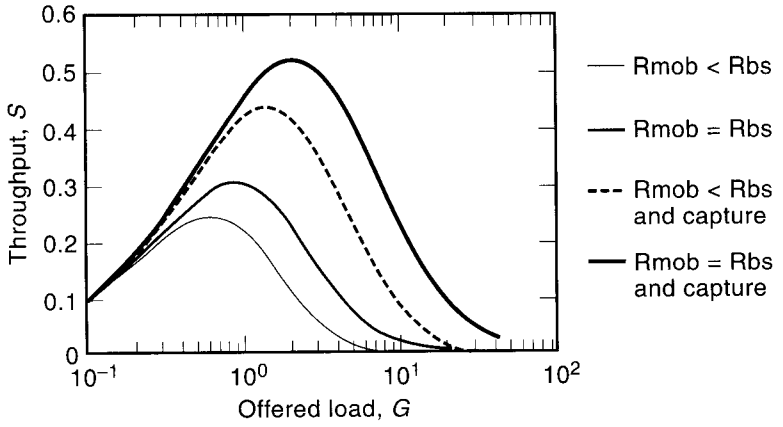

Figure 4.25 Throughput versus offered traffic for a wireless LAN with a large number of terminals.

throughput of the target terminal with respect to terminals in area I is the same as the throughput of a CSMA system. However, the throughput of the target terminal with respect to terminals in area II is the same as the throughput of ALOHA networks because carrier sensing does not work and the terminals transmit their packets without knowledge of the transmission from terminals in area II. Using these facts in the throughput at each point in the area of the coverage of the AP is calculated [ZHA96]. Then it is averaged over the entire coverage area of the AP for different coverage areas for the mobile terminals. Obviously, this average throughput will always remain between the throughput of CSMA and that of ALOHA. The lowest curve in Figure 4.25 shows the throughput when the hidden terminal problem is considered, and the coverage (radius R_{mob}) of each terminal is 70 percent of the coverage (radius *rbs*) of the AP. Because a number of termi-

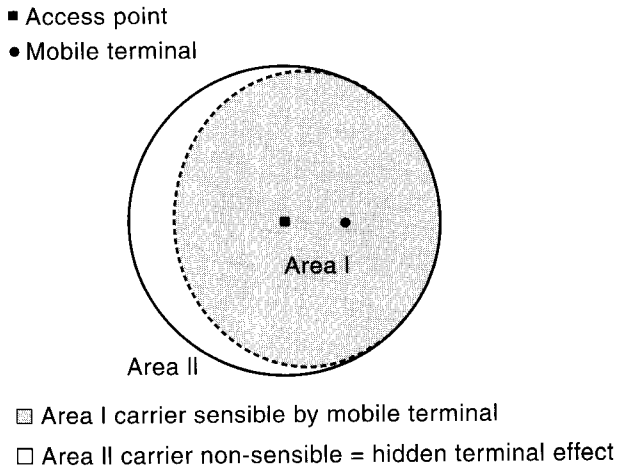

Figure 4.26 Coverage areas of an AP and a tagged mobile terminal in a WLAN.

nals cannot sense the transmission of others, the peak throughput has declined to less than 25 percent which is slightly higher than the 18 percent maximum throughput of the ALOHA and far below the maximum of CSMA. The second curve from below, with a peak value of around 30 percent, represents the same results where the coverage of the AP and the mobile terminals are the same. The third curve depicts the performance when both effects of the hidden terminal and capture are considered, and the coverage of the mobile terminals is smaller than that of the AP. The capture effect increases the throughput to more than 40 percent. The top curve is the same as the third curve where the coverage of the AP and the mobile terminals are the same. This situation has increased the throughput by another 10 percent to above 50 percent, which is getting closer to the performance of conventional CSMA.

4.4 INTEGRATION OF VOICE AND DATA TRAFFIC

4.4.1 Access Methods for Integrated Services

As the wireless communications industry moves toward 3G and 4G networks, one of the important objectives is the use of a single wireless system for multimedia applications to support a variety of communications services, including voice, data, and voice in various forms and combinations. A key technical problem to be dealt with in such integrated systems is that of multiuser access. As we saw earlier in this chapter, an access method that efficiently supports one category of service may be unsuitable for another category of service. In the 1G and 2G networks, as we saw in Chapter 1, the wireless industry evolved around two separate paths for voice- and data-oriented applications. If a data service can be efficiently integrated with a voice service, transmission resources that are otherwise wasted (because there is no voice transmission) can be used for data, which typically do not have stringent delay requirements. First, the voice-oriented networks evolved into supporting data. More recently with the popularity of voice over IP over the Internet and PSTN, supporting voice in data-oriented WLANs has become attractive as well.

As we saw earlier, in a packet communication environment, voice and data have different requirements. Voice packets can tolerate errors and even packet losses (a loss of 1 to 2 percent of voice packets has insignificant effect on the perceived quality of reconstructed voice [KUM74]), whereas data packets are sensitive to loss and errors but can generally tolerate delays. Also, the rate at which information is transmitted is constant in the case of voice, thereby making circuit-switching a viable and efficient approach, whereas information generated for data transmission is very bursty. As a result, voice- and data-oriented networks use different multiple access methods. In a wireless environment, the simplest approach is to assign different frequency bands to isochronous (voice) and asynchronous (data) packets. However, integration in one frequency band will result in a more efficient use of the bandwidth, a simpler radio interface, and an environment that provides a better control for synchronizing voice and video (e.g., lip-synch).

4.4.2 Data Integration in Voice-Oriented Networks

Fixed access methods such as FDMA, TDMA, and CDMA were basically designed for access to the circuit-switched voice-oriented networks. Later on, as we saw in Chapter 1, several data services evolved around these systems. The economical incentive for using this medium for mobile data services is to take advantage, either partially or fully, of the existing infrastructure, the terminals, and the frequency bands designed for the voice-oriented networks. This way the mobile data service provider saves in the major costs of deployment, which includes the cost of real estates and the installation of the antenna, and there no longer is a need to obtain new frequency bands for operation of the data service. If possible, using the same terminal for voice and data will reduce the cost and facilitates marketing of the service.

Example 4.22: Mobile Data over FDMA Analog Cellular

The CDPD packet data system introduced in the early 1990s uses available frequency channels in the existing analog FDMA cellular telephone network (AMPS) to provide an overlaid packet data service supporting data rates similar to voice-band modems (up to 19.2 kbps). In its present form, this system does not exploit the pauses between talk spurts but simply takes advantage of the frequency bands temporarily unused by mobile telephone users in each cell area. CDPD uses the unused AMPS channel to develop a communication link between a mobile data unit and a mobile data base station. Ideally, a CDPD terminal can use the RF and antenna of an AMPS terminal to communicate packet data bursts. However, the most important issue for the CDPD network is that it can use the same antenna site and the antenna towers and the frequency bands of an existing AMPS network. Because real estate, installation of the antenna post, and the frequency bands are perhaps the most expensive parts for implementation of a network, CDPD was perceived to provide a cost-efficient solution for a mobile data service with a comprehensive coverage. The air interface protocol and modulation technique used in the CDPD are, however, different from the AMPS system.

Example 4.23: Mobile Data over TDMA Systems

The GPRS packet data network, introduced in late 1990s uses the air interface and the infrastructure of the GSM network to provide a mobile packet data service that can support data rates of up to a couple of hundred kilobits per second. GPRS uses the same physical packet format and modulation technique as GSM. The logical channels used in GPRS do not use the dialing procedure used in the GSM. In a manner similar to CDPD, through the wired infrastructure of the network, the packets of data in GPRS are routed to the packet switched data networks rather than being switched through the PSTN. GPRS is designed to take advantage of the unused time slots of a TDMA voice-oriented GSM network.

Example 4.24: Mobile Data over CDMA Networks

In TDMA systems and FDMA systems, the data users may use free time slots and free channels, respectively, as they become available. In a CDMA system, the situation is somewhat different. The structure of CDMA is such that all active users

use the entire bandwidth-time space simultaneously. The resource to be managed is signal power. With the application of efficient power control algorithms, the signal levels transmitted by mobile stations and the base station are continually adjusted in response to the changing locations of mobiles and the number of users on the system at any given time. In a CDMA network, the integration of data calls with voice calls is straightforward in principle, because various numbers of both categories of calls are readily mixed together, with each call accessing the channel with its unique user signal code. Therefore, in a CDMA system no modification need be made to the channel access scheme to accommodate integration of voice and data "channels," and the information rate for voice or data traffic in any one channel can in principle be varied by a variable-rate scheme such as that used for voice service in the IS-95 standard. The integration of voice and data services in a single user channel is not necessarily straightforward.

From a technical point of view, there are two incentives for integration of data into voice-oriented fixed-assignment access methods:

1. The fixed-assignment access methods used in voice-oriented networks are designed to support a certain number of simultaneous users. When the number of active users falls below that number, some portion of the transmission resources is wasted.
2. A typical two-way conversation does not make full use of the call connection time, because only one of the parties talks at a time. Furthermore the flow of natural speech is because actually composed of *talk spurts* with intervening short pauses. It is generally estimated that in a two-way voice connection, the average *voice activity factor* for each party is in the vicinity of 40 percent and thus about 60 percent of available transmission time remains unused.

Assume that we have N voice channels available for a given area (e.g., the coverage area of a sectored antenna in a cellular deployment) to accommodate newly originated calls as well as calls handed off from other areas. Further assume that the overall calls in the area are generated according to a Poisson process with a normalized rate of $\rho = \lambda/\mu$ calls/unit channel and length of call holding is generally distributed with a unit mean. The number of idle channels, N_{idle}, according to renewal theory [BUD97] is given by

$$N_{idle} = N - \rho(1 - B(N,\rho)), \tag{4.10}$$

where $B(N,\rho)$ represents the blocking probability calculated from the Erlang B equation given by:

$$B(N,\rho) = \frac{\rho^N / N!}{\sum_{i}^{N} (\rho^i / i!)} \tag{4.11}$$

Figure 4.27 shows the average number of the idle voice channels per area N_{idle} versus the number of available channels N for a variety of call blocking rates. At call blocking rates of around 2 percent that is desirable for most cellular systems, a number of free channels are available for data communications. As the network

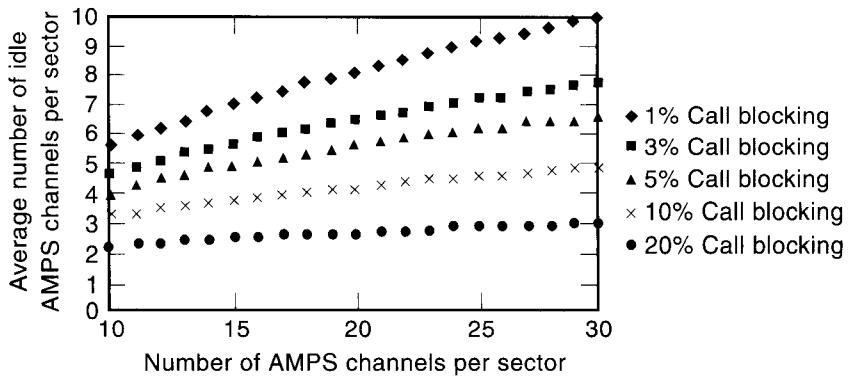

Figure 4.27 Average number of idle channels per area as a function of the number of available channels for a given call blocking rate [BUD 97].

accepts larger values of blocking rates, because most of the time all channels are in use, regardless of the number of channels, only a few channels will be available for data. If the integrated system uses the idle channels for data transmission, on the average, the maximum throughput available to the data user is N_{idle} times the encoding rate of the voice channel. If the system can take advantage of the silence periods in the two-way telephone conversations, as we indicated earlier, an additional throughput of up to 60 percent is available for data applications.

Another important parameter is the idle time for a voice channel. The idle time is the period of time that a voice channel is not occupied by a voice user. In an N channel voice-oriented network, assuming that each channel receives an equal fraction of the call load, one may calculate T, the average length of time a channel is idle, by [BUD97]:

$$T = \frac{N - \rho(1 - B(N,\rho))}{\rho(1 - B(N,\rho))} \tag{4.12}$$

Figure 4.28 shows the normalized average idle period per channel T versus the number of available voice channels for different blocking rates. For a typical holding time of around a couple of minutes and a blocking rate of around 2 percent, the average idle periods are fairly long implying that a data network has a reasonable time to detect the availability and send its bursts of data.

There are periods of time for which all the voice channels are occupied for the voice users, and there is no channel available to the data service. If these periods are short and infrequent, some data applications may accept the situation. Otherwise, specific channels should be allocated for data only usage. In these situations the data users have their own channel, as well as unused portions of the voice channels.

Assuming that we accept the block-out periods and assign no dedicated channel resources to the data application, we will have periodic operation between the available and blocked-out periods. Assuming that the holding time for the telephone conversations is exponentially distributed, it can be shown [BUD97] that the average active period where a channel is available for data, T_a, is given by:

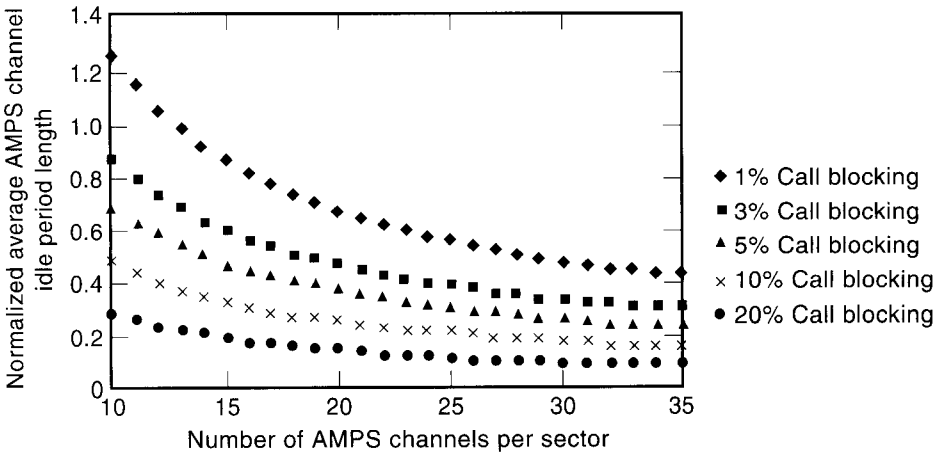

Figure 4.28 Normalized average idle period per channel as a function of the number of channels and call blocking probability [BUD97].

$$T_a = \frac{1 - B(N,\rho)}{NB(N,\rho)} \tag{4.13}$$

The mean length of the blackout period, T_b, for this case is independent of the call load and hence blockage rate and it is given by:

$$T_b = \frac{1}{N} \tag{4.14}$$

Figure 4.29 shows the normalized mean length of the active period T_a versus the number of channels. As the call blocking rate increases, the active period shortens, leaving the system in more blackout periods. For a blocking rate of 5 percent or less and with fewer than 25 channels, the system has the equivalent of one dedicated channel for data applications. At higher blocking rates and data applications that cannot tolerate blackout periods, the data service must be deployed with at least one dedicated channel. Some practical examples are helpful at this stage.

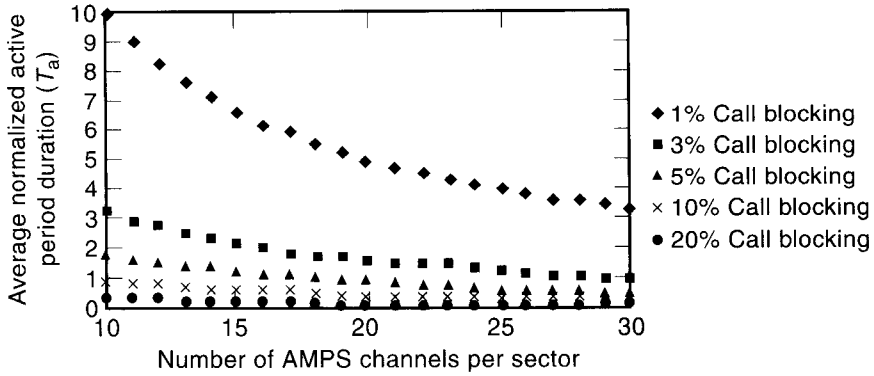

Figure 4.29 Mean length of available time for data as a function of the number of voice channels [BUD97].

Example 4.25: Data Overlay in FDMA Systems—CDPD

The above was actually developed for CDPD, and all the analyses and discussions are directly applicable to it. As we saw in Chapter 1, CDPD operates over analog FDMA cellular networks at the rate of 19.2 kbps per channel, which corresponds to digital transmission using GMSK modulation over 30 kHz AMPS channels. CDPD supports a channel hopping feature that allows a mobile data terminal to move to another channel during a communication session releasing the current channel for voice telephone conversation. This feature helps maintain the blockage probability at its nominal value and allows continual operation during the handoffs. The weakness of CDPD data overlay is that it does not assign several voice channels for one data user to support higher data rates. In general, in an FDMA system, the assignment of multiple voice channels to a single data user involves simultaneous operation of several RF channels by one terminal that is not practically attractive. For the same practical reason, data overlay in FDMA systems encounters difficulties in taking advantage of the silence periods during telephone conversations.

Example 4.26: Data Overlay in TDMA Systems—GPRS

An example of TDMA data overlay is GPRS. All the earlier analysis is applicable to GPRS applications as well. However, the format flexibility of TDMA allows multislot assignment to support higher data rates. GPRS also does not take advantage of silence periods in two-way telephone conversation.

An efficient method of integrating voice and data packets is a *movable boundary TDMA scheme with silence detection.* This method has been applied in the time-assignment speech interpolation (TASI) system used in T1-carrier telephone networks [FIS80] to maximize the number of voice users carried and to integrate data transmission into the channel. Using this basic idea, it is possible to design a *TDMA/framed-polling* protocol to integrate voice and data packets in a WLAN. This system consists of a number of voice and data terminals and a central station, which coordinates all the transmissions [ZHA90]. The protocol for integration of voice and data packets is a movable boundary TDMA scheme shown in Figure 4.30. A frame is divided into two regions with a boundary between them. The first region is used for both voice and data traffic where the voice traffic has priority. If not, voice packets occupy all the slots in this region, and the remaining slots are used for data traffic. The second region is reserved exclusively for data traffic. The boundary between the voice and data regions moves in accordance with the number of active voice packets in each frame. The maximum number of voice packets per frames is N_1, which is assigned an appropriate value to ensure some minimum data traffic capacity and to keep the blockage of voice packets below a selected value (2 percent in [ZHA90]).

The result of extensive analysis and simulation in [ZHA90] provides a simple experimental relation between the capacity that can be allocated for voice and data applications: $D = R_T - 0.032 \times N_v - 0.29$ where D is the data rate in Mbps available for data applications, N_v is the number of active 64 kbps PCM-encoded telephone conversations and R_T is the transmission rate available on the medium. We can

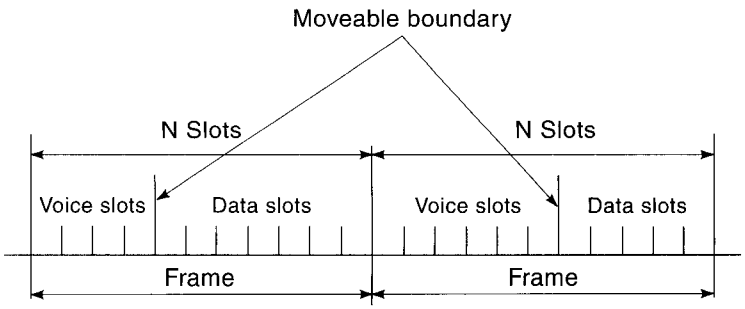

Figure 4.30 Frame structure in a moveable boundary frame-polling system.

apply this equation to the TASI system that uses this protocol to accommodate 30 voice users with some additional data in a traditional 24-voice channels system (T1 carrier). The transmission rate is $R_T = 0.064 \times 24 = 1.536$ Mbps to support 24 voice users at 64 kbps. Using the equation for $N_v = 30$ active users, the data throughput with maximum delay of 10 ms (used in the simulation) is 286 kbps. The new protocol supports 6 more voice users plus 280 kbps data. With the same 24-voice users, 487 kbps would be available for data applications that are around 30 percent of the overall transmission rate.

Example 4.27: Data Overlay on CDMA

As we saw earlier in this chapter, integration of bursty data with voice in the CDMA system is very simple, and CDMA systems already take advantage of the voice activity factor. Therefore, with the same infrastructure and terminals, data services can be overlaid on CDMA. If higher data rates are needed, one can either reduce the processing gain of the data channel or assign several parallel channels for one data link. Indeed the natural flexibility of CDMA to accommodate a variety of data services is one of the major reasons behind selection of the CDMA for 3G systems. Qualcomm has suggested *high data rate* (HDR) where asymmetric uplink and downlink data rates can be supported by simply using multiple carriers on the downlink for higher data rates.

4.4.3 Voice Integration into Data-Oriented Networks

Integration of voice and data has been discussed extensively in the literature. Most of these studies are concerned with protocols with explicit synchronization between the receiver and the transmitter. These approaches use assignment-based protocols for integration of voice and data, which allocate a fixed reference time such as a slot for transmission of packets. Synchronous systems provide more control on delay for voice traffic but less flexibility for bursty data traffic. Another approach that does not need explicit synchronization between the receiver and the transmitter is the asynchronous approach. The asynchronous packet approach mostly uses protocols extended from packet data networks, which are more suited for bursty

data traffic. The voice traffic in this approach requires relatively complicated handling to limit the delay.

Contention-based packet communications protocols such as ALOHA and CSMA are used for data-oriented wireless networks. They are especially well suited to networks comprised of many user stations, each with low-average data rate and potentially high-peak rates. These protocols can operate with little or no centralized control, and can generally accommodate variable numbers of users in the network. However, contention-based schemes can become very inefficient in sharing the communications resources when the traffic load is heavy, as the system throughput degrades and the transmission delays increase. The unpredictability of throughput and time delays make these access methods unattractive for a voice-dominated communication service, where a minimum throughput and delay is essential for user acceptance of the QoS. Up to recent times (the Internet and wireless age), wired telephone services and the PSTN were producing the dominant source of income for the telecommunications industry. In the past century, telephone users have accepted the quality of the PSTN wired voice services as a normal standard for telephone conversations.

4.4.3.1 QoS in Voice Services

In the language of digital packet communications, the QoS of the PSTN voice user is specified by a guaranteed 64 kbps PCM (or 32Kbps ADPCM) coded data rate and a maximum delay of around 100 ms. We refer to this QoS as *wire-line-voice quality*. With the introduction of cellular telephone services in the late twentieth century, users accepted a lower QoS that suffered from the effects of fading due to the radio channel and dropped calls due to handoff, lack of coverage, or other reasons. As we saw in Chapter 1, cordless telephony and PCS services were aimed at bringing their QoS close to that of wire-line quality and 3G cellular, that is, a merger of cellular telephony and PCS service follows the same pattern. If the quality of voice in a cordless telephone was far below that of wireline quality, users would have rejected the service. This is because they have the choice to receive a better QoS (no drops or fading effects) with their wired telephone that is also available at home or in the office. However, users had to accept lower QoS with cellular telephony because there was no other alternative service provision for vehicles and other mobile applications.

Another recent event deviating from the wireline QoS is the emergence of voice over Internet or, as most people refer to it, the voice over IP, phenomenon. The popularity of the Internet, its penetration into the home market, its capability to support multimedia, and the most important advantage, its uniform cost for local- and long-haul communications, encouraged development of Internet telephony. Operating on a packet switched environment with contention-access, the QoS of these services is not guaranteed at all, and in its present stage of technology, the quality of voice calls is well below that of wire-line quality. However, free international calls through an Internet connection have been an incentive for some users to try this option as well.

In wireless networks, voice-over IP (VoIP) does not make sense for mobile data applications because these services provide low data rates, and after all they

are evolving as an auxiliary network over already existing voice-oriented networks. However, VoIP can be considered for WLAN environments. Imagine a WLAN installation in a stock market hall supporting wireless terminals for the users working in the hall. It would be useful and beneficial if they had a VoIP service at the same terminal to use it for their telephone conversations as well. This incentive has initiated preliminary work on VoIP in a WLAN environment using contention-based access methods. The work in [ZAH00, FEI00] determines the number of supported voice terminals in a WLAN environment under a variety of conditions.

4.4.3.2 Capacity of a WLAN with Voice and Data

WLANs are becoming very popular in indoor applications such as in stock exchange halls where mobile users demand high-speed wireless data access to the network and voice capabilities for telephone conversations. To deploy such a network, a mathematical framework is very helpful to compare the capacity performance of WLANs with voice and data services in different scenarios. Therefore, for an asynchronous WLAN using the TCP/IP protocol suite, we need to find an answer to two questions:

1. What is the number of network telephone calls that can be carried with a given amount of data traffic?
2. What is the maximum data traffic per user for a given number of voice users?

A mathematical framework to answer these two questions is provided in [ZAH00] where the integration of voice and data with TCP/IP protocol that operates in an asynchronous CSMA access environment is analyzed.

To integrate voice packets in a TCP/IP environment, the first step is to select a speech coder. A variety of speech coding algorithms exist with different rates as discussed in Chapter 3. To reduce the load on the network generated by voice traffic, [ZAH00] adopted IMBE that is a popular low-data-rate vocoder (4.8 kbps) with an acceptable QoS. This vocoder has been used in INMARSAT-M and AUSSAT mobile satellite communication systems and is proposed for APCO-25 standard for narrow-band digital land mobile radio.

Using IP will provide two options for sending packets in the network: the transmission control protocol—TCP and the user datagram protocol—UDP. As a streaming protocol, UDP has no support for error correction, acknowledgment, sequencing, and flow control. Under high traffic conditions, the lack of flow control in UDP may cause bandwidth saturation in the Internet that should be prevented by the application program. In contrast, TCP has an error-correcting mechanism, uses acknowledgment, and guarantees in-order packet delivery. These requirements demand additional overhead that increases the average delay and reduces the overall throughput of the network. In general, data packets can tolerate delay but cannot tolerate packet loss, whereas the voice packets can accept packet loss of the order of 1 percent but cannot afford delays of more than 200 ms between consecutive packets. In the system described in [ZAH00], TCP is used for the data packets to guarantee accuracy of information and UDP for voice packets to handle the delay requirement. This approach is adopted in several products available in

the market, whereas other products use TCP for both voice and data. Although in this analysis TCP is selected for data transmission, there are some products, that use UDP for data transmission. In this case the upper layers are responsible for reliable delivery of data.

Figure 4.31 represents a general overview of the system where several voice and data terminals communicate with an AP of a WLAN. Because the human ear is sensitive to time delays larger than 200 ms (T_{th}) in a voice conversation, wireless terminals should provide some facilities to minimize voice time delay. Allocating higher priority for transmitting voice over the data traffic is one way to decrease the voice packet time delay. Therefore, voice and data packets should be stored in a queue and wait for transmission as shown in Figure 4.32. The total delay for each packet transmission consists of a *queuing delay* at the terminal and *channel transfer delay*. The work in [ZAH00] uses an *M/G/1* queue[3] with two priorities and a single server for node modeling. The arrival is modeled by a Poisson random variable. Figure 4.33 shows the system capacity and delay for T_{th} = 100 ms and 50 ms with a 1 and a 2 Mbps channel bandwidth. The maximum number of the voice users declines with reducing T_{th}.

Problem 13: Capacity of a WLAN with Voice and Data Users

Using Figure 4.33, find the number of users for T_{th} = 100 ms and T_{th} = 50 ms when the channel bandwidth is 1 Mbps.

Solution:

From the figure a maximum of 18 voice users are supported with T_{th} = 100 ms, and decreasing T_{th} to 50 ms will reduce the maximum voice users to 14. In this example the data traffic is less than 10 kbps.

4.4.3.3 IP Telephony Using WLANs

The previous section provided an analysis of how many voice channels can be supported using UDP protocol over a CSMA channel in a wireless LAN environment where data transmission is coexisting. In practice, there are a number of VoIP software packages and services, such as Speakfreely, Net2Phone, DialPad, and so on, that can be used to implement a real-time testbed to analyze the behavior of voice in a WLAN environment. The purpose of setting such experimental testbeds or simulations is to determine the number of voice users that can be supported and relate it to the design parameters. In this section, we provide a practical framework for the analysis of VoIP in a WLAN environment. The principles of operation of voice over contention-based packet-switched networks and assigned access circuit-switched networks are very different. In a circuit-switched network, a fixed connection between the two terminals is established during the call establishment procedure. This connection supports end-to-end communications with a fixed data rate and a controlled delay dominated by propagation delay and a negligible delay jitter. In packet voice communications with contention access and packet-switched

[3]An M/G/1 queue implies that packet arrivals are Poisson and the service rate of the queue has a general distribution with known mean and variance (see [BER87]).

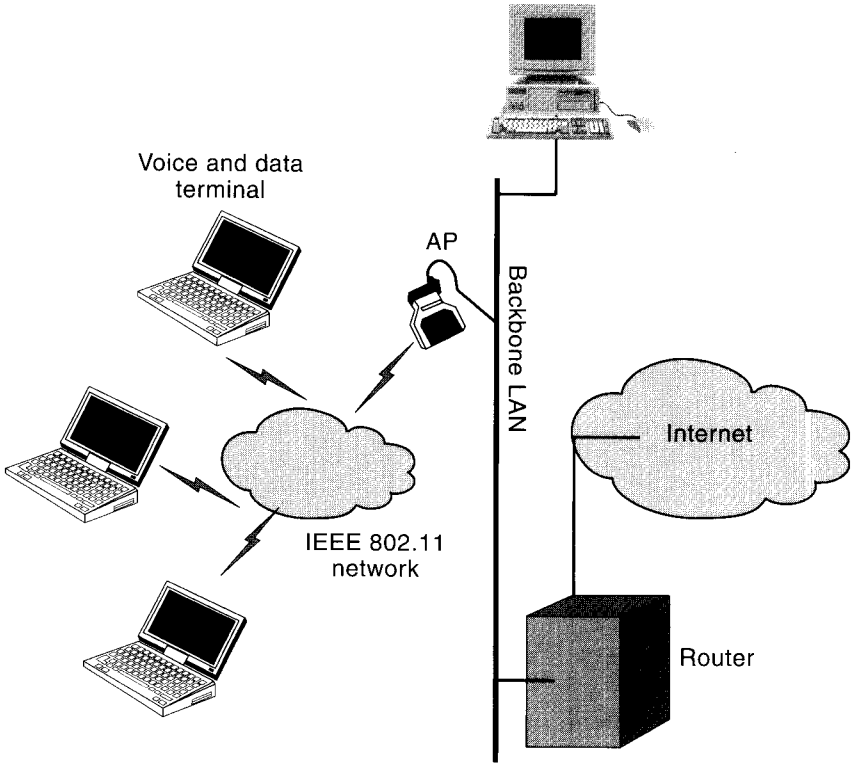

Figure 4.31 Schematic of a system employing voice and data terminals in a WLAN.

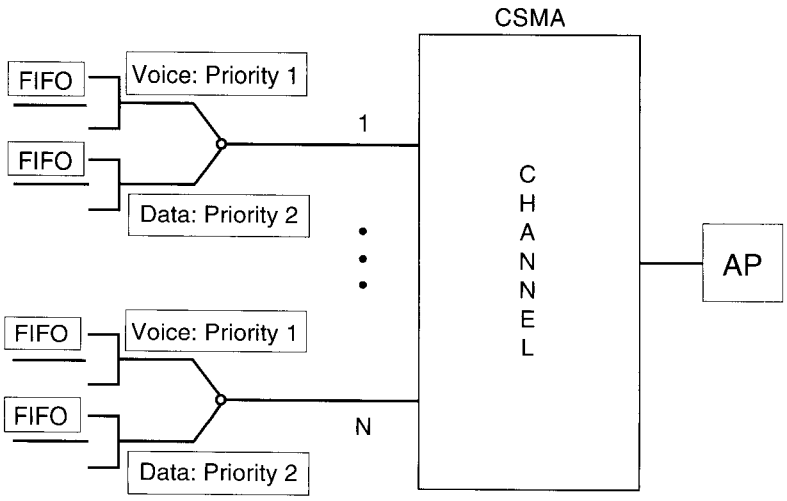

Figure 4.32 Queuing model for prioritizing voice traffic.

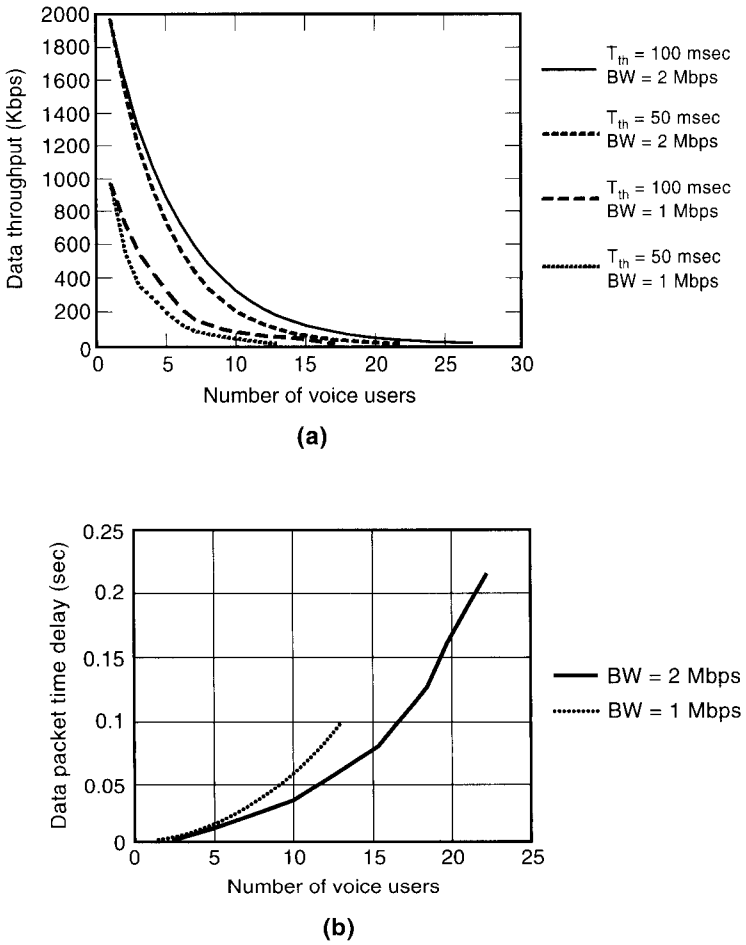

Figure 4.33 (a) Throughput of data versus number of voice users for variety of thresholds for acceptable delay in voice packets and (b) Data packet time delay versus number of voice users.

networks, delay and jitter are the dominant sources of the performance degradation. To regulate jitter, the receiver has a buffer to store the received packets at different delays but pump them to the user at constant intervals to reconstruct real-time voice. The performance of the system is then related to the size of the buffer at the receiver. This section provides a summary of the experimental work presented in [FEI00] to relate the throughput to the buffer size.

The first step is to describe the overall scenario in a practical situation for the implementation of the VoIP in a WLAN environment. Figure 4.34 describes the arrival of packets. Because of random access and packet-switched networking, the packets sent at fixed intervals arrive at a variety of delays.

The overall delay is the minimum network delay plus the individual jitter per packet that will be regulated with the jitter compensation buffer at the receiver.

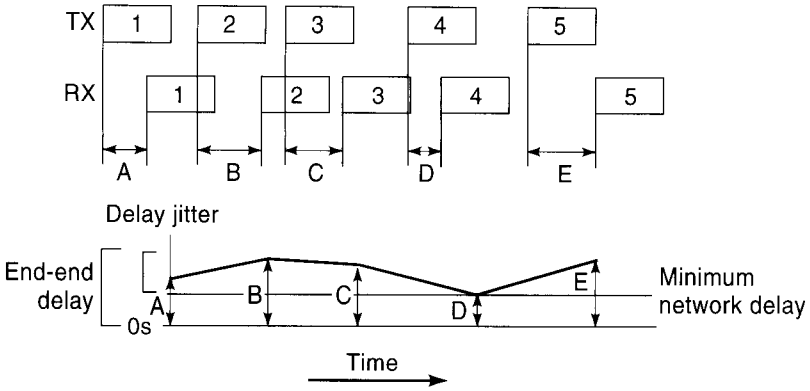

Figure 4.34 Illustration of the arrival of voice packets transmitted at a constant rate.

The packets arriving at different delays are stored in a buffer, and the application at the receiver reads this buffer periodically. When the packet with the right sequence number is available, the receiver reads the packet and plays it through the speaker. When the packet is not available, the application software at the receiver skips that packet. A simple example further clarifies the operation.

Example 4.28: Jitter in VoIP on WLANs

Figure 4.35 illustrates the details of the operation of the receiver and the relationship between the jitter compensation buffer, arrival, and playing time of the packets. When the first packet arrives, it is delayed in the receiver's jitter compensation buffer and after the maximum allowed delay, it is delivered to the user application to be played in the first slot. The second packet is discarded because it has arrived after the deadline for playing. The third packet arrives normally before the deadline, and it is delivered at its appropriate time to the speaker. The fourth and the fifth packets arrive off-sequence. The fourth one is late (arriving after its deadline), and it is discarded. The fifth packet has arrived before the deadline of the fourth packet, and it is shifted to its own time slot.

As we have discussed before, the user can accept a packet drop rate of around 1 percent. To reach the goal of 1 percent packet drop, the receiver has the choice of increasing the length of the jitter compensation buffer at the expense of additional overall delay at the receiver. Therefore, the length of the jitter compensation buffer is an important parameter in VoIP applications. This parameter is adjusted by changing the length of the buffer at the receiver. The delay observed by the user is the minimum network delay plus the jitter compensation buffering delay. Example 4.28 also illustrates that in VoIP applications. In addition to transmission packet losses occurring in the network, we have packet losses due to late arrival at the receiver (that is a function of the length of the jitter compensation buffer). Therefore, the length of the jitter compensation buffer and packet loss are interrelated.

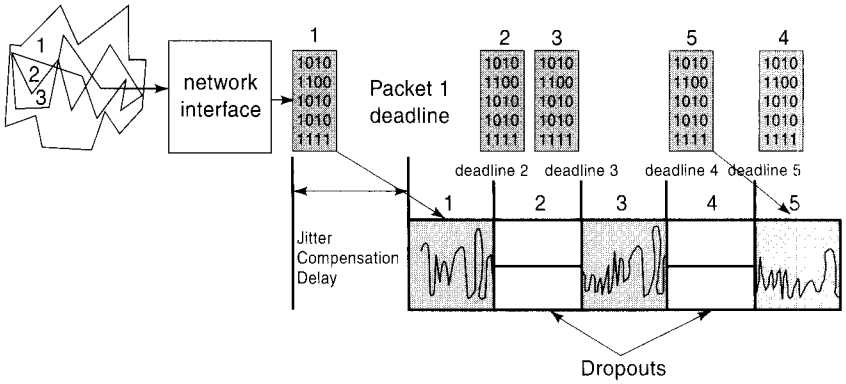

Figure 4.35 Reception of voice packets and buffering to maintain appearance of no jitter.

To determine the relationship between the jitter compensation delay and the packet loss rate, a testbed was developed in [FEI00] to implement the scenario shown in Figure 4.31. In this test bed, an infrastructure for wireless LAN operation using an AP and a number of laptops is used for measurement of the statistics of the delay jitter in a VoIP application. Figure 4.36(a) shows the statistics of the delay jitter for 1,800 packets. The measurement system transmits, time stamps, and stores the packet at the transmitting and the receiving laptops. The stored files are then post processed to eliminate the effects of differences between the clocks of the transmitter and the receiver laptops, and we extract the refined delay jitter measurements. Figure 4.36(b) shows the accuracy of the system (that is measured by comparing the results obtained from two separate laptops connected to the same point and receiving the same message from a receiver). The measurement error (difference of measurements in two identical laptops) has a mean of around .01 ms, whereas the mean of the measurements is around 1 ms, restricting the measurement error to around 1 percent. Using a delay jitter distribution, one can simply find the relationship between the packet loss and the jitter compensation buffer length. For any given jitter compensation buffer length, the probability of packet loss is the same as the probability of having a delay jitter larger than the jitter compensation buffer length. Figure 4.37 shows the experimental results for up to five stations operating in the wireless LAN test bed. If we fix the acceptable packet loss rate to 1 percent, the minimum buffer length increases from 0.5 ms to 7 ms when we increase the number of users from one to seven. The details of algorithms for the implementation of the test bed and the results of OPNET simulation for large number of voice users are available in [FEI99].

QUESTIONS

4.1 Name two duplexing methods and one example standard that uses each of these technologies.

4.2 What are the popular access schemes for data networks? Classify them.

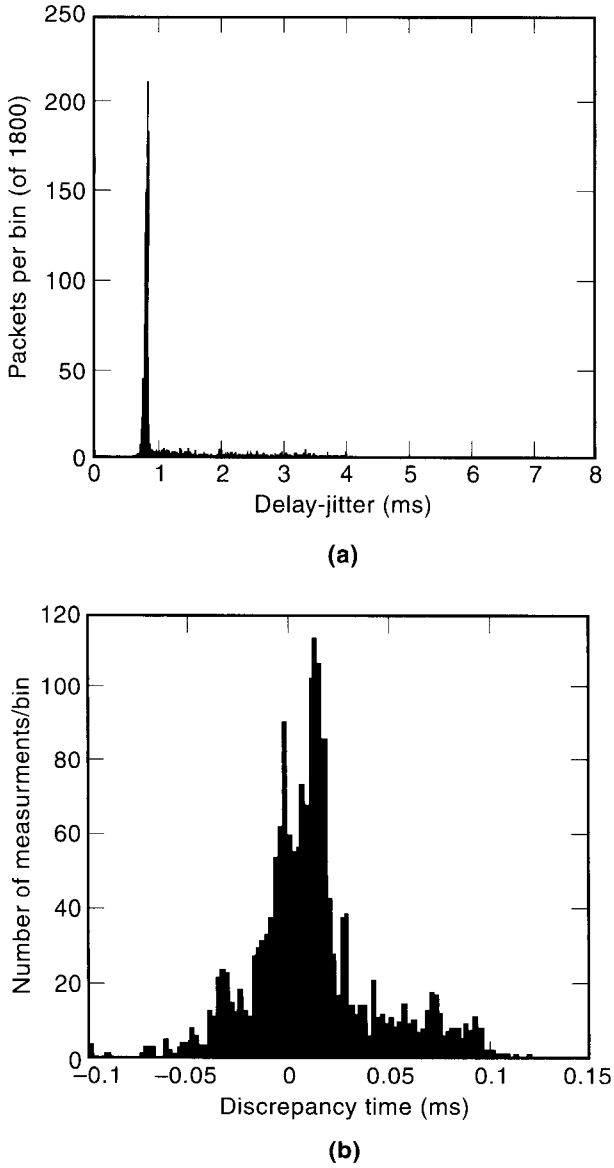

Figure 4.36 (a) Measured delay jitter in the wireless LAN testbed and (b) accuracy of the measurements.

4.3 Name a cellular telephony standard that employs FDMA.

4.4 What is binary exponential back-off algorithm, which standard uses it, and what is the purpose of using it? What is its weakness?

4.5 What is the purpose of the IEEE 802 standard committee? What are the steps taken to make its recommendations an international standard?

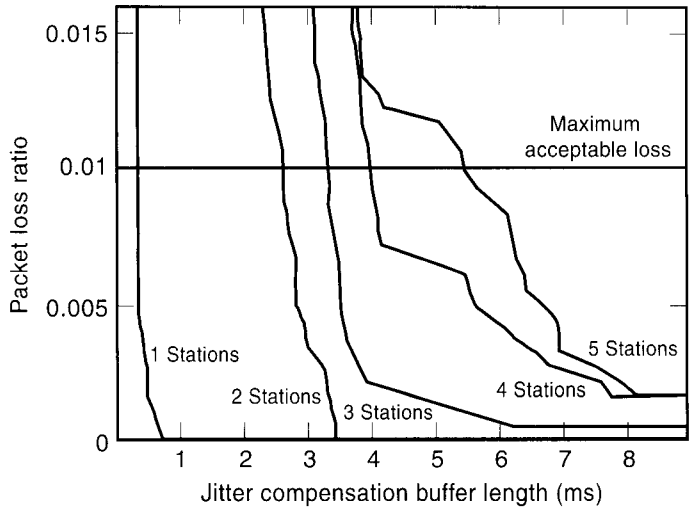

Figure 4.37 Packet loss versus jitter compensation buffer length.

4.6 What is the difference between the access techniques of IEEE 802.3 and IEEE 802.11?

4.7 Why do most PCS standards use TDD, while most cellular standards use FDD?

4.8 Why in the PSTN backbone hierarchy has FDM multiplexing lost its popularity to TDM multiplexing?

4.9 Why did the 2G cellular systems shift from analog FDMA to digital TDMA and CDMA?

4.10 Name three standards using TDMA/TDD as their access method.

4.11 What are the advantages of the CDMA access techniques?

4.12 What is the difference between performance evaluation of voice-oriented fixed assignment and data-oriented random access methods.

4.13 Explain the difference between the effects of power-control on the capacity of TDMA and CDMA systems.

4.14 In a radio ALOHA network, how does a terminal learn that its packet is collided?

4.15 What is the difference between the maximum throughputs of ALOHA and Slotted ALOHA networks? What causes this difference?

4.16 What is the difficulty of implementing CSMA/CD in a wireless environment?

4.17 Explain the difference between carrier-sensing mechanisms in the wired and wireless channels.

4.18 What is the hidden terminal problem and how does it impact the performance of a CSMA-based access method?

4.19 What is the capture effect and how does it impact the performance of the random access methods?

4.20 Explain the differences between integration of data into a voice-oriented network and integration of voice into a data-oriented network.

4.21 Explain the relation between the receiver buffer size and packet error rate in voice-over IP applications.

PROBLEMS

4.1 To provide public telephone access to commercial ferries a telephone company installs a multi-channel wireless telephone system in a ferry. This wireless radio system connects to a base station on the shore through the air. The base station is connected to the PSTN using wires.

a. If the telephone company installs a four-channel system, what is the probability of having a person come to the telephone and finding none of the lines are available? Assume that the average length of a telephone call is 3 minutes and each of the 150 passengers of the ferry make on the average one call per hour.

b. What is the average delay for accessing the telephones?

c. How many channels are needed to keep the blockage probability below 2 percent?

4.2 **a.** Neglecting the frequency spectrum used for control channels, what is the maximum number of two-way voice channels that can fit inside the frequencies allocated to the AMPS system.

b. What is the number of channels in each cell? Note that $K = 7$ was originally used in the AMPS.

c. Repeat (b) for IS-136 in which $K = 4$ and the number of slots per TDMA channel is three.

d. Repeat (b) for IS-95 CDMA, assuming the minimum required E_b/N_0 is 6 dB. Include the effects of antenna sectorization, voice activity, and extra CDMA interference.

e. Repeat (d) for broadband CDMA where 5 MHz bands are used in each direction.

4.3 **a.** Sketch the throughput versus offered traffic G for a mobile data network using slotted nonpersistent CSMA protocol. The packets are 20 ms long and the radius of coverage of each BS is 10 km. Assume the radio propagation speed is 300,000 km/second and use the worst delay for calculation of the a parameter.

b. Repeat (a) for slotted ALOHA protocol.

c. Repeat (a) for 1-persistent CSMA protocol.

d. Repeat (a) for a WLAN with access point coverage of 100 m.

e. Repeat (a) for a satellite link with a distance of 20,000 km from the earth.

4.4 Use the equations from Section 4.2.2 to reproduce Figures 4.27 and 4.29.

4.5 A cellular carrier has established 100 cell sites using AMPS with 395 channels and $K = 7$.

a. Use Figure 4.8 to calculate the total number of subscribers for a blocking probability of 0.02, average of 2 calls per hour, and average telephone conversation of 5 min.

b. Use computer software, such as MathLAB, or MathCAD to calculate the same values using the Erlong B equation directly.

c. Determine (either from Fig. 4.9 or calculation) the average delay for a call.

d. Repeat (a) for a blocking probability of 0.01.

e. Repeat (a) if IS-136 was used with $K = 7$.

f. Repeat (a) if IS-136 was used with $K = 4$.

4.6 We want to use a GSM system with sectored antennas ($K = 4$) to replace the existing AMPS system ($K = 7$) with the same cell sites.

a. Determine the number of voice channels per cell for the AMPS system.

b. Determine the number of voice channels per cell for the GSM system.

c. Repeat (b) if we were using a W-CDMA system with the bandwidth of 12.5 MHz for each direction. Assume a signal-to-noise ratio requirement of 4 (6 dB) and include the effects of antenna sectorization (2.75), voice activity (2), and extra CDMA interference (1.67).

4.7 We provide a wireless public phone with six lines to a ferry crossing between Helsinki and Stockholm carrying 100 passengers where, on the average, each passenger makes a 3-min telephone call every 2 hours.

a. What is the probability of a passenger approaching the telephones and none of the four lines are available?

b. What is the average delay for a passenger to get access to the telephone?

c. What is the probability of a passenger waiting more than 3 minutes for access to the telephone?

d. What would be the average delay if the ferry had 200 passengers?

4.8 A WLAN hop accommodates 50 terminals running the same application. The transmission rate is 2 Mbps and the terminals are using slotted ALOHA protocol. The commutative traffic produced by the terminals is assumed to form a Poisson process.

a. Give the throughput versus offered traffic equation for the system and determine the maximum throughput in Erlong.

b. What is the maximum throughput in bits per second?

c. What is the maximum throughput in bits per second for each terminal?

4.9 A local 3-hour tour boat with 50 passengers has one AMPS radio phone to connect to the shore. On the average, each user places one call per tour and the average holding time for the calls is 3 minutes.

a. What is the probability that a person attempts to use the phone and he/she finds it occupied?

b. Repeat (a) if the AMPS phone is replaced by three IS-136 phones using the three slots of the existing IS-136 TDMA over the same band.

c. Repeat (a) if this phone is replaced by six upgraded IS-136 phones using 6-slot upgraded IS-136 TDMA over the same band.

4.10 In a datagram packet switched network with

P: packet size in bits

N: number of hops between two given systems

B: data rate in bps on all links

H: overhead (header in bits per packet)

T: end-to-end delay

N_p: number of packets

L: message length in bits

D: propagation delay per hop

a. Give N_p in terms of L, P, and H.

b. Give T in terms of L, P, H, N, B, and D.

c. What value of P, as a function of N, B, and H, results in minimum end-to-end delay T? Assume that the message length is much larger than the packet size and propagation delay is negligible (D = 0).

4.11 An ad hoc 2 Mbps WLAN using ALOHA protocol connects two stations with a distance of 100 m from one another—each, on the average, generating 10 packets per second. If one of the terminals transmits a 100-bit packet, what is the probability of successful transmission of this packet? Assume that the propagation velocity is 300,000 km/sec and the packets are produced according to the Poisson distribution.