Wireless Networks

Local and Ad Hoc Networks

IVAN MARSIC

Department of Electrical and Computer Engineering and the CAIP Center

Rutgers University

THE STATE UNIVERSITY OF NEW JERSEY
RUTGERS

Contents

Chapter 1
Introduction

Contents

In this text I review basic results about wireless communication networks. Communication may be defined as the process of successfully transferring information between two or more parties. My emphasis is on computation rather than physics of the processes, while keeping in mind that understanding the physics is necessary to design the computation.

The recurring theme in this chapter is the ***capacity and its statistical properties*** for a communication channel. Capacity is modeled differently at different abstraction levels, but the key issue remains the same: how to increase the amount of information that can be transmitted over a channel in the presence of channel impairments. A complement of capacity is *delay*, and although it crops up sporadically, the treatment of delays is deferred to the next chapter.

The main difference between wired and wireless networks is that there are no wires (the air link) and mobility is thus conferred by the lack of a wired tether. This leads to both the tremendous benefits of wireless networks and the perceived drawbacks to them. Some of the key technical challenges in wireless communications come from (*i*) the hostile wireless propagation medium and (*ii*) user mobility. Most of the issues covered in this chapter arise in any data network, wired and wireless likewise. What makes the wireless different is the degree of importance inherent to the problems that often appear in both types of networks. Some examples are as follows.

- Errors due to channel noise occur both in wired and wireless networks, but in the former they play only a minor role. Conversely, in wireless networks, channel errors are a major concern. Moreover, due to dynamic changes in the transmission medium and user mobility, the error probability changes dynamically over a wide range.
- Both wired and wireless networks employ broadcast where multiple stations use the same medium for communication. However, in the wireless case “listening while speaking” is

complicated, so coordinating the medium access or MAC (Medium Access Control) is much more complicated.

- Both types of networks employ routing to bring a packet to the destination. However, wireless "link" is a relative (or "soft") concept. Dynamic changes in the transmission medium and user mobility often cause severance or emergence of a "link" and the associated changes in the network topology, i.e., network graph connectivity.
- The point of belonging to a network is in having relationship with other entities in the network. While a mobile device is moving, the related network entities may be moving in a different direction or not moving at all. This change of the network "landscape"—an uncommon phenomenon in the wired world—presents important challenges from the network management perspective if the network is to provide an undisrupted service to the roaming user.
- Energy expenditure is important in both types of networks (recall screen savers and coolers for wired computers), but battery energy plays a key role in wireless communication. Transmission power also determines the spatial volume within which a given radio transmission interferes with other transmissions. Radio resource management refers to the control signaling and associated protocols that negotiate the optimal usage of network radio resources (transmission signal strength and available radio channels) for communication.
- Application drivers are different (e.g., multimedia, sensors, military, etc.)

As a result, it is very hard to disentangle the concerns and responsibilities of different network layers, e.g., MAC, routing, and resource allocation, in wireless networks.

Protocol Layering

While information is represented and transmitted in the form of signal waveforms, communication between the sender and the receiver is usually viewed as taking place between different layers of abstraction, which are codified as ISO OSI reference architecture, see Figure 1-1. The key design principle is such that a protocol at layer i does not know about the protocols at layers $i-1$ and $i+1$. The aspects of abstraction include the type of *service* offered and the type of *representation* of messages exchanged between the communicating entities. The services offered by a higher abstraction layer are composed of the services offered by the lower abstraction layers. The type of information grows in abstraction from physical signals and streams of symbols to software objects.

In this chapter I consider the bottom three layers: *physical*, *link*, and *networking*. As will become apparent below, the separation of functions between layers is not as clean as it is with point-to-point links and wired networks in general.

The key service offered by the physical layer is a symbol stream with a transmission error rate. This is achieved by the physical transmission and reception of signals over the wireless channel, reviewed in Chapter 2. While the physical layer represents messages as discrete entities, usually called *frames*, the information exchanged between the communicating parties is essentially viewed as a continuous signal representing a stream of symbols (bits).

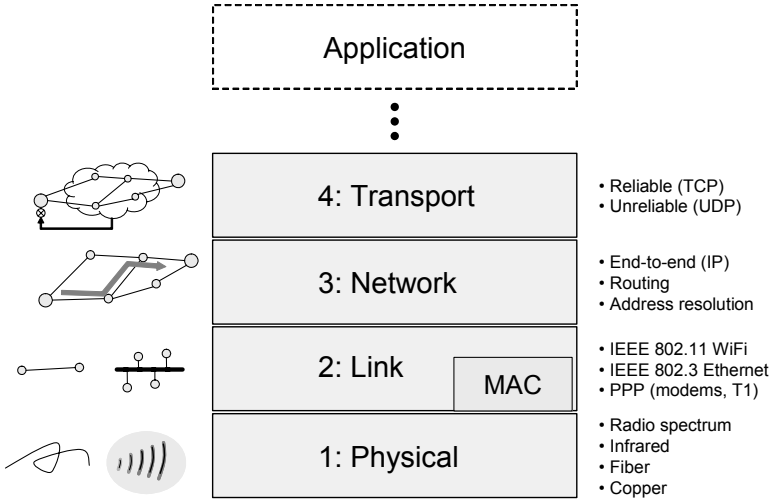


Figure 1-1: ISO OSI protocol reference architecture. Layers 5 (Session) and 6 (Presentation) are not shown.

The link layer deals with the messages at the output of the receiver. Functions exercised at the link layer can include radio resource management (e.g., power control, transmission rate allocation and error control), and network resource management (e.g., service scheduling and call admission control). Media access protocols controlling the access to shared medium are described in Chapter 3 and a detailed example of IEEE 802.11 standard is provided in Chapter 4. The networking layer comprises protocol stack that includes handoff management, location management, traffic management and control. Then ad hoc wireless networks are briefly reviewed in Chapter 5. Finally, some emerging future trends are described in Chapter 6.

Architectures for Wireless Networks

Mobility and lack of tethered links between the communicating entities present unique technical challenges. Wireless networks usually exploit the fact that the radio waves die out as they travel, so the same slice of spectrum (frequency channel) can be reused in different locations with minimal interference. A key choice is whether or not to rely on any fixed infrastructure, i.e., whether all system components are physically movable or stationary.

Depending on whether or not there is a fixed infrastructure, wireless systems can be categorized as either infrastructure systems or ad hoc networks (Figure 1-2). *Infrastructure systems* have fixed network resources in the form of a *base station* or *access point* (AP), which performs central administration for multiple mobile stations. The base station routes packets between mobile nodes, which do not communicate directly with each other. Infrastructure systems are organized in *cells*, which reuse portions of the spectrum to increase spectrum usage at the expense of greater system infrastructure. Base stations may have high-bandwidth connections to wired infrastructure networks. There is no direct ("peer-to-peer") communication between users. Rather, each mobile station sends its message directly to the base station in its cell[1], and this local base station routes

[1] In more sophisticated cellular systems, base stations in adjacent cells also receive and process the message, particularly when the user is near a cell boundary.

the messages to the base station in the cell of the intended recipient. In turn, this remote base station broadcasts the message within its cell for the receiving mobile station to pick up. Cellular networks provide the information transport platform for wireless local area networks (WLANs) and wireless wide area networks (WWANs).

Ad hoc networks have no pre-existing (fixed) infrastructure and the network architecture is configurable. They are formed by wireless stations which may be mobile and they route paths for each other. Every station in an ad hoc network can be set up as, and play the role of, a base station where it can directly transmit and receive from other stations in the network. Packets may need to traverse multiple links to reach a destination. Due to the mobility, the routes between stations may change dynamically. Ad hoc network can co-exist and co-operate, i.e., exchange data packets, with an infrastructure-based network.

Clearly, real world is not as rigid as the above codification would imply. Another useful concept is that of wireless mesh networks. *Wireless mesh network* is a multihop wireless network with mostly static routing nodes. It is a *stable* ad hoc network—a hybrid between the extreme cases shown in Figure 1-2, where one AP connects multiple (mostly stationary) nodes to the Internet. Mobility may or may not be a problem, and since static nodes can be powered by the electric grid, battery power is not necessarily a problem. Examples of wireless mesh networks include: an intra-home or inter-home wireless network; an enterprise-wide backbone wireless network of APs; and, cellular mesh.

The characteristics of a mesh are:

- "Grows" organically
- Does not necessarily require any "infrastructure"
- Increases overall network capacity
- Robust and resilient to faults
- Self-managing and self-administering
- Identity and security is a challenge

Naming and Addressing

Names and addresses play an important role in all computer systems as well as any other symbolic systems. They are labels assigned to entities such as physical objects or abstract concepts, so those entities can be referred to in a symbolic language. Since computation is specified in and communication uses symbolic language, the importance of names should be clear. The main issues about naming include:

- Names must be *unique* so that different entities are not confused with each other
- Names must be *resolved* with the entities they refer to, to determine the object of computation or communication

The difference between names and addresses, as commonly understood in computing and communications, is as follows. *Names* are usually human-understandable, therefore variable

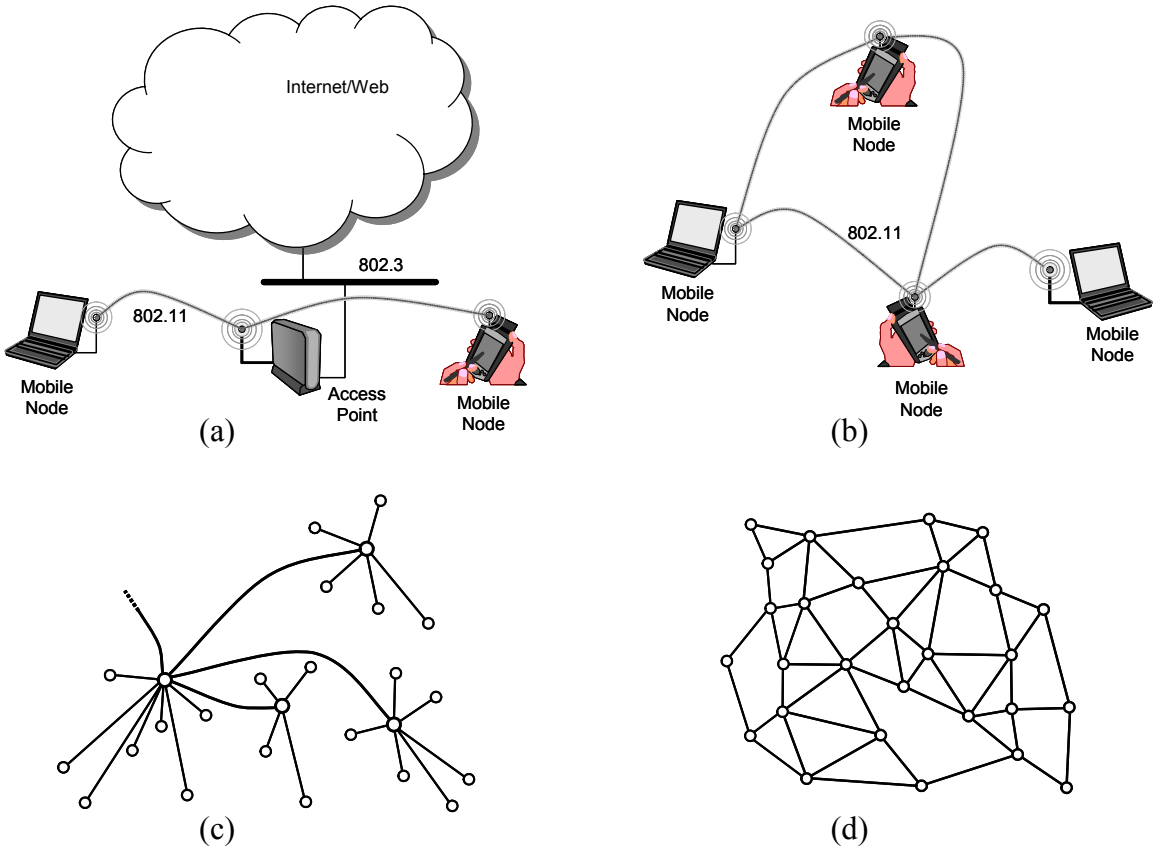


Figure 1-2: Examples illustrating different architectures for wireless networks: (a) infrastructure "cellular," and (b) infrastructure-less ad hoc systems, both based on IEEE 802.11 wireless LAN, also called Wi-Fi. The respective network topologies with nodes and links are hubs-and-spokes (c), and mesh (d).

length (potentially rather long) and may not follow a strict format. *Addresses*, for efficiency reasons, have fixed lengths and follow strict formatting rules. For example, you could name your computers: "My office computer for development-related work" and "My office computer for business correspondence." The addresses of those computers could be: 128.6.236.10 and 128.6.237.188, respectively.

Separating names and addresses is useful for another reason: this separation allows to keep the same name for a computer that needs to be labeled differently when it moves to a different physical place. For example, a telephone may retain its name when moved to a region with a different area code. Of course, the name/address separation implies that there should be a mechanism for name-to-address translation, and vice versa.

Two most important address types in contemporary networking are:

- Link address of a device, also known as *MAC address*, which is a physical address for a given *network interface card* (NIC), also known as *network adaptor*. These addresses are standardized by the IEEE group in charge of a particular physical-layer communication standard

- Network address of a device, also known as *IP address*, which is a logical address. These addresses are standardized by the Internet Engineering Task Force (http://www.ietf.org/)

Notice that a quite independent addressing scheme is used for telephone networks and it is governed by the International Telecommunications Union (http://www.itu.int/).

1.1 Summary and Bibliographical Notes

The recurring theme of this volume is the *capacity and its statistical properties* for a communication channel. Observe the probabilistic nature of capacity bounds. In other words, the bounds specify what is valid for many experiments, but cannot say much about an individual experiment. The capacity of a chain of wireless links is equal to the capacity of the "bottleneck link," i.e., the link with the lowest capacity. Notice that the raw channel capacity, considered in Chapter 2, is reduced due to contention inefficiency, Chapter 3, and routing overhead, Chapter 5. In addition, packet protocols add overhead (packet header and trailer). This considers only the communication loss and assumes that computing nodes (CPU cycles, memory space, etc.) are not bottleneck. Switches in computing nodes may be bottleneck in high-speed networks, see e.g., [Peterson & Davie 2003].

If the data at the receiver have errors that are too frequent for the desired use, the errors can often be reduced by the use of two main techniques:

- Forward error correction (FEC), using the digital signal processing techniques briefly reviewed in Chapter 2
- Automatic repeat request (ARQ), where the receiver directly requests retransmission of erroneous packets or sender indirectly realizes that a packet must be retransmitted since an acknowledgement has not arrived within a prescribed timeout period

The choice between using the ARQ or the FEC technique depends on the particular application. ARQ is relatively inexpensive to implement, but FEC techniques are preferred on systems with large transmission delays where ARQ would make delays much worse.

It could be observed that there exists a gap between *signal processing*, i.e., physical layer, and *network layer* communities. Network-level theoretical models of routing, broadcasting, connectivity, backbones, etc., are commonly based on unrealistic physical layer models.

Emerging research themes in wireless networking include:

- Hybrid ad hoc wireless networks, heterogeneous sensor and ad hoc networks
- More application scenarios
- Cross layering design

- More localized algorithms
- More testing with real equipment
- New technologies, e.g., ultra-wideband, software radio
- Sensor network issues, such as reliability
- Building working networks out of current fundamentals

This text covers many different topics aspects of computer networks and wireless communications. At many places the knowledge networking topics is assumed and no explanations are given. The reader should check a networking book; perhaps the two most regarded networking books currently are [Kurose & Ross 2005] and [Peterson & Davie 2003].

Protocol layering and OSI reference architecture is probably best covered by [Tanenbaum 2003]. This chapter considers only the bottom three layers: physical, link including MAC, and network layers. Most of the problems are common for both wired and wireless networks, but there is a shift in emphasis. A key issue with wireless is the interference—it causes a decrease in channel capacity due to relatively high- and dynamically varying noise. Another issue is link instability, which results in changes of network topology.

An interesting generalization of layered communication is available in [Benkler 2000].

The material presented in this chapter requires basic understanding of probability and random processes. [Yates & Goodman 2004; Bertsekas & Tsitsiklis 2002] provide excellent introduction and [Papoulis & Pillai 2001] is a more advanced and comprehensive text.

Chapter 2
The Radio Channel

Contents

2.1 Introduction

There are three important aspects to address about a communication channel:

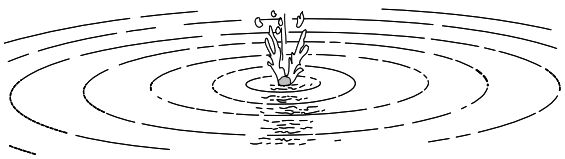
(1) How it works—the way information is "imprinted" on a medium of transmission

(2) How to build analytical or empirical model of channel errors

(3) How to minimize the effect of errors on messages (and thus, on the channel capacity)

I start with briefly reviewing regulatory constraints on implementing radio channels. Next, in Section 2.2 I start considering channel implementation and build an analytical model of error probability under very simple assumptions. Given the error probability, in Section 2.3 I derive the channel capacity for information transmission. I then revisit the channel error model in Section 2.4 looking more closely at the physics of radio transmission. Finally in Section 2.5 I review some advanced techniques for electronics design that help reduce errors in radio channels.

Communication is usually based on the propagation of electromagnetic waves. A wave is a disturbance that travels away from its source. Electromagnetic waves are generated by changing currents of electricity. When the molecules of matter vibrate with the energy of heat, relative motions of charges in them can emit electromagnetic radiation that can be described in terms of a stream of photons, each traveling in a wave-like pattern, moving at the speed of light and carrying some



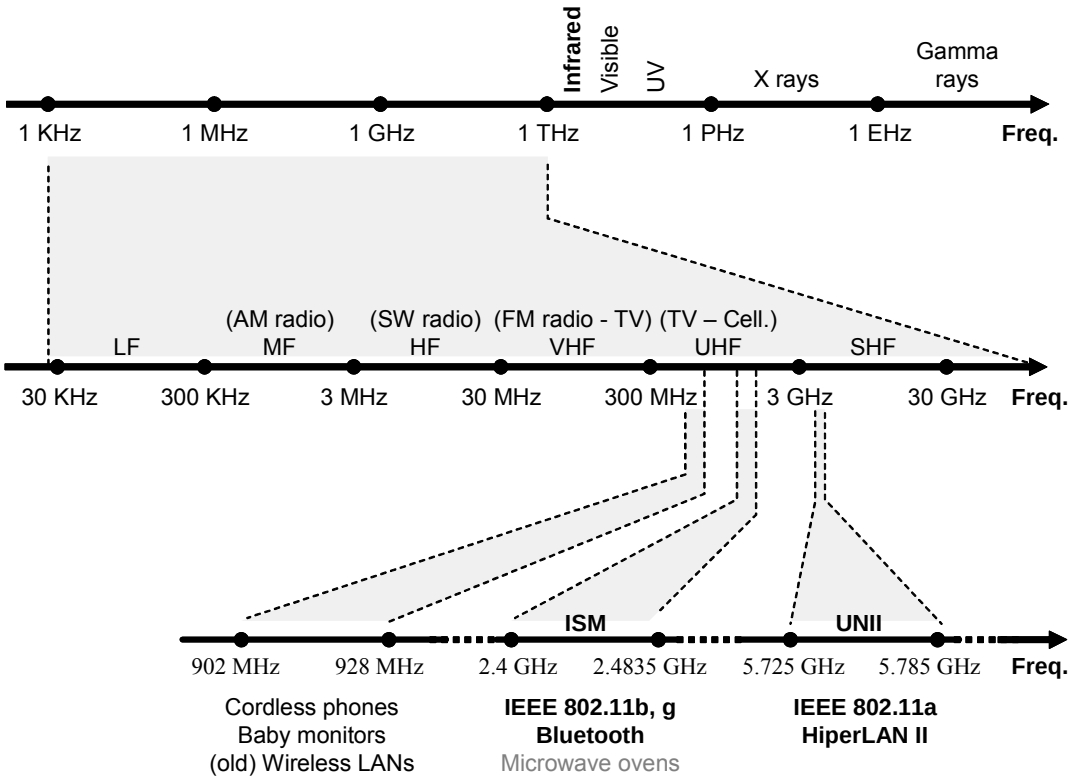


Figure 2-1: The electromagnetic spectrum allocation. ISM = Industry, Science & Medicine, UNII = Unlicensed National Information Infrastructure.

amount of energy. A property of a periodic wave that describes how many wave patterns or cycles pass by in a period of time is called *frequency*, *f*. Frequency is measured in Hertz (Hz), where a wave with a frequency of 1 Hz will pass by at 1 cycle per second. Another property of a wave is its *wavelength*, λ, which is the distance between adjacent peaks (crests). Since the wave travels the distance, λ, in a time for a complete vibration of the source, $1/f$, its speed, v, is given by the expression: $v = \lambda f$. [Wave velocity λ is also known as the phase velocity.]

Figure 2-1 shows the full range of frequencies (electromagnetic spectrum) and its uses for communication. Frequencies below 1 GHz are usually referred to as radio frequencies, and higher frequencies are referred to as microwave. Radio frequencies are further divided at 30 MHz. Below 30 MHz, long-distance propagation is possible by reflection from the ionosphere. Above 30 MHz, the ionosphere is transparent to electromagnetic waves and propagation is on line-of-sight paths. However, not all materials behave the same way to both light frequencies and radio frequencies, and radio frequencies can penetrate different materials to a different extent. In a sense, the waves with millimeter wavelengths are quasi-optical. For example cardboard is transparent to radio waves and is opaque to (blocks) light waves. Dielectrics such as cardboard, paper, clear glass, Teflon, some plastics, pure water and many building materials have low attenuation coefficients and radio waves reflect from them as well as pass through them. The prime portion of radio spectrum consists of the frequencies running from 30 to 3,000 MHz. These frequencies can penetrate buildings (which is why radios and cell phones work indoors) and transmit over long distances at low power (which lets batteries run longer and limits potentially harmful radiation). At higher frequencies a signal that is generated outdoors does not penetrate into buildings, and the signal generated indoors stays confined to a room. In all countries the

government regulates the spectrum licensing and the usage differs from country to country. Wireless networks operate in both licensed and unlicensed parts of the spectrum. For example, wireless telephone networks operate in the bands licensed to the telecommunications networking companies.

Of particular interest are the unlicensed bands, which are free to anyone to transmit without the need of a license or specific spectrum allocation from the spectrum regulatory body. The government mandates the maximum transmission power and sometimes the form of transmission. For example, to minimize interference between the uncoordinated devices, the devices in the unlicensed bands may be required to use wideband digital modulation (spread spectrum) technologies. Similar rules apply in different countries. Two important unlicensed bands for public wireless communications are ISM (Industrial, Scientific, Medical) and UNII (Unlicensed National Information Infrastructure).

The location of the ISM bands varies somewhat from country to country. Table 2-1 shows their location in the United States. Garage door openers, cordless phones, radio-controlled toys, wireless mice, and numerous other wireless household devices use the ISM bands. UNII bands are created to provide high-speed wireless networks that work over a short range. Typical uses are transfer of data, voice, graphics, teleconferencing, videoconferencing, and other multimedia services.

Table 2-1: Unlicensed spectrum location in the United States.

Unlicensed Bands	Spectrum	Typical Applications
ISM: Industry, Science and Medicine 902-928 MHz, 2.4-2.4835 GHz & 5.725-5.85 GHz	234.5 MHz	Cordless phones, (old) Wireless LANs (WLAN) and Wireless PBXs (WPBX)
UPCS: Unlicensed PCS Asynchronous: 1910-1920, 2390-2400 MHz Isochronous: 1920-1930 MHz	 20 MHz 10 MHz	 WLAN (IEEE 802.11b, Bluetooth) WPBX (Microwave ovens)
UNII: Unlicensed National Information Infrastructure UNII (5.15-5.25 GHz) UNII (5.25-5.35 GHz) UNII (5.525-5.825 GHz)	 100 MHz 100 MHz 100 MHz	IEEE 802.11a, HiperLAN II Indoor applications: WLAN, WPBX Short outdoor links, campus applications Long outdoor links, Point-To-Point links
Millimeter Wave (59-64 GHz)	5 GHz	Home networking applications

Radio spectrum is considered a scarce resource. Wired media provide us with an easy means to increase capacity—if affordable, we can lay more wires where required. With the wireless medium, we are restricted to a limited available band of operation, and we cannot obtain new bands or easily duplicate the medium to accommodate new users. Most of the prime spectrum is already licensed for different uses. However, there are indications that the majority of the licensed spectrum goes unused most of the time, thus creating an artificial shortage of spectrum, and there are efforts to free this spectrum for wireless networks, see e.g., [Woolley 2002]. Even if these efforts succeed, it will not happen in the near future and if it eventually does happen, the amount of freed spectrum will greatly vary from country to country. Thus, the designers of mobile applications must continue working under the assumption that the spectrum is a scarce resource.

A digital radio can either transmit a continuous bit stream or group the bits into packets. The latter type of radio is called a *packet radio* and is characterized by *bursty* transmissions: the radio is idle except when it transmits a packet. The first packet radio network, ALOHANET, was

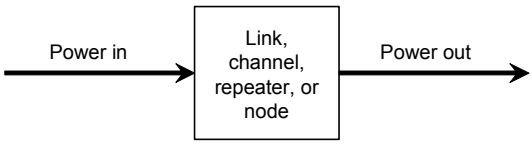


Figure 2-2: Description of model for defining decibel (dB).

developed at the University of Hawaii [Abramson 1970]. It enabled computer sites at seven campuses spread out over four islands to communicate with a central computer on Oahu via radio transmission. ALOHANET incorporated the first set of protocols for channel access and routing in packet radio systems. The underlying principles of these protocols are still in use today.

2.1.1 Decibels and Signal Strength

Communication systems handle information in the form of a *signal*—digital information must take analog form in a real world system, be it smoke signal, flashlight signal, or electrical signal. An important parameter in any transmission system is the signal strength, usually expressed in deciBels. *Decibel* or dB is the logarithm of ratio between two signal power levels. The reason for using dB rather than a linear unit is that all computation reduces to addition and subtraction rather than multiplication and division. Also, many things in nature obey logarithmical laws: signal strength often decays logarithmically; human sensory perceptions of brightness and loudness both depend logarithmically on physical stimuli; etc. For this reason the decibel scale of loudness, invented by the telephone pioneer Alexander Graham Bell (1847-1922), is also logarithmic. Decibel was originally a unit of sound, but now is commonly used to represent relative voltage or power gain. Every link, station, repeater, or channel can be treated as a *black box* (see Figure 2-2) with a particular decibel gain. The decibel gain of such a black box is given by

$$G_{dB} = 10 \cdot \log_{10}\left(\frac{P_{out}}{P_{in}}\right)$$

where P_{in} and P_{out} are the power of the input and output signals, respectively. This corresponds to the *relative* gain in output power with respect to the input power. For example, a gain of 3 dB means that the power has doubled. If the ratio in the above equation is negative, it is a decibel *loss*. For example, a gain of −3 dB means that the power has halved, indicating a loss of power.

Output power level is generally given in dBm, defined by the IEEE as decibels above or below one milliwatt. If the reference level is chosen to be P_1 = 1 mW, then $G_{dBm} = 10 \log_{10}(P_2 / 1\text{mW})$. By using dBm as a measure, it is relatively easy to calculate the cumulative effect of all passive and active system components. A unit's receiver sensitivity provides a measure of how weak a received signal can be while still providing error-free communications. Note that the decibel (dB) is the logarithm of a power ratio and not a unit of power, while dBm is a unit of power in the logarithmic system of numbers.

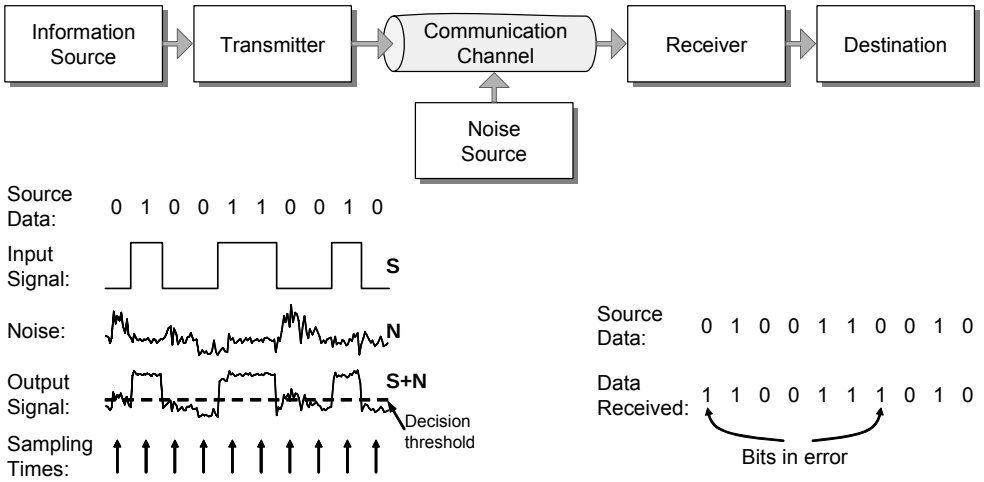


Figure 2-3: Shannon's diagram of the communication process.

2.2 Channel Implementation

A *channel* is the path, or link, that interconnects transmitter and receiver (see Figure 2-3). Ideally, communication channel should accept messages at the input and deliver them *unaltered* at the output. Unfortunately, every real channel suffers from noise, to a different degree. Noise is probably the most important problem of communication. The problem with noise is that it is unpredictable—it is a random process. Its effect on signal at the receiver is that one symbol could be mistaken for another.

Signals found in communication systems are complex waveforms. Any signal can be represented through Fourier transform as a sum of sine waves of different frequencies. Consider a signal with sinusoidal component frequencies extending from near zero to some upper limit, *B*. Such a signal is commonly called a *baseband signal*, and examples are audio signal from a microphone or a video signal from a camera.

2.2.1 Transmission Rate

For digital information, the *data rate* is the rate, in bits per second, at which data are communicated. The maximum rate at which data can be transmitted over a given communication channel, under given conditions, is referred to as the *channel capacity*. Under ideal conditions—a noiseless channel with no impairments to the signal—the channel capacity is limited only by the *bandwidth* of the transmitted signal[2]. The Nyquist's sampling theorem states that a signal of bandwidth *B* Hz can have 2*B* independent and successive amplitudes each second. In other words, a waveform low-pass limited to frequencies at most *B* Hz is completely determined by its values

[2] The *bandwidth* of a signal is the range of frequencies that must be used to represent it.

at each 1/(2*B*) seconds. Conversely, if it is required that a signal waveform have successively 2*B* independent amplitudes in a second, it must have a bandwidth of *B* Hz.

Higher transmission frequencies ideally offer greater channel capacity since it is easier to assign a greater bandwidth to the communication channel. This can be seen from the equation that links frequency and wavelength, $f = c/\lambda$. By taking a difference (discrete derivative) with respect to λ, we obtain

$$\frac{\Delta f}{\Delta \lambda} = -\frac{c}{\lambda^2} \tag{2.1}$$

By taking the absolute value, we have $\Delta f = (c \times \Delta\lambda)/\lambda^2$. Thus, given the width of a wavelength band, $\Delta\lambda$, we can compute the corresponding frequency band, Δf, and from that the data rate the band can produce. The wider the band, the higher the data rate. However, it is much more difficult to design a transmission system with high signal frequency and propagation characteristics of these frequencies unfavorable, as mentioned above.

Note that the Nyquist theorem specifies the *signal* rate. If the signals to be transmitted are binary (two voltage levels), then *B* Hz signal supports the maximum data rate, i.e., capacity, $C = 2B$ bps. However, signals with more than two levels are used, where each signal element can represent more than one bit. With multilevel signaling, the Nyquist limit is:

$$C = 2B \cdot \log_2 M \tag{2.2}$$

where *M* is the number of discrete signal (e.g., voltage) levels. Although increasing the number of different signal elements increases the data rate, it also increases the complexity of the receiver, which now must distinguish one of *M* possible signals.

2.2.2 Symbols To Signals

The communication process can be summarized as in Figure 2-4. The techniques commonly employed in the physical layer of wireless networks are modulation and error-control encoding of source signals. *Modulation* is the transformation of messages into signals suitable for transmission over the medium. *Encoding* is an accurate representation of one thing by another. As will be seen below, modulation involves encoding but does not include redundancy as a means of error-control. The latter is the task of channel encoding. Both modulation and error-control encoding are employed at the transmitter and inverse processes, demodulation and decoding are employed at the receiver.

Notice that some communication systems do not employ modulation. For example, digital logic circuits communicate in the baseband, both within the VLSI chip and between the chips in the computer. Also, many wired networks communicate in the baseband, e.g., most Ethernet (IEEE 802.3) networks. In such cases, the rectangular pulse waveform is placed directly on the medium.

2.2.3 Modulation

Modulation is the general name for the process by which a signal is transformed into a new form, with bandwidth translated and increased or decreased, while preserving the information content in a retrievable form. It is needed since at the source the messages are rarely produced in a form

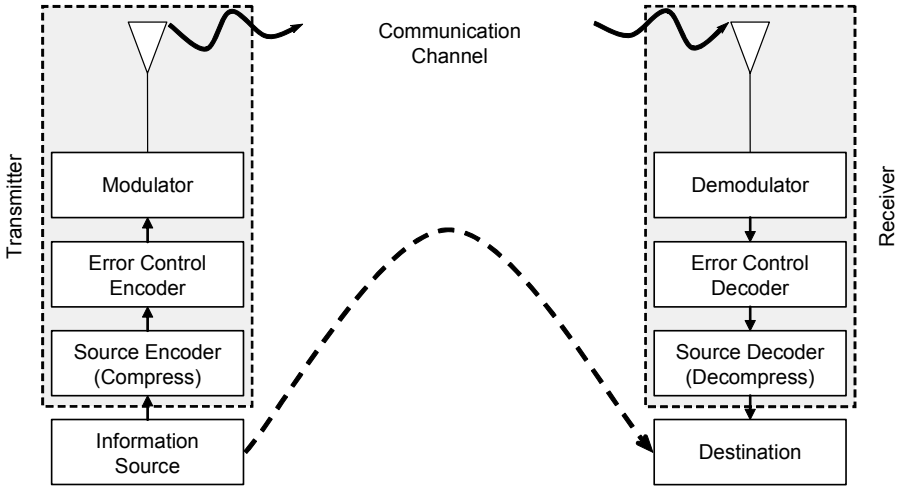


Figure 2-4: Components of a wireless transmission and receiving system.

suitable for direct transmission over a channel. Modulation shifts the signal from its original frequency range to another part of the frequency spectrum. A signal can be shifted up or down in frequency range by shifting the frequencies of its component sine waves. Shifting a signal from its baseband range to a higher frequency range is a technique commonly used to comply with regulatory constraints on radio spectrum usage (Figure 2-1). It is also easier to assign a greater bandwidth at higher transmission frequencies, Eq. (2.1), which may be needed to add redundancy in channel encoding.

Modulation is a process of combining an input signal and a carrier at frequency f_c to produce a signal whose bandwidth is usually centered on f_c. Input signal can be either an analog or a digital signal. In the analog case the amplitude, frequency, or phase of the carrier varies continuously in response to the message waveform. However, in the binary digital case these three parameters switch between one of two possible values, depending on whether a 0 or 1 pulse is transmitted. Examples of signal modulation are shown in Figure 2-5. The left part shows modulation of analog signals. An *amplitude-modulated* (AM) signal is a high-frequency wave whose peak amplitudes vary in accordance with an input signal. Analogously, in *frequency modulation* (FM), the instantaneous frequency of the carrier is varied proportionally to the amplitude of the input signal. In *phase modulation* (PM), the phase angle of the carrier is varied proportionally to the amplitude of the input signal.

For digital input signals, the amplitude of the carrier sine wave is shifted from one level to another depending on whether a logic 1 or a logic 0 is sent, which is called amplitude-shift keying (ASK). Similarly, in *frequency-shift keying* (FSK) the frequency can take on one of two predetermined constant frequencies. In *phase-shift keying* (PSK) the phase can be one value or another and the signal transitions the phase shifts by 180°.

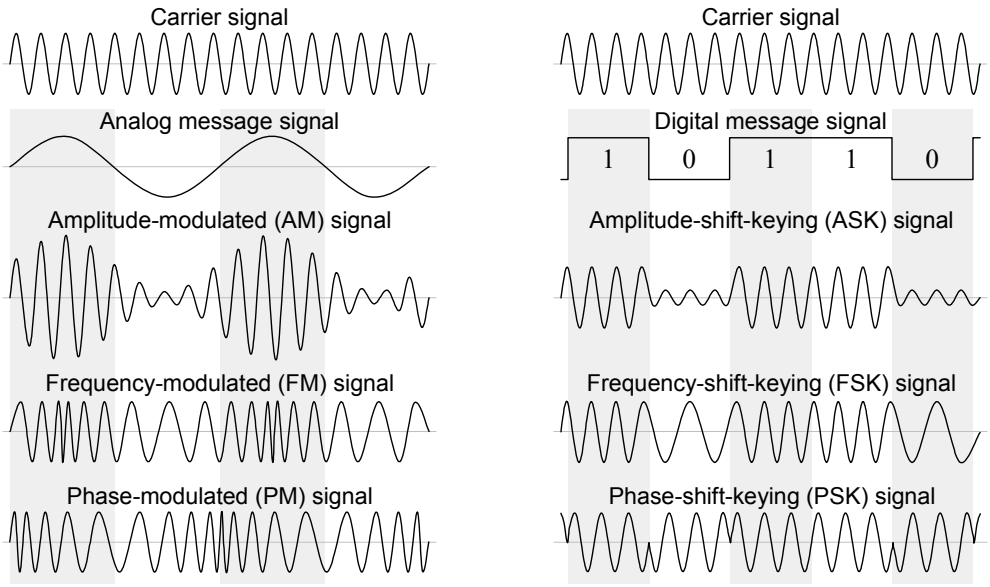


Figure 2-5. Modulation of analog carrier signals for analog and digital messages.

PSK is a preferred modulation technique in wireless communications. The absolute phase of the waveform is not relevant in PSK; only changes in the phase encode data. PSK is not susceptible to the noise degradation that mostly affects amplitude and thus ASK, nor does it have the bandwidth limitations of FSK. This implies that a receiver can detect small variations in the signal. Therefore, instead of using only binary variations (BPSK), one can establish, for example, quadruple (QPSK) and octal phase variations (8-PSK). One can represent the phase and amplitude relationships by a constellation or phase-state diagram, as illustrated in Figure 2-6(a) and Figure 2-6(b), where different combinations for amplitude and phase represent different symbols. Thus, QPSK works with four symbols, represented by binary numbers as "00," "01," "10," and "11." Generally, in *M*-ary PSK (Figure 2-6(c)), the carrier phase takes on one of *M* possible values, namely, $\theta_i = 2(i - 1)\pi / M$, and it works with *M* symbols.

The simplest form, BPSK, uses two carrier waves, shifted by a half cycle (i.e., half wavelength) relative to each other. One wave, the reference wave is used to encode a "0," and the half-cycle shifted is used to encode a "1." Since absolute phase is difficult to determine, the phase is usually shifted from one value to another and the difference in phase is measured. This method is called *differential phase-shift keying* (DPSK).

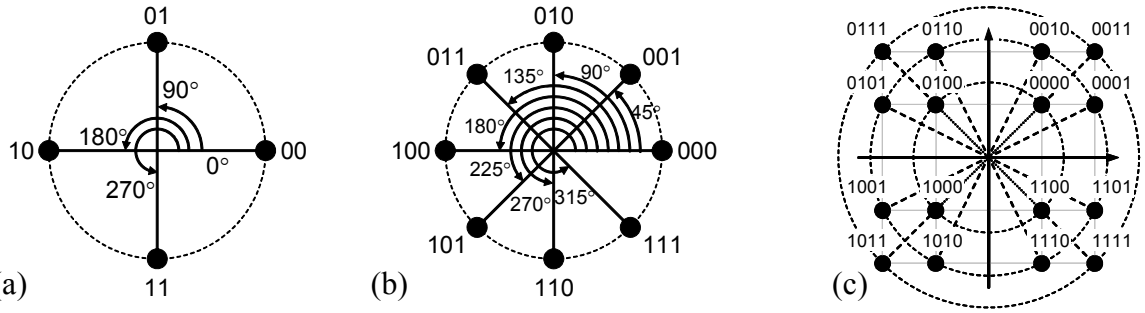


Figure 2-6. Representation of the phase and amplitude relationship by a constellation of symbols, or phase-state diagram, for QPSK (a) and 8-PSK (b). (c) Constellation diagram of an *M*-ary QAM (*M*=16, 3 amplitudes + 12 phases) signal set. [Note that there are other possible 16-QAM constellations].

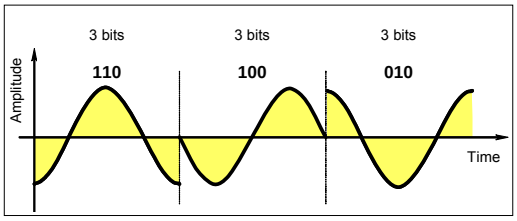
As a further variation one can combine ASK and PSK to form *quadrature amplitude modulation* (QAM). Figure 2-6(c) shows an example of 16-QAM, where sixteen possible combinations of amplitude and phase allow the encoding of a pulse carrying four bits.

Example 2.1 8-PSK

Suppose 8-PSK modulation is employed for a binary sequence "110100010." Provide the phase states and the coordinates of message-symbol points in a tabular form. The first column should show the tribits (3-tuples of bits) of the input message, the second column the phase-space states (polar coordinates: amplitude and phase angle), and the third column the Cartesian coordinates of message points. Draw the waveform of the modulated signal.

The solution is as follows:

Tribit	Phase States	Cartesian Coordinates
110	$\varphi_1 = 270°$, $r_1 = 1$	$x_1 = \sin 270°$, $y_1 = \cos 270°$
100	$\varphi_2 = 180°$, $r_2 = 1$	$x_2 = \sin 180°$, $y_2 = \cos 180°$
010	$\varphi_3 = 90°$, $r_3 = 1$	$x_3 = \sin 90°$, $y_3 = \cos 90°$



The diagram on the right shows one carrier cycle per three bits, which is only to simplify the illustration. In reality, the carrier frequency is orders of magnitude higher than the bit rate, so we normally have hundreds or thousands of cycles per bit.

What should come across from the above examples is that in modulation we are associating a sequence of bits (binary representation of a symbol) to a point in a three-dimensional space. The dimensions of the 3D space are defined by the aspects of the analog carrier signal: amplitude, frequency, and phase. The number of different symbols that can be encoded corresponds to the number of valid points to choose from. For example, with 16 valid points (Figure 2-6(c)), we can encode 16 different symbols or 16 different sequences of 4 bits.

The most critical modulation design parameter is the separation distance of the encoded symbols. When several symbols (*M*) are encoded in a narrow bandwidth, they should be carefully separated to minimize the cross-correlation—or symbol overlap—between them, defined as the cross-correlation metric. A well-designed distancing of symbol decreases the chance that noise will confuse the receiver when deciding on which symbols was transmitted. The system will have

a consistent performance and operating range without demanding excessive transmitter power. With the present electronic technologies, the maximum feasible number of different symbols is 64, which corresponds to 6 bits per symbol.

If symbols are transmitted one at a time, non-simultaneously, why does this overlap matter? To detect symbols, a digital receiver will correlate the energy of the phase, frequency, or amplitude, depending on the modulation scheme. This is a straightforward process when the symbol energy is well above the noise—there will be few errors in choosing one symbol from another. However, in a low-signal, high-interference environment, this gets difficult for the receiver, unless enough separation exists between symbols to make the proper choice.

Orthogonal symbol sets are necessary to achieve optimal symbol separation. If all the symbols cannot be orthogonal, then at least they should be consistently spaced to minimize differences between their metric values.

Symbol separation is a key parameter for predicting the probability of bit error of a digital system. The probability of bit error is proportional to the distance between the closest points in the constellation. This probability is calculated using the statistical *Q*-function to relate the cross-correlation and bit energy-to-noise ratio to the bit error performance (BER, defined below). This is explained next.

2.2.4 Noise and Error Probability

Electrical *noise* may be defined as any unwanted energy that accompanies a signal in a communication system. Noise in communication systems falls into two types. Some is manmade (or artificial) and could be eliminated through better design, perhaps by suppression at source, or screening of offending sources or sensitive circuit elements. Artificial noise arises from sources such as electrical machinery and switches. The other type of noise occurs naturally and is unavoidable. Natural sources include cosmic noise (e.g., due to geomagnetic storms) and atmospheric noise (e.g., due to lightning discharges). In addition, many electrical and electronic components naturally introduce noise into a system. Every physical body at a temperature greater than absolute zero radiates electromagnetic waves due to thermal agitation of electrons, which is termed *thermal noise* because its energy increases with temperature. Due to its additive nature, another name for thermal noise is additive white Gaussian noise (AWGN). The word "white" implies that the noise contains all frequencies equally. Gaussian comes from the fact that the noise amplitudes are distributed according to the Gaussian distribution. Other types of natural noise include shot noise, which occurs in active devices (diodes and transistors), and flicker or pink noise, which occurs in semiconductors.

The noise may be added to the signal, in which case it is called *additive*, or the noise may multiply the signal, in which case the effect is called *fading*. The multiplicative nature of a channel means increasing signal power may not yield a proportional improvement in performance.

Noise is usually expressed using a measure called *noise figure*, which is the ratio of a device's (or channel's) input signal-to-noise ratio (SNR) to its output SNR expressed in decibels:

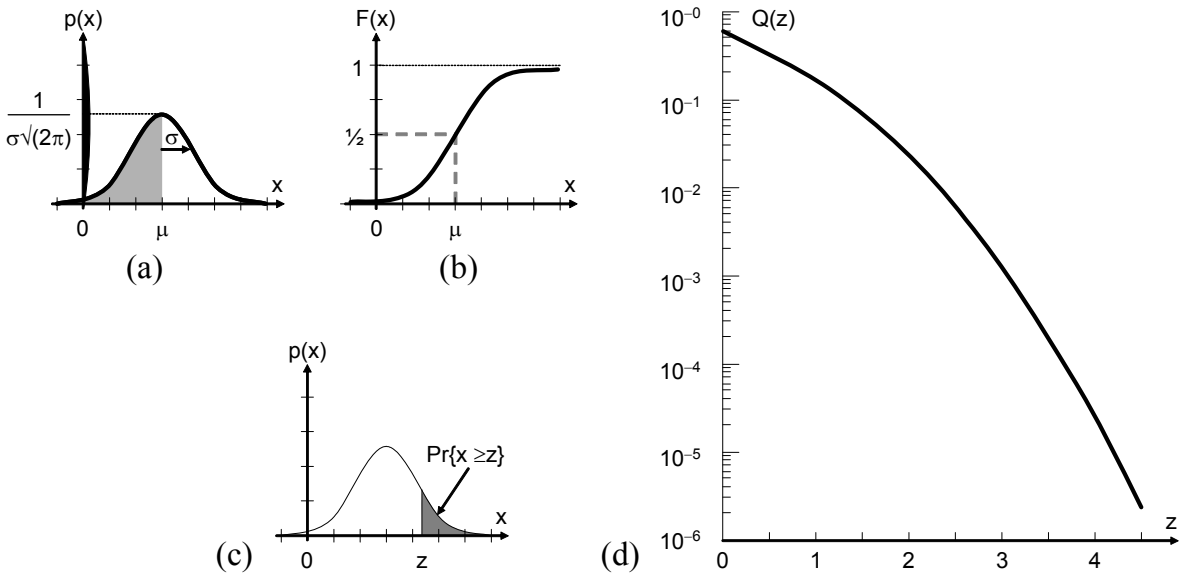


Figure 2-7: (a) Probability distribution function of a Gaussian random variable. (b) Cumulative distribution function of the same r.v. (c) $Q(z) = Pr\{x \geq z\}$ for a Gaussian r.v. is represented by the shaded region. (d) A plot of the *Q*-function. [Note the logarithmic scale.]

$$NF = 10 \cdot \log_{10}\left(\frac{S_{in}/N_{in}}{S_{out}/N_{out}}\right) \text{ dB} \tag{2.3}$$

For every symbol transmitted over a noisy channel, there is a non-zero probability that the receiver will make mistakes in interpreting the symbol from the received signal. Given the signal energy, we are interested in knowing what the probability is that noise will result in a bit error. The *bit error rate* (BER) is the percentage of bits in error relative to the total number of bits received in transmission.

The simplest case of noise, which is always present, is AWGN. The power levels of AWGN form a Gaussian random variable. The *probability distribution function* (pdf) of a Gaussian or normally distributed random variable (r.v.) is:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean and σ^2 is the variance of the random variable. The abscissa in Figure 2-7(a) represents the noise amplitude length and the ordinate represents the probability of occurrence. The *cumulative distribution function* (cdf) of the same random variable is:

$$F(x) = \int_{-\infty}^{x} p(u)du$$

Both pdf and cdf of a Gaussian r.v. are illustrated in Figure 2-7. An important concept is the *Q*-function, which is a convenient way to express right-tail probabilities for normal (Gaussian) random variables. Its use in communications will become apparent below. For a real number *z*, $Q(z)$ is defined as the probability that a standard normal random variable (zero mean, unit variance) exceeds *z*:

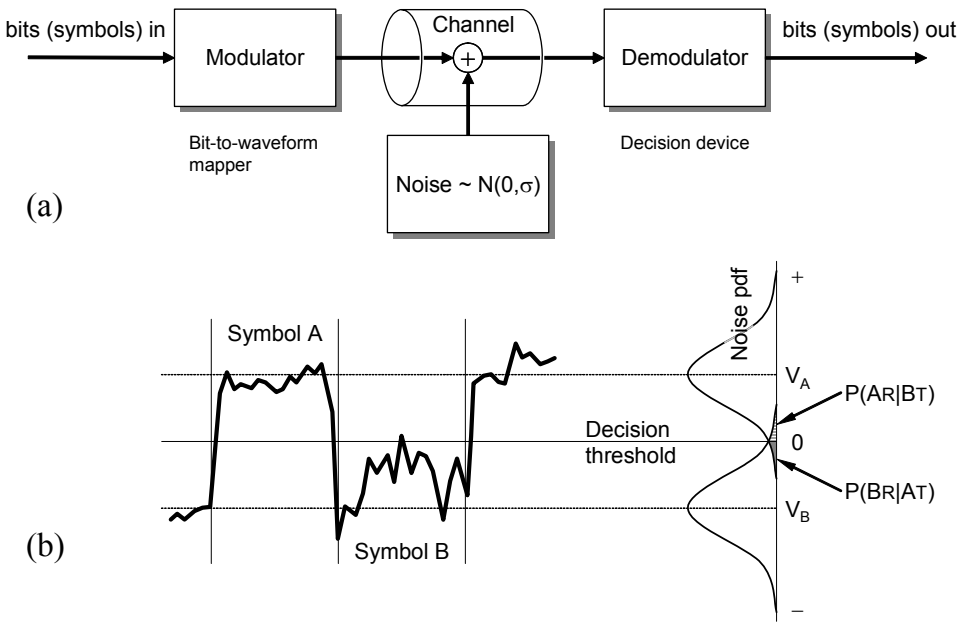


Figure 2-8: (a) Model for received signal passed through an AWGN channel. (b) Effect of noise on a binary signal (BPSK modulation, two symbols, "A" and "B").

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{y^2}{2}}\, dy, \qquad z \geq 0 \tag{2.4}$$

Q is a mapping from a real-numbers line to $\{0,1\}$. One may also define $Q(-\infty) = 1$ and $Q(\infty) = 0$. If $F(z)$ denotes the cumulative distribution function of a standard normal, then clearly $Q(z) = 1-F(z)$. For this reason Q is also called the *complementary cumulative distribution function*. The integral in (2.4) cannot be evaluated in closed form for arbitrary z, and it is evaluated by numerical methods. The Q-function is useful because it offers a concise notation for the tail probability integral. [A note of caution: Some authors define the Q-function in a different way, one alternative being $Q(z) = F(z) - F(0)$. In addition, the name Q-function is used for several different things, so some authors prefer the "tail probability" designation.]

Figure 2-8 shows the model of signal transmission through an AWGN channel and effects of noise on signal. In the case of binary signaling using BPSK modulation there are two symbols "A" and "B," corresponding to a logic 1 or a logic 0. The pdf for a zero-mean Gaussian noise is superimposed upon the nominal signal levels V_A and $-V_B$ (also called antipodal symbols). If, when symbol A is being transmitted, the associated, the accompanying noise voltage amplitude is less than $-V_A$ at the moment when the receiver makes decision, the signal will be interpreted as that of symbol B and an error produced. Let A_T denote the event of symbol A being transmitted and B_R denote the event of symbol B being received. Then, the joint probability of these two events can be expressed via a conditional probability:

$$P(A_T, B_R) = P(B_R \mid A_T) \cdot P(A_T)$$

This is the probability of an error if symbol A is transmitted. In a similar manner we may deduce that the probability of an error when symbol B is transmitted is given by:

$$P(B_T, A_R) = P(A_R \mid B_T) \cdot P(B_T)$$

In a general communication system it is reasonable to assume that symbols A and B are equiprobable, and thus $P(A_T) = P(B_T) = 0.5$ and the decision threshold is set midway between voltage levels V_A and $-V_B$. Therefore the total probability of an error in a BPSK-based system is:

$$P_e = P(A_T, B_R) + P(B_T, A_R) = P(B_R \mid A_T) \cdot 0.5 + P(A_R \mid B_T) \cdot 0.5 = P(B_R \mid A_T) = P(A_R \mid B_T)$$

The last equality is due to the symmetry of the Gaussian pdf. As indicated in Figure 2-8(b), this is the tail probability of a Gaussian pdf, i.e., the probability that noise amplitude exceeds the signal voltage, which is given by the Q-function (assuming that $|V_A| - |V_B| = 2 \cdot V$):

$$P_e = Q(V/\sigma) = Q\left(\sqrt{\text{SNR}}\right) = \text{BER}$$

where σ is equivalent to the rms of the noise voltage and $z^2 = (V/\sigma)^2 = \text{SNR}$ represents the *signal-to-noise* ratio per bit. This probability is the *bit error rate* (BER) and looking at the chart in Figure 2-7(d), a BER of 10^{-6} simply means that on average one bit will be in error out of every million transmitted. Note how rapidly $Q(z)$ decreases as z increases—this leads to the threshold characteristic of digital communication systems. The Q-function is maximum, ½ (50% errors), when its argument is zero (no signal), and approaches its minimum value (almost zero errors) when the signal is much greater than the channel noise.

RMS VALUE

♦ The concept of "root-mean-square value" (rms value) of noise is based on the fact that the "mean-square value," $\overline{V_n^2}$, is proportional to the "noise power," N_0. The square root of noise power, $\sigma = \sqrt{N_0}$, is the rms value of the noise voltage, which is the "effective value" of the noise voltage. For the sine wave, the rms value is equal to 0.707 times the peak value of the sine wave. This is the equivalent DC value of the sine wave.

In digital communication systems we usually express SNR in terms of the average signal energy per transmitted bit, E_b, and single sided noise power spectral density, N_0. We have:

Signal_power = Energy_per_bit × Data_rate, or $S = E_b \cdot r_b$

Noise_power = Single_sided_noise_PSD × Noise_bandwidth, or $N = N_0 \cdot B_N$, or equivalently using the double sided noise PSD, $N = ½\, N_0 \cdot (2B_N)$

If we assume that the noise bandwidth equals the minimum possible bandwidth for the digital communication system, i.e., $B_N = r_s/2 = r_b/2$ Hz, then $(V/\sigma)^2 = S/N = (E_b \cdot r_b) / (N_0 \cdot 0.5\, r_s)$. Therefore, $(V/\sigma)^2 = 2 \cdot E_b / N_0$, so $P_e = Q\left(\sqrt{2 \cdot E_b / N_0}\right)$.

There are two important observations to make. First, the error probability depends upon the message only through its energy E_b. The waveform is of no consequence. Second, the probability of error may be expressed in terms of symbol separation. In the above example, the distance between the two symbols A and B is $d_{AB} = 2V = 2\sqrt{E_b}$ and by substitution, $P_e = Q\left(\sqrt{d_{AB}^2 / 2N_0}\right)$. The symbol separation metric scales the E_b/N_0 ratio up or down and adjusts the Q-function's argument, improving or degrading P_e. Binary antipodal symbols (M=2), like BPSK, have a maximum symbol separation angle of 180° that scales the E_b/N_0 ratio by a factor of 2 and

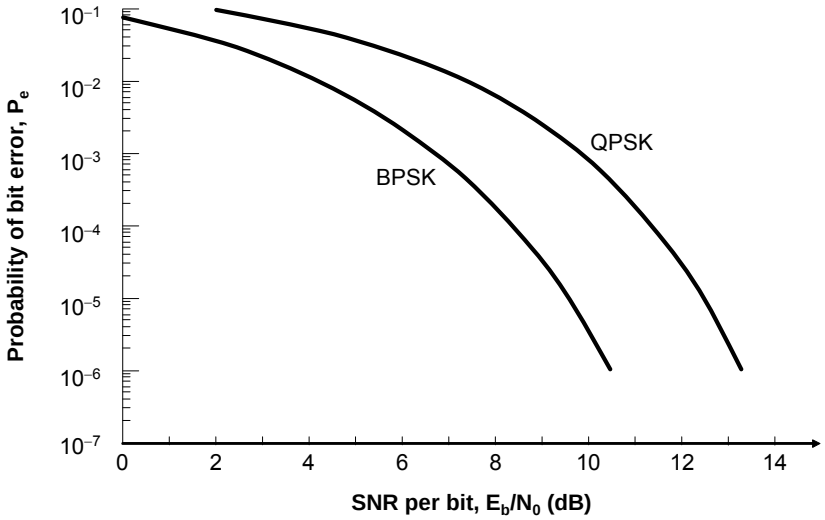


Figure 2-9: Probability of error (BER) for binary signals, also called *Q*-functions for BPSK and QPSK modulation schemes.

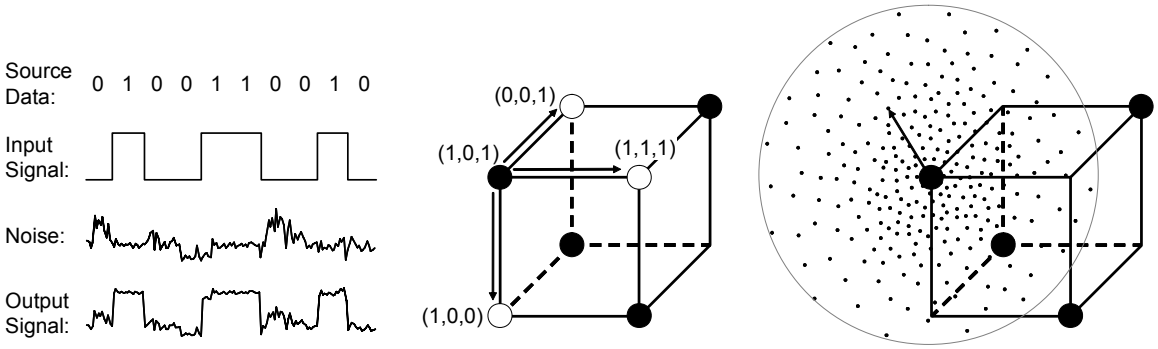
improves the P_e. The receiver derived here is called the *optimum binary receiver* when the channel noise is white.

BPSK antipodal modulation has one of the best P_e performance curves in digital design because of this large symbol separation. For *M*-ary multiple symbol schemes (M>2), however, the distance between the symbols is smaller than for BPSK, see Figure 2-6. Accordingly, their P_e performance curves are worse than for BPSK. For example, for the case of binary orthogonal signals of QPSK, $P_e = Q\left(\sqrt{E_b/N_0}\right)$. This means that QPSK requires a factor of two increase in energy to achieve the same error probability as BPSK. The error probability vs. $10 \cdot \log_{10}(E_b/N_0)$ for these two types of signals is shown in Figure 2-9. Notice that these theoretical *Q*-functions are somewhat different from empirical BER vs. SNR functions obtained for actual transceiver chips, see Section 4.3 below.

The above analysis derives the channel error probabilities P_e for an AWGN channel under different modulation techniques. Error probability on a channel can be derived analytically, as above, for simple channels, or it must be modeled from empirical observations for complex channels. Given the error probability, we next look at the channel capacity for transmitting information. Then, in Section 2.4 we will consider in more detail the radio channel and see why it does not fit the AWGN model. Lastly, in Section 2.5 we describe some advanced techniques designed to reduce P_e for radio channels.

2.3 Continuous Noisy Channel Capacity

Physical channels carry continuous waveforms, so the discrete symbols must be converted to signal waveforms. Physical channels are also noisy, so the received signal differs in an unpredictable way from the signal transmitted (see Figure 2-3). The presence of noise can corrupt



(a) (b) (c)

Figure 2-10. (a) Symbol-to-signal conversion and superposition of noise in transmission. (b) For a discrete message, noise can only "displace" it to another discrete message. (c) For a continuous message, noise can "displace" it anywhere in the space.

one or more bits (symbols). Figure 2-10 illustrates the effect of noise and contrasts discrete and continuous channels. Unlike discrete bit sequences, where the distance can be measured by Hamming metric (Chapter 1), the distance of continuous waveforms is usually measured by Euclidean metric.

Since we deal with continuous random variables, it would be desirable to extend the definition of entropy accordingly. Since the channel capacity derivation can be carried out without it, it is omitted and the interested reader should consult, e.g., [Lathi 1989; Wozencraft & Jacobs 1990].

Suppose that the power received over the channel is a signal power S and a noise power N from noise added during transmission. These are typically measured at a receiver, since this is where an attempt is made to process the signal and eliminate the unwanted noise. Shannon's theorem states that the maximum channel capacity, in bits per second, is:

$$C = B \cdot \log_2(1 + S/N) \tag{2.5}$$

For example, a channel of 3000-Hz bandwidth with a signal-to-thermal-noise ratio of 30 dB (typical parameters for the analog part of the landline telephone system) can transmit up to about 30,000 bps. This is the theoretical maximum, in practice much lower rates are achieved. One reason is that the formula assumes only white Gaussian noise (thermal noise). Impulse noise (due to lightning and switching effects), nonlinear distortions (due to electronic devices used to build the transmitter and receiver), etc., are not accounted for in (2.5). Each of these noise sources has different statistical properties from that assumed by Shannon's theorem. The principal difference is that the errors caused by these extra noise sources tend to occur in bursts of arbitrary length, rather than being spread continuously.

Formula (2.5) suggests that increasing either signal strength S (relative to the noise N) or bandwidth B can increase the data rate. However, as the signal strength increases so do the effects of nonlinear distortions in the system, leading to an increase in other (non-AWGN) types of noise. The higher the data rate, the shorter the time duration of the signal waveform for one bit, hence more bits are affected by a given noise waveform. Thus, for a specified noise pattern, the higher the data rate, the higher the error rate. And, the wider the bandwidth, the more white noise is admitted to the system. Thus, the data rate increase would not be as high as initially expected.

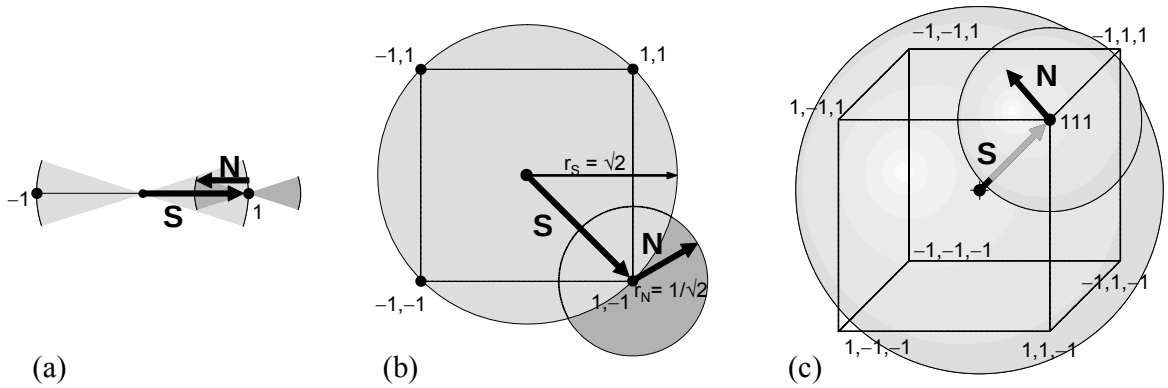


Figure 2-11: Geometrical representation of signals consisting of (a) one-, (b) two-, or (c) three binary symbols −1 and 1. The example assumes that $N = \frac{1}{2}S$.

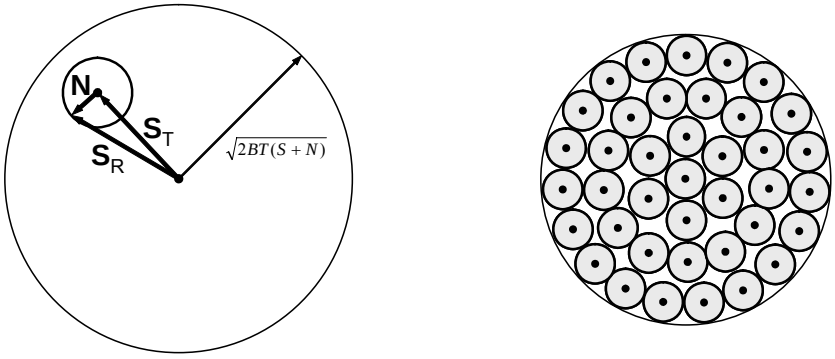
One way to deal with noise is to increase signal power—we all know that in a noisy environment we need to raise our voices to hold a conversation. This corresponds to the term S/N in Shannon's formula (2.5) and what matters is, obviously, signal power relative to noise or signal-to-noise ratio (SNR), rather than signal's absolute value. Clearly, when the signal exceeds the noise by an adequate margin communication may occur. Where signal and noise levels are similar, and certainly if noise power exceeds that of the signal, communication fails. The margin that would be suitable depends on the type of transmission and is generally determined by means of subjective tests. Typical values of acceptable SNR are 50–60 dB for high-quality music listening, 16 dB for low-grade speech and up to 30 dB for commercial telephony systems, and about 60 dB for good quality TV transmission.

Formula (2.5) also tells us that, in addition to increasing the signal strength, there is an alternative for increasing the transmission rate of a communication channel, that is, by increasing the channel bandwidth B while keeping SNR constant. The key here is to consider the effect of noise on messages, rather than individual symbols. A symbol is atomic and cannot have redundancy, whereas a message can; thus "messages" comprising a single symbol are disregarded. A message affected by noise is distorted to another message and the receiver's goal is to detect the distortion.

An analogy can be established between message signals and vectors, whereby a signal $s_i(t)$ consisting of a sequence of n symbols sampled at the times $t = 1, 2, \ldots, n$, can be represented as an n-dimensional vector (see Problem 2.3 at the end of this chapter). This suggests a geometrical representation of signals as well. Any vector $\mathbf{S}_i$ in this hyperspace is specified by n numbers (s_{i1}, s_{i2}, …, s_{in}) which represent the magnitude of components of $\mathbf{S}_i$ along the n basis vectors respectively. Figure 2-11 shows examples of one-, two-, and three-dimensional signals, consisting of one-, two-, or three binary symbols −1 and 1, respectively. Thus, if we let the amplitudes of the samples of a band-limited signal be the coordinates of a point in hyperspace, then this point represents the complete signal. The *energy* of a signal $s_i(t)$ is given by:

$$E_i = \int_{-\infty}^{\infty} s_i^2(t)dt = |\mathbf{S}_i|^2 \tag{2.6}$$

where $|\mathbf{S}_i|^2$ is the square of the length of the vector $\mathbf{S}_i$. Given all possible vectors of n symbols (samples), we can compute the average energy per signal. For example, for two-dimensional binary signals, the possible vectors are {(−1, −1), (−1, 1), (1, −1), (1, 1)}, and the average energy is $(2+2+2+2)/4 = 2$ and for three-dimensional binary signals it is $8\times 3/8 = 3$. Generally, the average



(a) (b)

Figure 2-12. (a) Signal space representation of transmitted signal S_T, received signal S_R and noise signal N. (b) For error-free communication, the valid signals must be chosen as the centers of the embedded noise spheres, and all the other points (signals) are declared invalid.

energy per binary signal is $\overline{E} = n$, which means that the average distance of a signal point from the origin is $\sqrt{n}$.

Let $S = \overline{E}/n$ denote the average energy per individual sample (symbol) of the signal and N denote the average energy of the white Gaussian noise added to the signal in the channel. [Symbols E_b and N_0 are sometimes used instead of S and N, respectively.] For a band-limited signal of bandwidth B, we know from Nyquist sampling theorem that we need $2B$ samples. The total number of samples received over a time T is $n = 2 \cdot B \cdot T$ and their total energy is $2 \cdot B \cdot T \cdot (S+N)$. The volume of an n-dimensional hypersphere of radius r is proportional to r^n. The received signal may be thought of as the transmitted signal surrounded by a spherical region of uncertainty due to the noise, Figure 2-12(a). Transmitted signals may be confused at the receiver if these regions of uncertainty overlap. Since the probability distribution function of the white Gaussian noise extends to infinity, they always overlap, but if the dimensionality of the hypersphere is high, practically all of the noises will lie within a rather sharply defined sphere of radius $\sqrt{2 \cdot B \cdot T \cdot N}$. The problem of reliable communication then reduces to the problem of packing the maximum number of small, noise spheres into the large signal-plus-noise sphere, see Figure 2-12(b). The upper limit of distinguishable messages we can transmit in time T is equal to the ratio of the signal-plus-noise sphere to the volume of the noise sphere:

$$m \le \frac{(r_S + r_N)^n}{r_N^n} = \left(\frac{\sqrt{2BT(S+N)}}{\sqrt{2BTN}} \right)^{2BT} = \left(\frac{S+N}{N} \right)^{BT} \tag{2.7}$$

Taking the logarithm of this number and with $T = 1$ sec, we obtain Eq. (2.5). The trick to redundant encoding is to leave out some of the possible signal points, i.e., to declare them invalid. The valid points are the centers of the noise spheres (N-spheres) of radius $\sqrt{2BTN}$, as shown in Figure 2-12(b). Another important observation from Eq. (2.7) is that for the same relative values of average signal and noise energy per symbol, the number of non-overlapping N-spheres that can be embedded in the $(S+N)$-sphere increases exponentially with the increasing dimension of space (i.e., message length).

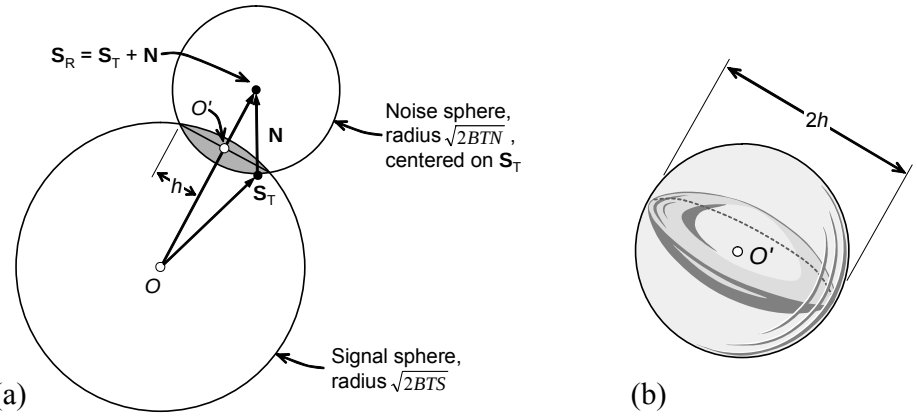


Figure 2-13: (a) Projection of the signal and noise hyperspheres with the intersection containing the error causing signals in the shape of a lens. (b) The enlargement shows the lens enclosed within a third hypersphere, of radius *h*, centered on *O′*.

To show that information can actually be transmitted at the rate of *m* messages, Eq. (2.7), I follow the argument proposed by Shannon. Instead of choosing *m* valid messages as the centers of non-overlapping spheres, Shannon proposed to select *m* points randomly located in the hypersphere of radius $\sqrt{2BTS}$. Although the *n*-dimensional vector $\mathbf{S}_i$ is equally likely to lie in any point within this sphere's volume, it is *not* equally likely to lie at any radius ρ, $0 \le \rho \le \sqrt{2BTS}$, as follows. For any $\Delta > 0$, the probability a signal lies in the sphere of radius $\sqrt{2BTS - \Delta}$ is equal to the ratio of the smaller sphere to that of the larger one:

$$P\left(|\mathbf{S}_i| \le \sqrt{2BTS - \Delta}\right) = \frac{K_n\left(\sqrt{2BTS - \Delta}\right)^n}{K_n\left(\sqrt{2BTS}\right)^n} = \left(1 - \frac{\Delta}{2BTS}\right)^{\frac{n}{2}}$$

This quantity approaches zero for increasing *n* (remember that $n = 2BT$). (K_n is a positive constant that depends only on *n*.) Therefore, most of the signal points are located around the surface of the hypersphere of radius $r = \sqrt{2BTS}$. Consider a particular transmitted signal $\mathbf{S}_T$ and a particular received signal $\mathbf{S}_R$, which consists of the transmitted signal plus random noise vector, $\mathbf{S}_R = \mathbf{S}_T + \mathbf{N}$, as illustrated in Figure 2-13(a). We shall use a maximum-likelihood receiver, which works as follows. If the received signal $\mathbf{S}_R$ does not equal one of the *m* "valid" messages, the receiver shall make the decision that "$\mathbf{S}_T$ was transmitted" provided none of the remaining $(m - 1)$ valid signal points are closer to $\mathbf{S}_R$ than $\mathbf{S}_T$. The transmitted signal lies, usually (not always, but in most of the cases!), near the surface of a sphere of radius $\sqrt{2BTS}$, and the received signal lies near the surface of a sphere of radius $\sqrt{2BT(S + N)}$.

Now combine the fact that a transmitted signal with high probability lies near the surface of the sphere of radius $\sqrt{2BTS}$ and, in order for it to result in $\mathbf{S}_R$, it must also lie within a sphere of radius $\sqrt{2BTN}$ centered on $\mathbf{S}_R$. The locus of the signal points that satisfy both of these conditions simultaneously is a lens, as shown in Figure 2-13(b). The volume of the lens is not easily computed, but it is certainly less than the volume of a sphere of radius *h*, where *h* is the distance indicated in Figure 2-13(a). This can be easily found to be:

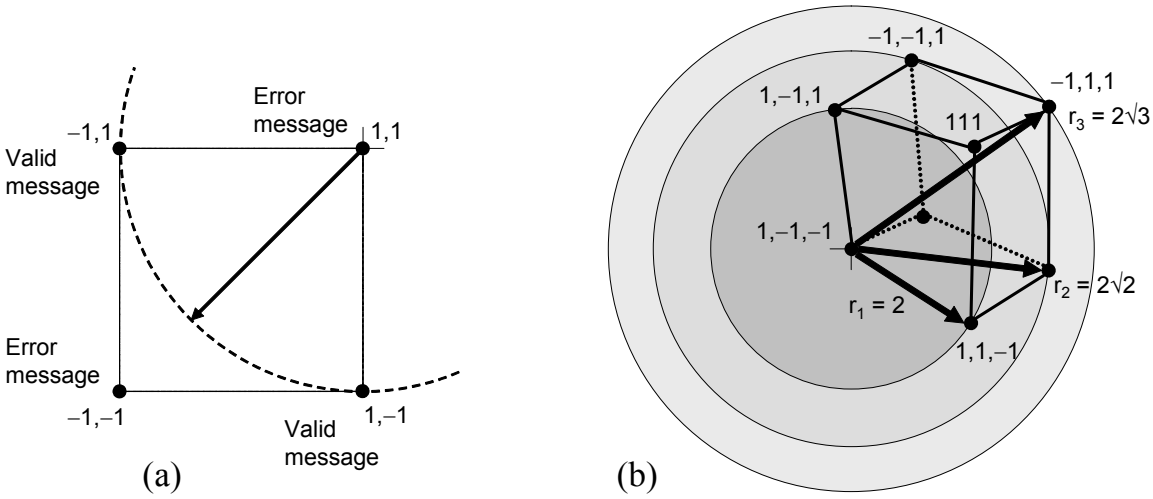


Figure 2-14: Error detection and correction. (a) In the two-dimensional case, the receiver is able to detect but not correct errors since all valid messages are equidistant from the error. (b) In the three-dimensional case, the receiver is able to both detect and correct some errors. See text for details.

$$h = \sqrt{2 \cdot B \cdot T \frac{S \cdot N}{S + N}}$$

Thus the probability of finding any signal in the lens is less than the volume ratio of spheres of radius h and $\sqrt{2BTS}$, or less than $[N/(S+N)]^{BT}$. Hence the probability that all m "valid" signals except the one that caused $\mathbf{S}_R$ (namely, $\mathbf{S}_T$) are outside the lens volume satisfies the inequality:

$$\text{Probability of no confusion} > \left[1 - \left(\frac{N}{S+N}\right)^{BT}\right]^{m-1} > 1 - (m-1) \cdot \left(\frac{N}{S+N}\right)^{BT}$$

The probability of an error in detection is:

$$P_e = (1 - \text{Probability of no confusion}) = (m-1) \cdot [N/(S+N)]^{BT} < m \cdot [N/(S+N)]^{BT}$$

If we choose $m = [k \cdot (1+S/N)]^{BT}$, then $P_e < [k]^{BT}$. This means that P_e can be made arbitrarily small by increasing T, provided that m is chosen arbitrarily close to $(1+S/N)^{BT}$. This means,

$$C = \frac{1}{T} \log_2 M = \left[B \log_2 \left(1 + \frac{S}{N}\right) - \varepsilon \right] \text{ bits/second} \qquad (2.8)$$

where ε is a positive number chosen as small as desired (depending on how low error rate you want to achieve). Since the average error probability with these randomly chosen signals is less than ε, there must be some choices of the m valid signals (that is, some coding procedures) that give a probability of error less than ε. The following example illustrates the "right" type of redundancy to produce signal with immunity against noise.

Example 2.2 Channel Encoding for Error Detection and Correction

Figure 2-14 shows examples of one-, two- and three-dimensional messages. Noise distorts the transmitted sequence into the received one and, if they are different, the noise caused error. Increasing the signal power causes further separation of the signal points and reduces the probability that the

distortion will result in an error. Redundant encoding excludes some of the *possible* signal points from a selected set of *valid* points, which is another way of increasing the distance between the points. Let us assume two-dimensional messages and we select (−1, 1) as a valid message. Then the Euclidean distances from other candidate messages are $d_E[(-1, 1), (1, 1)] = 2$, $d_E[(-1, 1), (-1, 1)] = 2\sqrt{2}$, and $d_E[(-1, 1), (-1, -1)] = 2$. The best candidate is (−1, 1), since it requires the largest noise power to confuse these two messages. The messages (1, 1) and (−1, −1) are left unused. If either of them is received at the receiver, the receiver knows it is an error. Thus, by introducing redundancy, errors are encoded as invalid messages. As a result, the receiver is able to ***detect*** an error. What to do with it? A natural approach is to mutate it to the closest valid message. Suppose the received message is (1, 1). As shown in Figure 2-14(a), it is equidistant from both valid messages, so the receiver does not know whether the source message was (−1, 1) or (1, −1), i.e., it is unable to ***correct*** the error.

The effect of introducing redundancy in this manner is a reduction in the probability of an error because the noise energy is spread over more bits. The higher the dimensionality of a message, the greater the error probability reduction is possible. For example, in the case of three-dimensional messages, the distance of the antipodal points is $2\sqrt{3}$, and generally, in the n-dimensional case, the distance of antipodal points is $\sqrt{n}$. Figure 2-11(c) is redrawn as Figure 2-14(b), but with a different set of hyperspheres. Let us assume that (1, −1, −1) and (−1, 1, 1) are the valid messages. Suppose (1, −1, −1) is transmitted and affected by noise so the receiver receives (1, 1, −1). In this case, the receiver detects an error, but it can also *correct* it by mutating the received message to the closest valid message, which is (1, −1, −1). If the noise was of a greater magnitude, the received message could be (−1, 1, −1), in which case the receiver would mistakenly mutate it to (−1, 1, 1), which is the closest valid message.

However, interestingly enough, it takes a noise of magnitude 1 or higher to mutate a valid message to an invalid one with only one wrong symbol. It takes a noise of magnitude $\sqrt{2}$ or higher to mutate a valid message to an invalid one with two wrong symbols. The volume of the smallest noise sphere in Figure 2-14(b) is $V_1 = 4/3 \times \pi(\frac{1}{2}r_1)^3 = 4\pi/3$ and for the medium one it is $V_2 = 4/3 \times \pi(\frac{1}{2}r_2)^3 = (8\pi\sqrt{2})/3$ and the ratio of the volumes is $V_1/V_2 = 0.35$. This means that 65 % of the points in the larger sphere are unreachable by the noise of small- to moderate magnitude. Only the strong noise would cause a disastrous error, and for white Gaussian noise very few noise magnitudes are very large. Thus, the probability of an error in two symbols is small. With increasing n, the error probability diminishes, but no matter how large we make n, the error probability never becomes zero. Of course, with longer messages we have more choices for valid messages, so we do not need to use only the antipodal points as valid messages. The selection of messages that are most resistant to noise is studied in the theory of error-correcting codes. Some of this is mentioned in Chapter 1 and I will briefly review some of the powerful error-control encoding schemes below.

The key observation is, instead of dealing with individual symbols, we should deal with sequences of symbols, i.e., messages. The received message is declared error-free if it is one of the predefined "valid" messages and incorrect otherwise. By considering longer symbol sequences, we give the noise a chance to cancel itself out over many symbols. Having the noise energy spread over a higher-dimensional space lowers the noise impact.

2.4 Radio Propagation: Multipath and Doppler

Thermal or AWGN noise is present in every communication channel. Techniques presented above deal with AWGN noise and the channel capacity formula (2.5) applies only to AWGN channels. The radio propagation channel exhibits many different forms of channel impairments, which result in different types of noise, in addition to AWGN. Unlike a wire link, wireless link is very unstable and changes dramatically even if the stations are not moving and have direct line-of-sight (LOS) communication. This section aims at obtaining the characteristics of noise and random signals in radio channels from knowledge of the physical processes generating them.

The *air link* is the radio or infrared link between the transmitters and the receivers; therefore "link" is a relative (or "soft") concept. Radio signals are sent across the air by inducing a current of sufficient amplitude in an antenna whose dimensions are approximately the same as (but no less than 1/10 of) the wavelength of the generated signal. (The signal wavelength $\lambda_c = c/f_c$, where f_c is the signal's carrier frequency and $c \approx 3\times10^8$ m/s is the speed of light in the air.)

In real life transmitter and receiver are almost never located in an empty space—normally they are surrounded by objects in the environment and even the air is filled with molecules. As EM waves propagate from transmitter to receiver, they interact with these objects and get modified in the process. Waves also interact with waves from other sources or from the same source and with ambient noise, and again get modified in the process. In addition, the movement of the transmitter and receiver relative to each other impacts the transmitted EM waves. The most important aspects of these interactions from the communication viewpoint are transmission impairments that result from *multipath transmissions*, *Doppler effect*, and *interference of multiple sources*. The ultimate effect is loss of transmitted information, which can be interpreted as decrease in channel capacity due to increased noise.

2.4.1 Large-Scale Path Loss

Electromagnetic waves emanating from a point source experience a reduction in amplitude as they travel away from their source. The reason for the loss is that signals spread over an increasingly larger area (it is not that they get tired traveling the distance from the transmitter!). Each advancing wave crest is an expanding circle. Since the wave energy contained in each crest is a fixed amount and the crest is expanding, the energy per unit length of crest must decrease. Therefore, the amplitude of the wave must diminish. [Actually, since it is an expanding sphere for spatial sources, the effect is even greater.] For example, the apparent power of a signal from a point source will decrease by the square of the distance simply because it is spread over a circle centered at the broadcast source, as shown in Figure 2-15(a). This is referred to as *large-scale path loss*, which is the ratio of received power to the transmitted power for a given propagation path.

As a receiver moves farther away from the transmitter, propagation loss will increase and cause the signal-to-noise ratio at the receiver to decrease to the point where the radio is unable to distinguish the data signal from the noise. When this occurs, the receiver is operating in a fringe area, the maximum distance from the transmitter.

Free space is the simplest propagation model. If P_R is the received signal power and P_T is the transmitted power, then in free-space propagation:

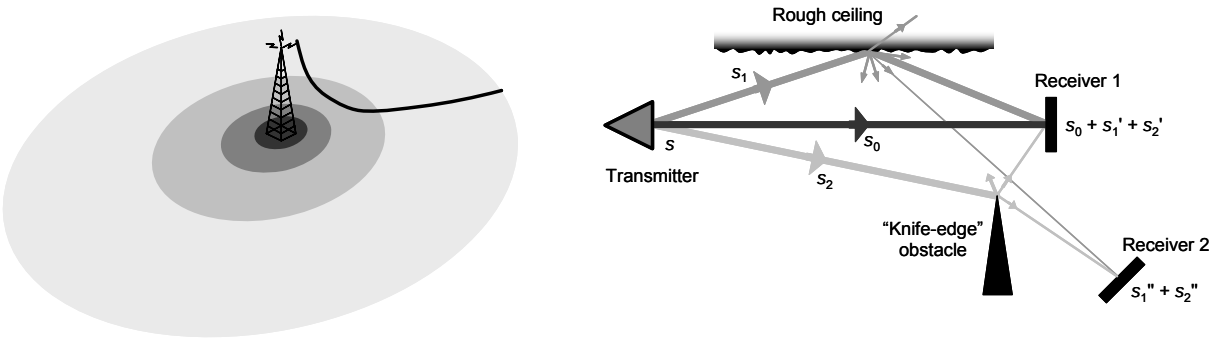


Figure 2-15. Wave propagation and interactions. (a) Propagation loss or large-scale path loss. (b) Multipath waves caused by reflection (s_1'), scattering (s_1''), and diffraction (s_2', s_2'').

$$P_R(d) \propto \frac{G \cdot P_T}{f_c^2 \cdot d^{\alpha}} \tag{2.9}$$

where d is the propagation distance, f_c is the carrier frequency, G is the power gain from the transmit and receive antennas, and α is the propagation loss exponent. It is known in physics that radiated power confirms to an inverse square relationship, corresponding to α=2. Although this describes radio waves propagation in free space, there is usually additional attenuation due to environment factors. Attenuation is caused by obstacles such as foliage, buildings and rain. Table 2-2 gives typical values for the propagation loss exponent in different propagation environments.

Table 2-2: Propagation loss exponent for different environments.

Environment	Propagation loss exponent α
Free space	2
Urban cellular radio	2.7 to 3.5
Shadowed urban cellular radio	3 to 5
In building with line of sight	1.6 to 1.8
Obstructed in building	4 to 6

A problem with propagation loss is that it becomes much more significant at higher frequency, see equation (2.9). For example, increasing the signal frequency by a factor of 10 reduces the received power by a factor of 100. What this means in practice is that the networks operating at a higher frequency require larger number of "cells" to cover the same area. E.g., IEEE 802.11a wireless LAN, which operates at 5 GHz, requires much greater number of access points than an IEEE 802.11b 2.4 GHz system to cover the same area.

When using frequencies above 10 GHz rain attenuation affects radio propagation, depending upon the rain rate, usually expressed in mm/h. For carrier frequencies in the 63-64 GHz band the oxygen absorption introduces a further attenuation, whose specific value α_{OXY} is about 11 dB/km for f_c=63 GHz.

Antennas provide signal gain, which increases the possible distance of a link. Most of the mobile wireless stations use omnidirectional antennas, but directional antennas can improve the distance covered and the signal quality.

2.4.2 Types of Wave Interactions

Waves interact with the environment and other waves through the phenomena of *reflection*, *refraction*, *diffraction*, *scattering*, and *interference*. Some of these are shown in Figure 2-15(b).

Waves are turned back, or **reflected**, when they encounter an abrupt change in the nature of the medium in which they are traveling. Reflection occurs when a propagating EM wave impinges upon an object which has much larger dimensions than the wavelength of the propagating wave, such as surface of the earth, buildings, and walls. Reflection may be partial or complete, depending on the severity of the change at the reflecting boundary. If there is no boundary mechanism for extracting energy from the wave, the entire energy incident with the wave is reflected back with it. If the second medium is a perfect dielectric, part of the energy is transmitted into the second medium and part of the energy is reflected back into the first medium, and there is no (loss of energy to) absorption. If the second medium is a perfect conductor, then the wave is reflected back into the first medium without loss of energy.

Refraction is the bending of the path of wave disturbance as it passes obliquely from one medium into another of different speed of propagation. This is the case with dielectric materials, such as glass or cardboard. The bending of a wave around the edge of a barrier is called **diffraction**. Diffraction occurs when the propagation path en route to the receiver is obstructed by a surface that has sharp irregularities (edges or corners).

As the mobile station moves in uneven terrain, it often travels into propagation shadow behind a building or a hill or other obstacle. The secondary wavelets resulting from diffraction from the obstructing surface are present throughout the space and even behind the obstacle. In a way, the "shadow" cast by the obstacle is fuzzy, so reception is possible even in the shadow. The longer the radio waves the fuzzier shadows they cast, and shorter radio waves cast sharper shadows, almost like visible-light waves. The associated signal level received inside the shadow is attenuated significantly and the inverse square law of propagation loss does not hold. Random signal variations due to these obstructing objects are called *shadow fading* or *shadowing*.

Scattering occurs when a wave impinges upon an object with dimensions on the order of its wavelength or less, causing the reflected energy to spread out or "scatter" in many directions. An incoming wave is scattered into several outgoing waves of smaller amplitudes. Scattered waves are produced by rough surfaces, small objects (chairs and metal bins), foliage, rain, dust, smoke, and other irregularities in the channel.

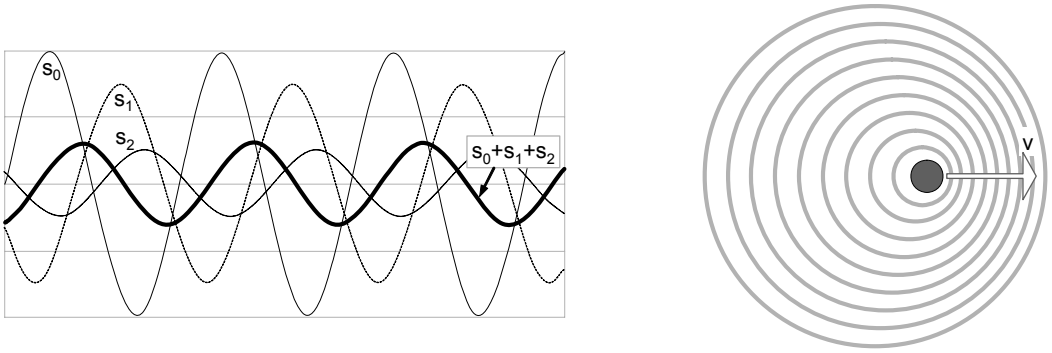


Figure 2-16: (a) Interference of three sinusoidal waves of the same frequency and different amplitudes and phases. (b) Doppler effect produced by a point source moving from left to right with velocity *v*.

It is possible that two or more wave disturbances move through a medium at the same time. The multiple waves could originate from the same or different sources. The interactions defined above (reflection, diffraction, and scattering) cause multiple copies of the same wave to arrive at the receiver, as illustrated in Figure 2-15(b). The principle of superposition states that when multiple waves meet in the same medium, the instantaneous displacement of the medium is given by the algebraic sum of the instantaneous displacements of the individual waves. The effect of superposing multiple wave trains is called **interference**. An example is shown in Figure 2-16(a), where a source emits a sinusoidal signal, which arrives as a line-of-sight (LOS) and two non-line-of-sight (NLOS) copies at the receiver. Since different copies usually traverse paths of different lengths, they will arrive slightly out of step with each other. In Figure 2-16(a), the waves are almost opposite in phase, meaning that the crests and troughs of the waves have reversed their positions, and they *interfere destructively*. Had the waves arrived in phase, i.e., they fall neatly on top of each other, they would *interfere constructively*.

Equation (2.9) represents the average signal strength, and the power measured at a specific location is likely to be above or below this average. Based on propagation loss alone, the received signal power at a fixed distance from the transmitter should be constant. However, wave interactions with the environment and other waves cause the received signal power at equal distances to be different, since different locations exhibit different wave interactions. Figure 2-17 shows ray-tracing simulation of signal intensity in a closed office environment, a room with a doorway and a metal desk. It can be seen that the path loss has alternate minima and maxima as the distance from the transmitter increases, and in general, the path loss increases with distance.

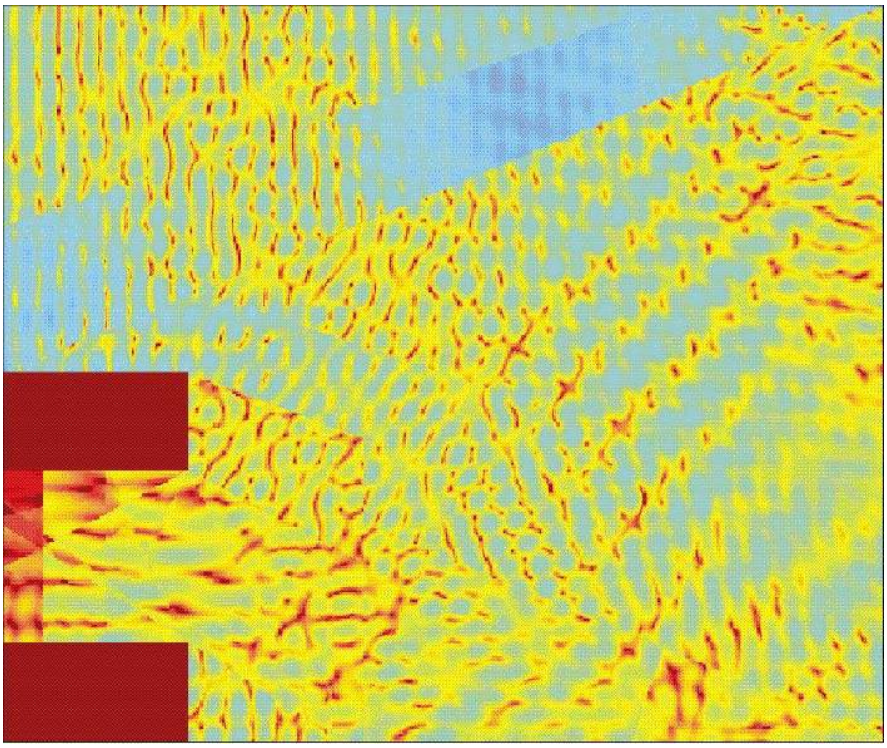


Figure 2-17: Signal intensity simulation [from IEEE 802.11 standard]. Darker (red) points symbolize a stronger signal and lighter (blue) points symbolize a weaker one.

2.4.3 Doppler Shift

If the source and receiver move relatively to each other, the receiver experiences a *Doppler shift* for the arriving waves, as illustrated in Figure 2-16(b). If the stations are moving towards each other, the receiver experiences an increase in the apparent wavelength of the signal; conversely, if the stations are moving away from each other, the receiver experiences a decrease in the apparent wavelength of the signal. Waves arriving at right angles to the direction of motion experience no shift in wavelength. Of course, the wavelength of the signal emitted by the source remains unchanged, as does the velocity of the wave in the transmitting medium. Suppose the velocity of the relative motion is v and the motion is towards each other. Then wavelength of the wave reaching the receiver is:

$$\lambda_c' = (c - v \cdot \cos\theta) / f_c$$

where c is the speed of light, θ is the arrival angle between the line connecting the transmitter and receiver and the direction of motion, and f_c is the carrier frequency. The frequency of the wave reaching the receiver is $f_c' = (v \cdot \cos\theta) / \lambda_c'$. (In case of the motion away from each other, the formula for the wavelength is modified so that the motion velocity is added to the speed of light.)

The *Doppler spread*, f_m, is the maximum Doppler shift, $f_m = v/\lambda_c$.

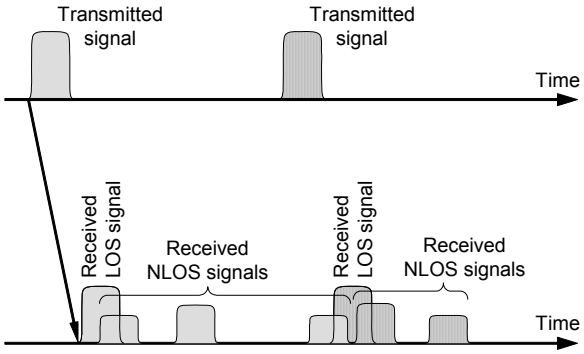


Figure 2-18: Illustration of multipath fading, which produces a single distorted copy or multiple copies of each transmitted signal.

2.4.4 Multipath or Small-Scale Fading

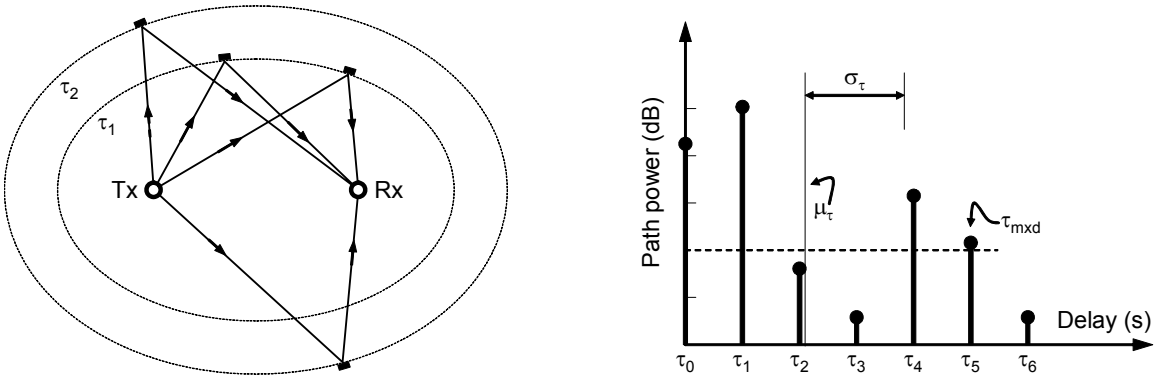
Most antennas are omnidirectional and even directional antennas emit signal over a spatial angle. Therefore, when a communication signal is transmitted to a receiver, the wave interactions described above will cause that signal to arrive to the receiver along several different paths. For example, in Figure 2-15(b) signal s_0 is LOS wave, and s_1, s_2 are NLOS waves. The resulting unpredictable set of reflections and/or direct waves is called *multipath*. Multipath effects are a major problem for wireless communications. They are a major factor limiting the capacity of wireless channels.

On one hand, this is beneficial, since it is possible to receive reflected signals in places a direct LOS signal cannot reach. For example, reflections can allow reception in between office buildings or under overpasses that completely block a direct signal (Receiver 2 in Figure 2-15(b)).

However, the problem with multipath signals is that the length of the path can greatly vary, changing the received signal's timing, strength, and quality. Figure 2-18 illustrates the effects of multipath fading on a baseband signal. (Baseband signal is used just for illustration purposes; in reality digital information is modulated on an analog carrier, see Figure 2-5.) Multipath causes three major effects:

- Rapid changes in signal strength over a short distance or time
- Random frequency modulation due to Doppler shifts on different multipath signals
- Time dispersion caused by multipath delays

Different parts of the signal are delayed differently, causing a spread in arrival times at the receiver, which is equivalent to the variation in phase. This phase variation is characteristic of a channel under consideration. If transmitter, receiver, or reflecting objects are moving relative to each other, then the phase variation is also *time varying*. Modeling such a channel is important to determine the expected channel error rates so proper countermeasures can be undertaken. If we were to produce an exact channel model for a particular environment, we would need to know the attributes of every reflector in the environment at each moment in time. Multipath channels are generally very difficult or impossible to model exactly and it is usual to develop a *statistical*



(a) (b)

Figure 2-19. (a) The signals from secondary sources located on the same ellipse have the same excess propagation delay. (b) Path delay profile of a specific time-invariant wireless channel. Indicated are the channel parameters: mean excess delay (μ_τ), rms delay spread (σ_τ), and maximum excess delay (τ_{mxd}).

model, which emulates the typical or average behavior of the channel. Here I review some of the channel features that decide on the type of the statistical model to be employed.

The first feature is the ***power delay profile*** of a channel. The additional delay of a multipath (NLOS) signal relative to the LOS signal is proportional to the length of its propagation path. The physical structures, which in effect are secondary sources that cause multipath signals, can be grouped to equidistant groups as in Figure 2-19(a). These groups form confocal ellipses with the transmitter (Tx) and receiver (Rx) as the foci. All objects located on the same ellipse are associated with the same excess total path length and cause the same additional delay.

A useful concept here is that of the Fresnel zone. Fresnel zones represent successive regions where secondary signals have a path length from the transmitter to receiver which exceeds by $n\lambda_c/2$ the length of the LOS path, where λ_c is the wavelength of the carrier and $n = 1, 2, 3, \ldots$. The sinusoidal waves arriving along the paths in successive Fresnel zones alternately have opposite or equivalent phase with the LOS wave. Thus, they result in alternating constructive and destructive interference to the total received signal. The multipath signals from secondary sources in between of these provide correspondingly smaller interference effects. The secondary sources within the first Fresnel zone ($n = 1$), have the most severe destructive effects since they produce the least attenuated data ghosts. As an example, for 2.4 GHz IEEE 802.11b and 802.11g, $\lambda_c/2 = ½\ v/f_c = ½\ (3\times10^8/2.4\times10^9) = 6.25$ cm; similarly, for 5 GHz IEEE 802.11a, $\lambda_c/2 = 3$ cm.

Channel sounding is used to determine the delay profile of a channel. Conceptually simplest channel sounding technique is to send a very short spike (impulse) and measure the received signal strength. This produces *impulse response* of a multipath channel; an example is shown in Figure 2-19(b). Because the system (channel) is considered to be linear, the impulse response of the channel is enough to characterize it. The first detectable arriving spike has delay τ_0 (if there is a line of sight between the transmitter and receiver then this is a LOS signal). All other delays are measured relative to τ_0, which is set to zero, and are spaced equally, $\tau_{i+1} - \tau_i = \Delta\tau$. The parameters that grossly quantify a multipath channel are: mean excess delay (μ_τ), rms delay spread (σ_τ), and maximum excess delay (τ_{mxd}), which is defined as the last delay path with a "noticeable" amplitude (typically 20 dB below the peak amplitude).

The second feature is the ***Doppler spread*** of a channel, which is relevant only for time-varying channels, i.e., if communicating entities and/or reflectors are in relative motion.

Multipath causes two significant channel impairments:

- Flat fading
- Intersymbol interference

In addition, moving stations also experience:

- Fast fading
- Slow fading

Flat fading describes the rapid fluctuations of the received signal power over short time periods or over short distances. Such fading is caused by the interference between different multipath signal components that arrive at the receiver at different times and hence are subject to constructive and destructive interference. At some locations, the rays reinforce each other and the received signal is relatively strong. At other locations, the rays cancel each other out and the received signal is weak. This constructive and destructive interference generates a standing wave pattern of the received signal power relative to distance or, for a moving receiver, relative to time.

The real difficulty is when the delay causes previous symbols to interfere with the one currently being processed. For example, in Figure 2-18 the last NLOS signal of the first symbol interferes with the LOS signal of the second symbol. These secondary symbols or data ghosts cause what is referred to as *inter-symbol interference* (ISI). If the data rate is increased, the symbols come closer together and ISI becomes even more of a problem.

Fading is a random linear time-varying transformation of the transmission signal. As a result, the receiver may not be able to detect the data correctly, especially with higher transmission rates. The higher transmission rate means that the bandwidth of the signal will be large compared with the coherence bandwidth of the propagation channel. When this is the case, different frequency components of the signal will experience different fading characteristics, i.e., different parts of the signal spectrum will be affected differently by channel fading. Figure 2-18 illustrates effects of time-varying flat-fading multipath channel. "RF hostile" areas, such as manufacturing plants, can offer significant multi-path distortion, requiring, for example, 11 Mbps IEEE 802.11b wireless LANs to operate at the lower, 1 or 2 Mbps.

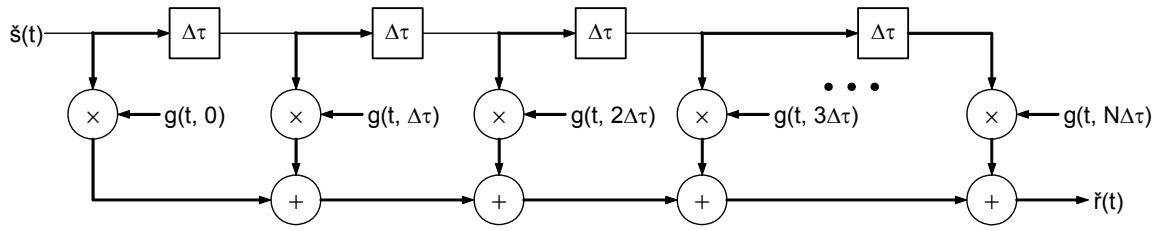


Figure 2-20: Discrete-time tapped delay line model for a multipath-fading channel.

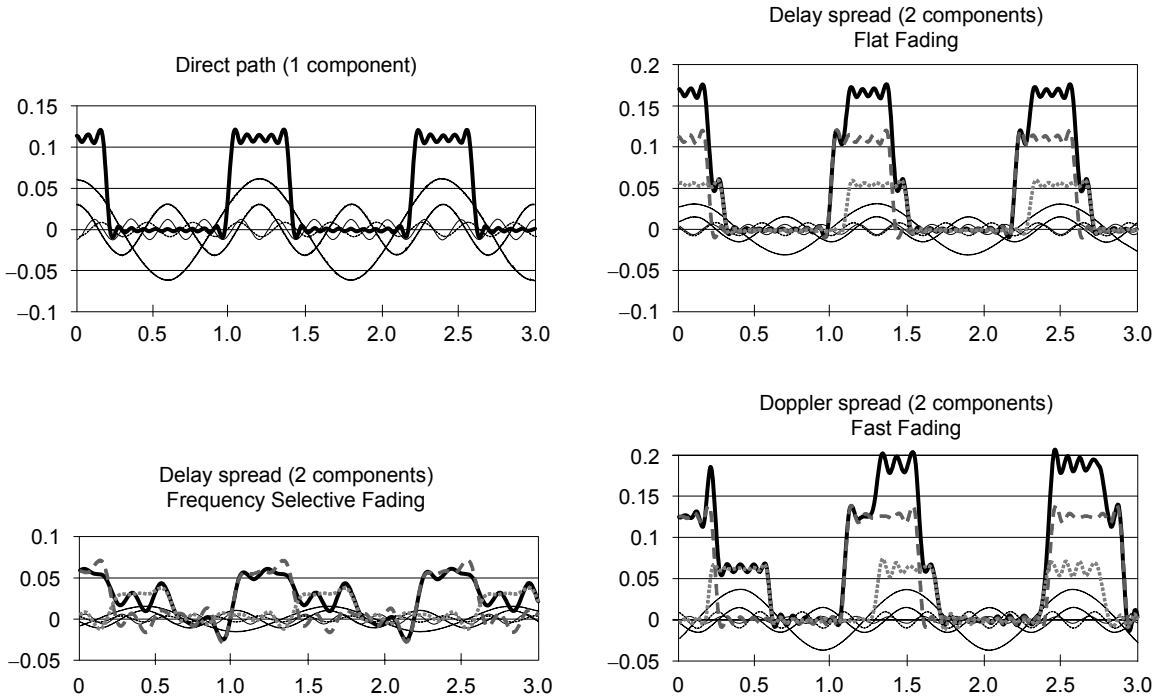


Figure 2-21: Types of small-scale fading.

The different Doppler shifts in the many rays arriving at the receiver cause the rays to arrive with different relative phase shifts that depend on the exact position of the station. Again, different multipath signal components are subject to constructive and destructive interference. The phenomenon is referred to as *Rayleigh fading* or *fast fading*.

Coherence time, T_C, is the time domain dual of Doppler spread. Coherence time is the time duration over which two received signals have a strong potential for amplitude correlation. Two signals arriving with a time separation greater than T_C are affected differently by the channel.

Equation is an approximation using a correlation of 0.5. If the symbol rate is greater than $1/T_C$, the channel will not cause distortion due to motion.

Coherence bandwidth is a statistical measure of the range of frequencies over which the channel can be considered constant, or 'flat'. The bandwidth over which two frequencies have a strong potential for amplitude correlation. Coherence bandwidth is best measured, but it can be approximated as:

2.5 Equalization, Coding, and Diversity

Wired communications typically force a retransmit of information based on the corruption of as little as a single bit of the transmitted information. In wireless systems, bit error rates are several

magnitudes higher than in a typical wired system. Applying this approach to wireless communications would result in overwhelming numbers of retransmissions. Something different has to be done.

The philosophy of wireless communications has been to treat the communication channel as an unreliable transmission medium. The emphasis is on *error prevention*, rather than simply detection and repeated transmission. Different methods of error prevention are the topic of communication engineering, and here I provide a brief overview.

Wireless communication systems deploy a large collection of signal processing techniques to overcome inevitable transmission impairments. Adaptive equalization, channel coding, and diversity techniques are three techniques which can be used independently or in tandem to compensate for the errors and distortions introduced by multipath fading.

2.5.1 Adaptive Equalization

Adaptive equalization is used to combat intersymbol interference. The process of equalization involves some method of gathering the dispersed symbol energy back together into its original time interval.

Figure 2-?? illustrates a basic approach using a linear equalizer circuit. In this example, for each output symbol, the input symbol is sampled at five uniformly spaced intervals of time, separated by delay τ. These samples are individually weighted by the coefficients C_i and then summed to produce the output. The circuit is referred to as adaptive because the coefficients are dynamically adjusted. Typically, the coefficients are set using a training sequence, which is a known sequence of bits. The training sequence is transmitted. The receiver compares the received training sequence with the expected training sequence and on the basis of the comparison calculates suitable values for the coefficients. Periodically, a new training sequence is sent to account for changes in the transmission environment.

For Rayleigh channels, or worse, it may be necessary to include a new training sequence with every single block of data. This represents considerable overhead but is justified by the error rates encountered in a mobile wireless environment.

2.5.2 Channel Coding or Forward Error Correction

The emphasis is on *error prevention*, rather than simply detection. One of the ways to do this is to encode the information sent over a wireless link. An 8-bit number will typically be encoded into a 12- or 16-bit number. The trick is to spread the 255 values of the data byte as widely as possible over the larger range of values. The receiver then decodes the received value into the equivalent data byte that is closest to the value. An example of such an encoding scheme is shown in Table ??. Any single bit in the encoded values can change without causing it to decode into a wrong data value. The number of bits that can change without this happening is called the Hamming distance between the encoded values. The trick is to balance factors like Hamming distance, required data throughput, and the complexity of the decoding algorithm against each other to achieve acceptable throughput and reliability of communications.

A basic approach to combat frequency selective fading is to partition the signal into contiguous frequency bands, each of which is narrow compared with the coherence bandwidth of the channel. Frequency hopping spread spectrum (FHSS) offers a good resistance to multipath fading because the direct signal always arrives at the receiver first. Reflected signals follow a longer path and arrive later. By then the receiver may have changed frequency and no longer accepts signals on the previous frequency, thus eliminating interference between the direct and reflected signals.

2.5.3 Diversity Techniques for Fading Channels

Fading channels are not memoryless. Thus, sequence estimation is required for optimum detection.

In a Rayleigh-faded channel, the BER performance predictions use the equation:

where E(a2) is the Rayleigh-fading term. This Rayleigh term predicts performance degradations similar to the (1–r) cross-correlation term. Therefore, nonorthogonal systems like the 4-FSK example above may have static BER performance curves similar to the faded-channel performance of an ideal orthogonal system.

TO COVER:

Receiver structures for optimum detection on wireless channels (including effects of MAI (multiple access interference) and ISI (intersymbol interference)), bandwidth efficient coding, link quality and channel estimation techniques for wireless communications

2.5.4 Packet Error Rate

The relationship between packet error rate (PER) and BER depends on the channel coding scheme. Assume that there is no error-correction coding applied and the number of bits in a packet is *n*, then

$$\mathrm{PER}(t) = 1 - \prod_{i=1}^{n} \left[1 - \mathrm{BER}(t_i)\right] \tag{2.10}$$

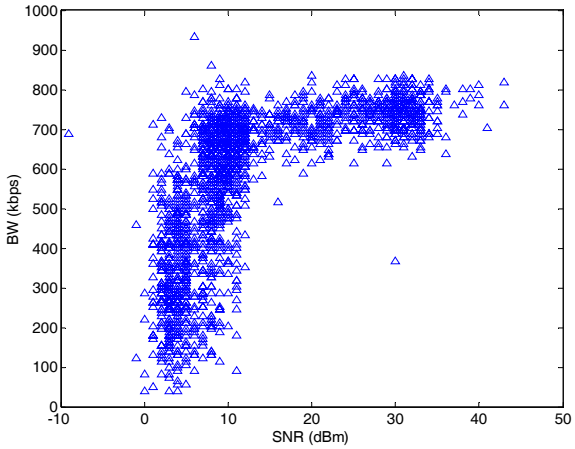


Figure 2-22: SNR-capacity relationship for a typical dataset clearly shows almost abrupt drop in capacity.

where t_i is the time instant the receiver received the i^{th} bit. If the wireless channel fading variation is very slow compared to the packet transmission time (true for walking and driving speeds), then PER can be approximated as $\text{PER}(t) = 1 - [1-\text{BER}(t)]^n$. Thus PER($t$) gives the probability of packet loss at time t on the channel and can be viewed as $\text{PER}(t) = P[\text{BER}(t)]$, where $P[\cdot]$ is a nonlinear operator.

Eq. (2.10) assumes that bit errors are independent events, which is known to not be the case, see e.g., [Berger & Mandelbrot 1963; Mandelbrot & van Ness 1968; Kopke *et al*. 2003].

2.6 Summary and Bibliographical Notes

This is merely a brief digest of the vast field of wireless digital communications. Spectrum management information is available from [18,50]. History of radio and wireless communications is available from [5,45].

[Shannon & Weaver 1949] is the original text that introduced information theory. Detailed treatment of the geometrical approach to digital communication system analysis is in [Wozencraft & Jacobs 1965/1990]—a classic text on communication systems. Noise is present in all communications, but the probability of noise-caused errors is much higher and varies dynamically for wireless channels. Complex coding and signal processing techniques are employed to prevent transmission errors. An in-depth text on signal processing techniques for digital communications is [Proakis 2001]. A somewhat older, but perhaps more accessible text is

[Lathi 1989]. An inquiring reader is well advised to browse older textbooks to get feeling of how the field evolved since Shannon's original work.

Wireless channel models and efficient encoding are covered in [Bertoni 2000; Rappaport 2002; Stüber 2001]. Sklar [1997(a); 1997(b)] provides a comprehensive tutorial on fading channels and mitigation techniques.

Recent work on the capacity of wireless channels is summarized in [Goldsmith & Chua 1997; Biglieri *et al*. 1998].

Problems

Note: Solutions of selected problems can be found on the back of this text.

Problem 2.1

Consider a telephone line with a bandwidth of 3000 Hz and a signal-to-noise ratio of 20 dB. What is the maximum theoretical transmission rate that can be obtained?

Problem 2.2

A system has an input noise of 20 pW, output signal of 0.4 mW and output noise of 4 nW. If the gain of the system is 40, determine the signal-to-noise ratio at the input.

Problem 2.3

Let $p(t)$ denote the rectangular pulse used in construction of the digital waveform $s_i(t)$ for a block of n binary digits consisting of 0's and 1's. The pulse amplitude is 1 and let T denote the bit duration, i.e., the duration of binary symbol 0 or 1. Write a single formula to represent any block using the basis vectors, such that −1V is used for binary symbol 0 and +1V is used for binary symbol 1. (See illustration in Figure 2-23.)

Problem 2.4

Let $\{\phi_k(t)\}_{k=1}^{K}$ be K orthonormal waveforms over $[a, b]$. Suppose we have a function $s(t)$ defined over $[a, b]$, and that we approximate $s(t)$ by

$$\hat{s}(t) = \sum_{k=1}^{K} a_k \cdot \phi_k(t)$$

Show that the choosing $a_k = \langle s(t), \phi_k(t) \rangle$ minimizes $\varepsilon_e = \|s(t) - \hat{s}(t)\|^2$.

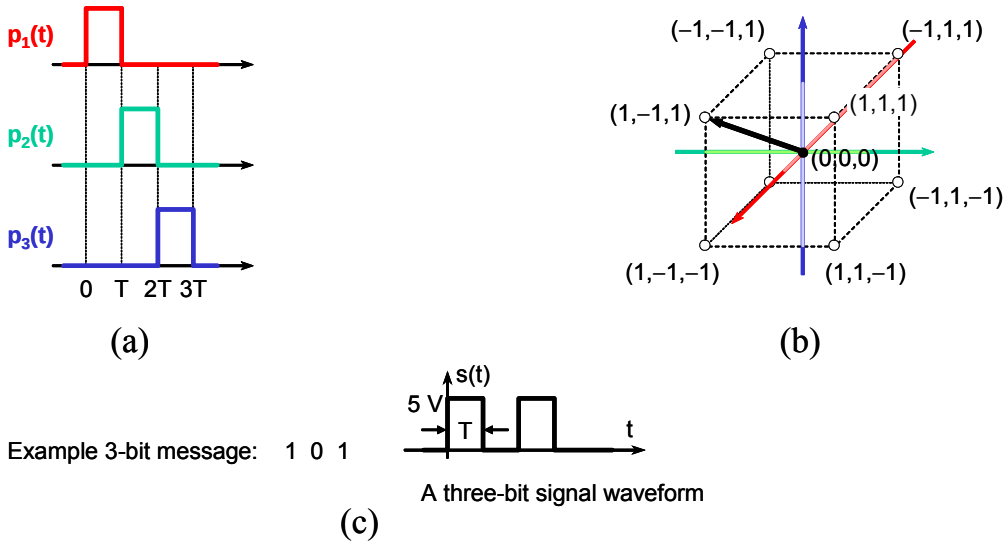


Figure 2-23: (a) Orthogonal function set (basis vectors) for *n* = 3. (b) Vector representations for different *n*-blocks. (c) A 3-bit signal waveform for the example 3-bit message "1 0 1".

Problem 2.5

Derive the relationship among transmission rate D (in bits), symbol rate R, and the number of symbols M in an M-ary system.

Problem 2.6

The diagram in the lower-right corner of Figure 2-5 shows the signal waveform for BPSK encoding of a binary sequence "101101." Construct the waveform of the QPSK signal for the same binary sequence as well as for "11010001."

Problem 2.7

Suppose 16-QAM modulation is employed for a binary sequence "11010001." Provide the phase states and the coordinates of message-symbol points in a tabular form, similar to Example 2.1 above. Draw the waveform of the modulated signal.

Problem 2.8

A baseband digital system is to employ binary signaling over a channel which produces AWGN at the input of the receiver with rms of 100 mV. If BER is to be better than 10^{-5} determine the minimum voltage separation between the voltage levels of each symbol. (*Hint*: see Figure 2-8(b))

Problem 2.9

Suppose a transmission system sends out information at 200 kbps and that the rms noise voltage is 0.2 V. If 1 and 0 digits are equally likely to be transmitted, what is the average time between error occurrences?

Problem 2.10

In the caption of Figure 2-6 we noted that there are other possible 16-QAM constellations in addition to the one shown in Figure 2-6(c). Draw a 16-QAM constellation that uses two amplitudes and eight phases.

Problem 2.11

Derive the relationship among transmission rate D (in bits), symbol rate R, and the number of symbols M in an M-ary system.

Problem 2.12

Chapter 3
Multiaccess Communication

Contents

3.1 Introduction

A transmission medium can be used in a *point-to-point* mode, where the medium directly connects two stations, or it can be used in a *broadcast* mode, where multiple stations share the same medium to communicate with each other. In the latter case, the messages from one station can be heard by every other station in its listening area. A key issue in networks where transmissions are broadcast is how to determine who gets use of the medium when there is competition for it. In literature, broadcast channels are sometimes referred to as *multiaccess* channels or *random access* channels.

When multiple stations use the same medium to transmit packets, a successful transmission will occur only if exactly one station transmits at any given time. If two or more stations transmit, the reception is garbled. If none transmit, the channel is unused or idle. The problem is illustrated in Figure 3-1. Systems in which multiple stations share a common medium in a way that can lead to conflicts are known as *contention* systems. A key issue in multiaccess communication is *coordination* between the competing stations accessing the medium.

The protocols used to determine whose turn is next on a multiaccess channel belong to a sublayer of the link layer in the OSI reference architecture, called the *MAC (Medium Access Control)* sublayer (Figure 1-1). The dilemma is similar to the one faced by a group of people who are about to pass, one by one, through a revolving door. The greatest difference is that the stations do not have the benefit of human senses to help perceive the potential collisions and avoid them

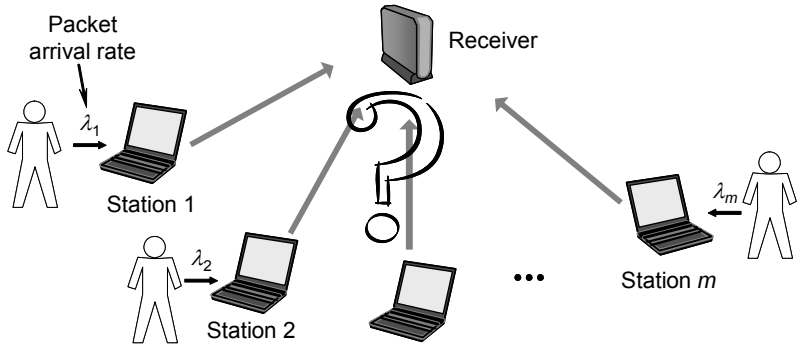


Figure 3-1. Multiaccess contention problem: Who gets to transmit when and how to avoid "collisions" of packet transmissions. Packet arrivals at each station are assumed to be Poisson distributed with rates $\lambda_i = \lambda / m$, $i = 1, \ldots, m$, for an aggregate rate λ.

before they happen. This chapter deals with the techniques employed by the stations to coordinate transmissions.

The MAC layer is particularly important in local area networks (LANs), many of which use broadcast as the basis for communication. Multiaccess communication is commonly used in wired and wireless networks alike, but the problems associated with the wireless link are significantly more complex as will be seen below.

The key issue in designing a MAC protocol is the degree of **control** exercised over the transmitting stations. At one extreme, there is no control at all—stations send new packets immediately as they arrive. In this case, the interesting question is when and how packets are retransmitted when collisions occur. At the other extreme, total control is exercised over packet transmissions—stations receive a schedule designating the reserved intervals for channel use. The interesting questions here are: (*i*) what determines the scheduling order (it could be dynamic), (*ii*) how long can a reserved interval last, and (*iii*) how are stations informed of their turns.

In controlled schemes, channel allocation can be done in a static or dynamic way. *Static schemes* divide the medium into several different portions, whether in time or frequency domain, and assign each portion to a different station. Figure 3-2 illustrates time division multiple access (TDMA) and frequency division multiple access (FDMA). Deterministic schemes avoid contention between the competing stations, but support only a small number of stations and they are suitable if the stations offer heavy traffic load with a "regular" character (i.e., all packets

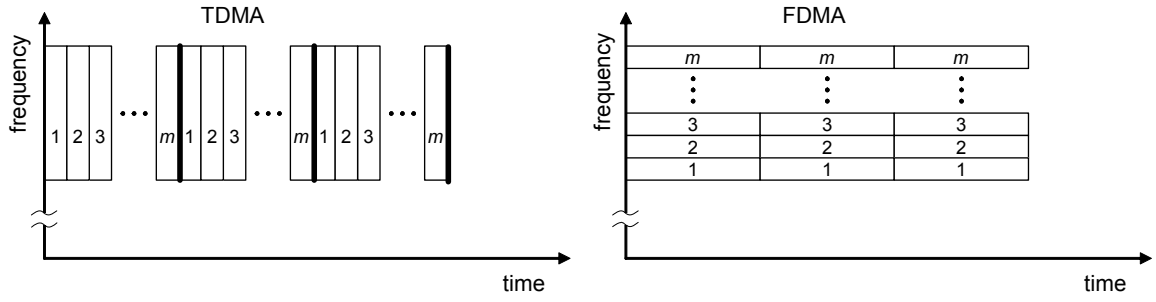


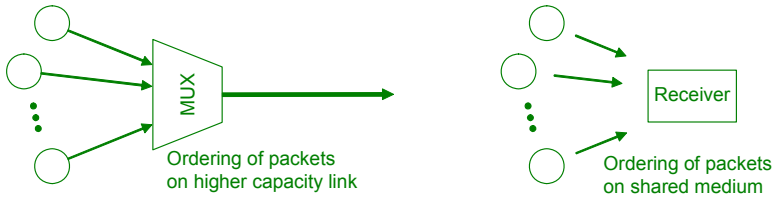
Figure 3-2. Static or deterministic multiaccess schemes. A wideband channel capacity is divided into m channels and each channel is assigned to a different station. In TDMA, a station periodically gets hold of the wideband channel capacity for one time unit, whereas in FDMA the station gets continuous hold of a narrowband subchannel of the capacity $1/m$.

arrive at approximately equally spaced times). An example is voice communication.

In case there is a large number of stations offering a relatively low traffic load, which arrives randomly in bursts, then deterministic schemes are not appropriate and a dynamic channel allocation scheme should be used. Examples are web browsing, file download, and video streaming. Consider an automobile traffic analogy. In the static case (TDMA or FDMA), cars are not allowed to cross over to other lanes even if those lanes may be currently empty. Allowing random access to the channel increases the channel utilization and provides support for larger number of stations, but in this case we need to deal with the contention problem.

MULTIACCESS vs. MULTIPLEXING

♦ Related to multiple access (MA) is *multiplexing* since in both cases multiple sources share a common medium. The difference is that in multiplexing the packets from different sources are scheduled by the multiplexor in serial order, thus assuring that there is no contention between different sources. (The contending packets are simply forced to wait for their turn.) In MA, there is no single physical device that gathers packets before marching them on the shared medium, so the contention problem must be solved.



A simple control solution is to have a *centralized* moderator to poll each station in turn, asking whether it needs to transmit. The problem with this solution is that a station that needs to transmit must wait for the moderator's poll, even if no other station is transmitting. Moreover, if the moderator fails, no one can transmit until the problem is detected and corrected. A more robust, *distributed* solution is for a station to start transmission as soon as another station stops, hoping that no one else is about to do the same. This avoids the time wasted in polling and is robust against failed connections, but it has its own downsides. First, an aspiring transmitter should actively listen to the medium so to know when a previous transmission stops, which requires special-purpose electronics. Second, two stations waiting for a third one to stop transmission are guaranteed to *collide* with each other. If collisions are not carefully resolved, they will continue colliding and no station will make progress. The coordination should be done in an *efficient* manner, so that the number of successful transmissions is maximized. However, notice that increasing control increases the *complexity* of the protocol, thus making it more difficult and expensive to implement and maintain. The MAC schemes reviewed below employ increasing level of control and coordination.

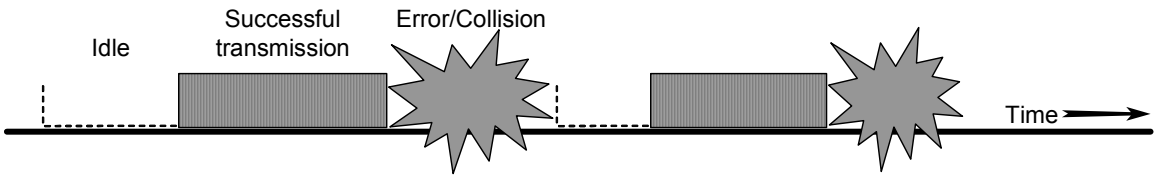


Figure 3-3. Broadcast communication channel can be in one of the three possible states. Assuming that at least one station always needs to transmit, the design objective is to maximize the fraction of the "Successful transmission" state.

3.1.1 MAC Protocol Performance Measures

Given the variety of choices for MAC protocols, it is natural to ask how to choose between them. The most important performance metrics are *capacity*, *delay*, and *fairness*. Our focus in this chapter is on capacity, and although delay and fairness are occasionally mentioned, the proper treatment is postponed to the next chapter. The capacity of a MAC protocol is usually expressed in terms of its *efficiency*—the fraction of transmitted packets that escape collisions (Figure 3-3). In other words, the efficiency identifies the maximum *throughput* rate for a MAC protocol. A first approximation of the throughput can be obtained as follows. Suppose there is always at least one packet ready for transmission and the network channel state probabilities seen at the receiver (p_{idle}, $p_{success}$, and p_{error}) are known. Then, the average total delay per successful packet transmission is:

$$T_{xs} = \frac{p_{idle}}{p_{success}} \cdot \bar{t}_{idle\_contention} + \frac{p_{error}}{p_{success}} \cdot \bar{t}_{error} + \bar{t}_{success} \qquad (3.1)$$

where $\bar{t}_{idle\_contention}$, $\bar{t}_{error}$, and $\bar{t}_{success}$ are the average times the channel spends on idle contention, failed- and successful transmission, respectively. Errors can occur on a multiaccess channel due to channel impairments and due to collisions. In this chapter I first consider an error-free channel and assume that errors happen only due to collisions of multiple simultaneous transmissions, so $p_{error} = p_{collision}$ and $\bar{t}_{error} = \bar{t}_{collision}$. The channel throughput is $\Gamma = 1 / T_{xs}$, and the throughput as seen by an individual station is $\gamma = \Gamma / m$, assuming that all m stations are active.

Eq. (3.1) is only a first approximation since it assumes that the collision probability remains the same, regardless of how many collisions a packet has already suffered. In some cases below, we will be able to derive more accurate expressions for throughput.

Note that if a single station is active on the network, then it transmits without ever colliding with other transmissions. In such a case, the fraction of transmitted packets that escape collisions can be close to 100 percent. The efficiency measure is for a network where many stations compete for transmitting on the common channel. It is under such difficult operating conditions that we need to verify how well a protocol performs. Much of this chapter is dedicated to the derivation of channel state probabilities, p_{idle}, $p_{success}$, and $p_{collision}$, for different MAC protocols.

Note also that efficiency measures only the ratio of the successful transmission to the total number of transmissions. However, it does not specify the delays introduced as a result of reducing the number of collisions.

The following analysis of MAC protocols assumes that the MAC protocol is obliged to resolve all *contention-related* packet errors, i.e., collisions. A packet is kept in the transmitter queue until it is successfully received and acknowledged by the receiver.

COLLISION vs. NOISE

♦ Since wireless stations usually detect collision indirectly, through the lack of acknowledgement, this implies that packet loss due to channel noise cannot be distinguished from packet loss due to collisions—both are assumed to be due to collisions. This has implications on protocol efficiency, since the remedies for collisions are not best suited for channel noise. For example, the exponential increase of the backoff delay does not make sense if the packet was lost due to channel noise.

Looking from the perspective of individual nodes (stations), collision vs. noise makes no difference. However, from the network standpoint there is difference, since collision affects multiple nodes. In effect, channel load increases and the delay would increase for every node.

The following analysis will at first ignore the transmission errors and interference due to near transmissions that have a too weak impact to be considered contending transmissions, but nevertheless impair the channel quality. I will revisit this issue in Section 3.5.

Before analyzing efficiency of contention based protocols, here I state the assumptions about the statistical characteristics of the traffic that is generated by the stations. Different types of traffic yield different efficiencies for a given protocol. It is common to model the sequence of times at which the packets arrive for transmission as a random process. For simplicity, these arrivals are usually modeled as occurring at random points in time, independently of each other and of the arrivals at other stations. This is a Poisson type of arrival process (see Chapter 1).

3.1.2 Propagation Time and Parameter β

The time (in packet transmission units) required for all sources to detect a start of a new transmission or an idle channel after a transmission ends is an important parameter. Intuitively, the parameter β is the number of packets (or a fraction of a single packet) that a transmitting station can place on the medium before the station furthest away receives the first bit of the first packet (Figure 3-4). In the original paper [Kleinrock & Tobagi 1975], this was called parameter a, and both notations are common.

Signal propagation time is $t_{prop} = d/v$, where d is the distance from source to destination and v is the velocity of electromagnetic waves. When electromagnetic waves travel through a medium, the speed of the waves in the medium is $v = c/n$, where n is the index of refraction of the medium and $c \approx 3 \times 10^8$ m/s is the speed of light in vacuum. Both in copper wire and glass fiber or optical fiber $n \approx 3/2$, so $v \approx 2 \times 10^8$ m/s. The index of refraction for dry air is approximately equal to 1. Propagation time is between 3.3 and 5 nanoseconds per meter (ns/m).

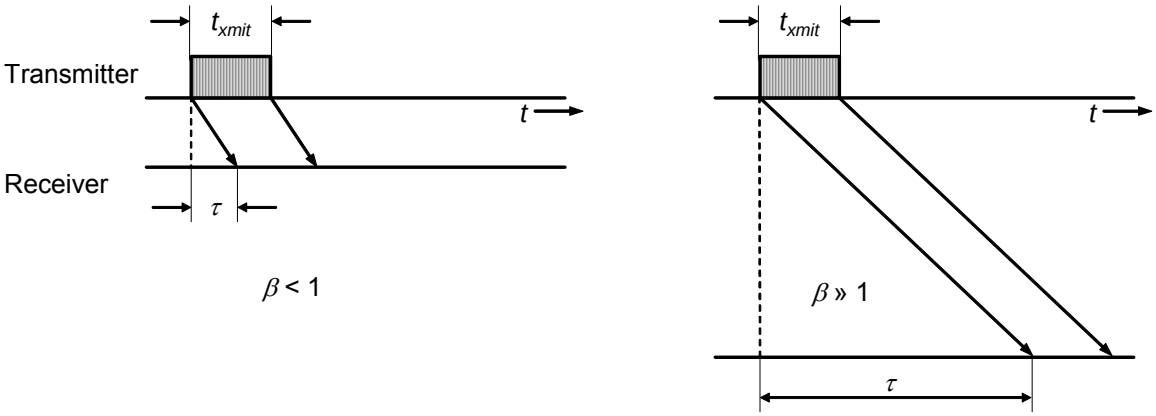


Figure 3-4: Propagation constant β for different ratios of propagation and transmission times. τ combines both propagation and detection delay.

Thus if t_{xmit} is the packet transmission time, then

$$\beta = \frac{t_{prop}}{t_{xmit}} = \frac{t_{prop} \cdot C}{L} \tag{3.2}$$

where C is the raw channel bit rate (channel capacity), and L is the expected number of bits in a data packet. For a 1 Kbytes packet and a transmission speed of C =1 Mbps, the transmission time is t_{xmit} = (8 × 1000 bits) / (1 × 10^6 b/s) = 8 ms. The time taken by the electronics for detection should also be added to the propagation time when computing β, but it is usually ignored as negligible.

Figure 3-4 illustrates different cases for β.

3.1.3 Vulnerable Period

The "vulnerable period" is an important parameter for contention-based MAC protocols. For a given a transmitted packet, the "vulnerable period" is the period of time during which transmission of another packet unavoidably leads to a collision. The length of the vulnerable period determines the protocol performance—the shorter the vulnerable period, the lower the number of collisions, and therefore the better the channel throughput.

Let us assume, for simplicity, that packets are fixed-length. Then, the "packet time" denotes the amount of time needed to transmit a packet, which is the packet length divided by the channel bit rate, $t_{xmit} = L/C$. Consider the most general case of random transmissions. A packet will not suffer collision if no other packets are sent within one packet time of its start, as shown in Figure 3-5. (Notice that for simplicity I ignore the propagation time.) If the packet transmission starts at t_{start}, for any other packet that is transmitted between ($t_{start} - t_{xmit}$) and t_{start}, the tail of that packet will collide with the head of the current packet, shaded in Figure 3-5. Similarly, any other packet started between t_{start} and ($t_{start} + t_{xmit}$) will collide with the tail of the current packet.

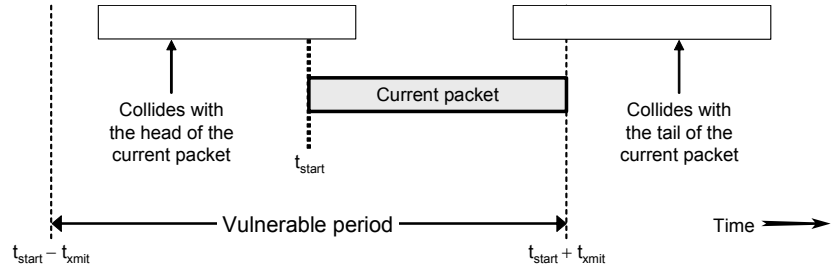


Figure 3-5: Vulnerable period for random packet transmissions.

A common means of improving a protocol's performance is to shorten the vulnerable period by introducing the constraints on packet transmission start time. Some common techniques are:

- Limit the transmission start time to the beginning of discrete time slices, called "slots"
- Listen before talking to avoid having a collision with an ongoing packet transmission

This will become apparent as I introduce different contention-based MAC protocols below.

3.2 ALOHA Protocols

ALOHA is a simple random access protocol: whenever a station has a packet to transmit, it simply transmits without waiting or checking if some other station is already transmitting. This results in occasional collisions and the colliding packets are lost. Assuming that the station can receive feedback about the successfulness of its transmission, the station can then adjust its transmission strategy in the case of collisions.

There are two versions of ALOHA: pure and slotted. They differ with respect to whether or not time is divided into discrete slots into which all packets must fit (Figure 3-6). In the slotted system, all transmitted packets have the same length and each packet requires one time unit (called a slot) for transmission. The benefit of introducing slots is that slots ensure some order in chaotic transmissions. In the slotted version, the stations are allowed to commence transmission only at the beginning of a slot and not at any other time. In effect, this shortens the vulnerability period by half (Figure 3-5). The drawback is that, unlike pure ALOHA, slotted ALOHA requires global time synchronization which is a difficult problem in itself.

Figure 3-7 shows the state diagram for both versions of the protocol. A station in ALOHA sends a packet immediately and in S-ALOHA on a slot boundary, in both cases without checking to see if any other station is already transmitting. After sending the packet, the station waits for an implicit or explicit acknowledgement. The timeout period is set to $2\times t_{prop}$, which is a round-trip time (RTT) for data and acknowledgement packets. If the station receives no acknowledgement, it assumes that the packet got lost to a collision. Each packet involved in collision must be transmitted in some later slot, with further such retransmissions until the packet is successfully received. A station holding a packet that must be retransmitted is said to be *backlogged*. After too many failures, the link is declared down and transmission is aborted.

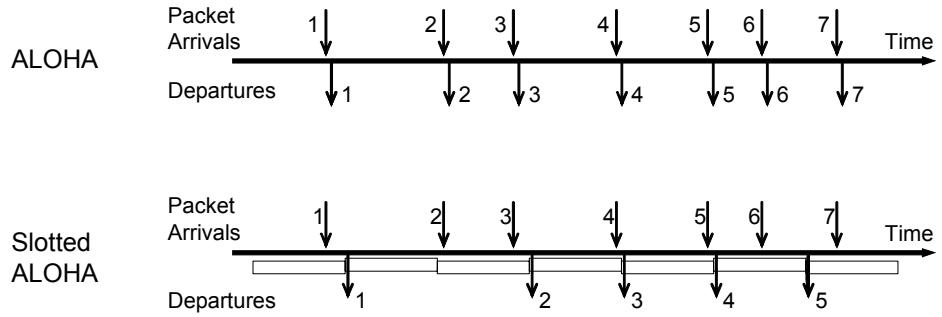


Figure 3-6: Packet arrivals and departures for pure ALOHA and Slotted ALOHA MAC protocols. In Slotted ALOHA, a packet is transmitted only at the beginning of a slot.

When a packet gets lost to a collision, the station waits for a random time and then retries. This is called *random backoff* and it can be selected in a number of ways. A usual way is to select a random number of time units, or slots, uniformly distributed between 1 and CW_{max}, where CW_{max} is an integer number called contention window size.

3.2.1 Throughput Analysis

I first discuss the slotted version for its simplicity. Here are the assumptions about the idealized model with *m* stations (Figure 3-8):

1. *Slotted system*: Assume that all transmitted packets have the same length and each packet requires one time unit (called a *slot*) for transmission. All transmitters are synchronized so that the reception of each packet starts at an integers time and ends before the next integer time.
2. *Poisson arrivals*: The packets arrive for transmission at each of the *m* stations according to independent Poisson processes. Let λ be the overall arrival rate to the system, and let λ/m be the arrival rate at each station, as indicated in Figure 3-8.
3. *Collision or perfect reception*: Any time two or more stations transmit simultaneously, the packets are lost to collision; if exactly one station transmits at any time, the packet is correctly received. (In other words, the packet loss due to erroneous channel is ignored.)

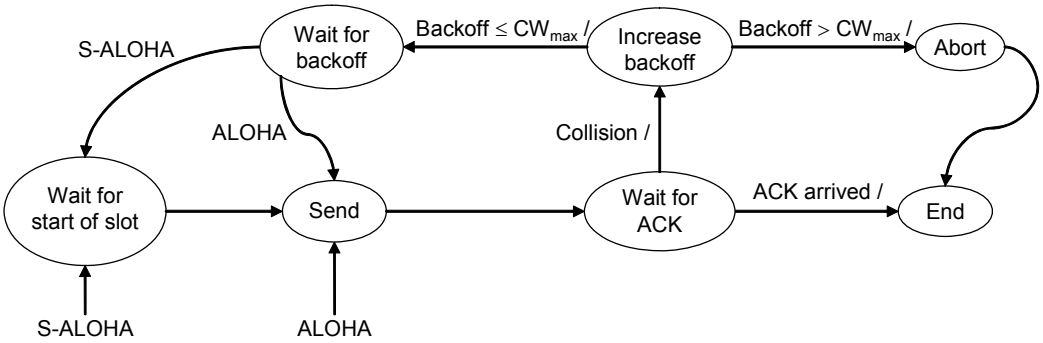


Figure 3-7: State diagram for ALOHA and Slotted ALOHA MAC protocols. The time to wait for acknowledgement is slightly greater than the round-trip time (RTT).

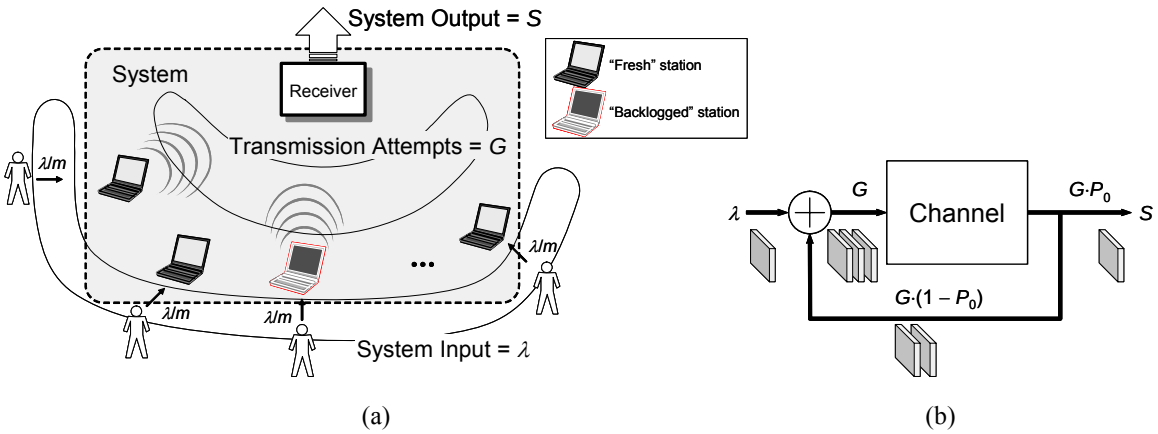


Figure 3-8: (a) ALOHA system representation. (b) Modeled as a feedback system.

4. *Immediate feedback* (0, 1, *e*): Every station obtains feedback immediately at the end of each slot, specifying whether 0 packets, 1 packet, or more than one packet (*e* for error) were transmitted in that slot.
5. *Retransmission of collisions*: Packets involved in collisions become backlogged and are transmitted in a later slot, until successfully received.
6. *No buffering* or *infinite set of stations* ($m = \infty$): In order to ensure that λ does not decrease as the stations become backlogged, we need to assume that: (*a*) the backlogged stations discard the new arrivals until they succeed in retransmission; or, (*b*) the system has an infinite number of stations and each newly arriving packet arrives at a different station.

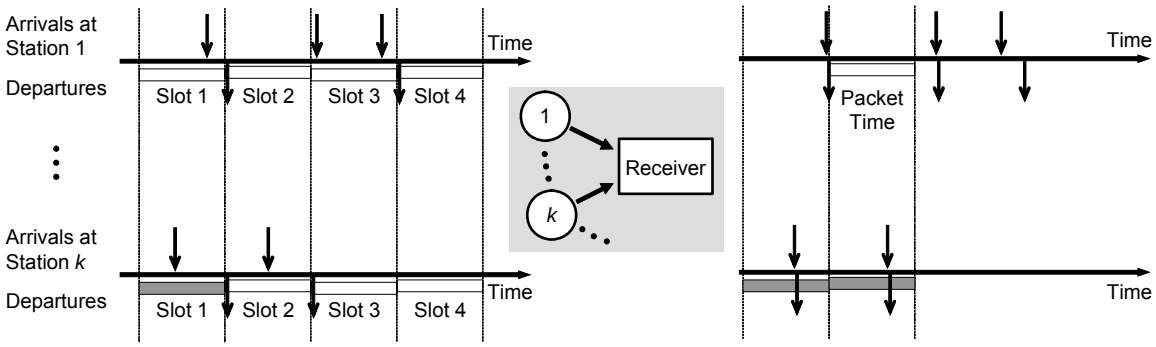
For a reasonable throughput we would expect $0 < \lambda < 1$ since the system can successfully carry at most one packet per slot, i.e., only one station can transmit at a time. Also, for a system to function, the departure rate of packets from the system should be equal to the arrival rate in equilibrium. In equilibrium, the departure rate cannot physically be greater than the arrival rate; if it is smaller than the arrival rate, all the stations will eventually become backlogged.

The following "quick-and-dirty" derivation yields a reasonable approximation, which will be strengthened in the second iteration. In addition to the new packets, the backlogged stations generate retransmissions of the packets that previously suffered collisions. If the retransmissions are sufficiently randomized, it is plausible to approximate the total number of transmission attempts per slot, retransmissions and new transmissions combined, as a Poisson random variable with some parameter $G > \lambda$. With this approximation, the probability of successful transmission (i.e., throughput *S*) is the probability of an arrival times the probability that the packet does not suffer collision; since these are independent events, the joint probability is their product. The probability of an arrival is

$P_a = \tau \cdot G$, where $\tau = 1$ is slot duration and *G* is the total arrival rate on the channel.

The packet will not suffer collision if no (other) packets are sent in the slot $[t, t + 1)$, and from the Poisson distribution formula, Eq. (1.1), $P_0 = P\{A(t + \tau) - A(t) = 0\}$. With $\tau = 1$, we have

$$S = P_a \cdot P_0 = (1 \cdot G) \cdot P\{A(t+1) - A(t) = 0\} = G \cdot e^{-G} \tag{3.3}$$



(a) – Slotted ALOHA (b) – Pure ALOHA

Figure 3-9. A packet transmitted by Station 1 will suffer collision if one or more packets *arrive* at other stations during a vulnerable time period (shaded). The difference between ALOHA and Slotted ALOHA is since the arrived packets are transmitted differently.

This is illustrated in Figure 3-9(a), where we assume that only Station 1 has a packet to transmit in a given slot. Finally, in equilibrium, the arrival rate, λ, to the system should be the same as the departure rate, $G \cdot e^{-G}$. The relationship is illustrated in Figure 3-10.

We see that the maximum possible throughput of $1/e \approx 0.368$ occurs at $G = 1$. This is reasonable, since if $G < 1$, too many idle slots are generated, and if $G > 1$, too many collisions are generated. At $G = 1$, the packet departure rate is one packet per packet time, of which $1/e$ are newly arrived packets and $1-1/e$ are the successfully retransmitted backlogged packets.

Example 3.1 Slotted ALOHA Throughput

Assume a slotted ALOHA system with 10 stations, a channel with transmission rate of 1500 bps, and the slot size of 83.33 ms. What is the maximum throughput achievable per station if packets are arriving according to a Poisson process?

The packet size is $1500 \times 0.08333 = 125$ bits, so the channel can support a maximum of 12 packets per second ($G = 1$ per slot). Of these, some will end up in collision and the effective throughput will be equal to $0.368 \times 12 = 4.416$ packets per second. Since this is aggregated over 10 stations, each station can effectively transmit $4.416 / 10 = 0.4416$ packets per second, or approximately 26 packets per minute, at best.

The figure also shows the throughput for pure (unslotted) ALOHA. In this case, packet transmission can start at any time and is not limited to slot starts. Since in ALOHA a station does not check whether the channel is busy before transmitting, any packet whose transmission started within one packet time of the current transmission will collide with the current transmission, Figure 3-5 and Figure 3-9(b). Thus, a packet will not suffer a collision if no other packet is sent in an interval of two packet-times: $S = G \cdot P_0 = G \cdot e^{-2G}$. In this case, the maximum possible throughput of $1/2e \approx 0.184$ is achieved at $G = 0.5$ packets per packet time. In effect, the twofold size of the vulnerable period of pure Aloha yields in the twofold reduction in the maximum achievable throughput.

We can note that for any arrival rate less than $1/e$, there are two values of G for which the arrival rate equals the departure rate. We need a more accurate model to better understand the behavior of the system. Some of the problems with this simple model are:

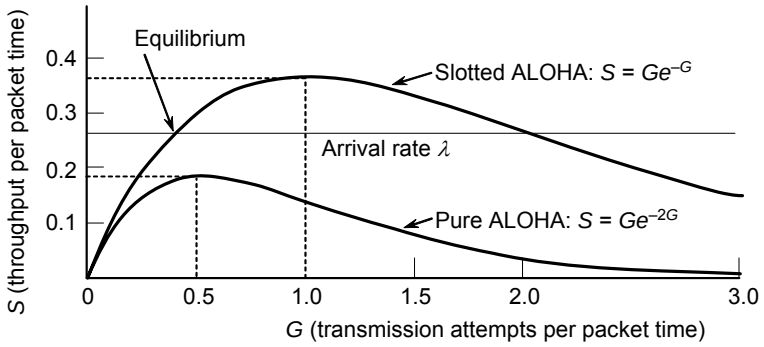


Figure 3-10: Efficiency of the ALOHA MAC protocol. (In case of Slotted ALOHA, the packet time is equal to the *slot time*.)

- It assumes constant arrival rate, but the arrival rate decreases with increasing number of backlogged nodes (stations) since there is no buffering
- It does not specify what is G—it simply assumes that the composite process (new arrivals plus retransmissions) can be approximated as a Poisson process with arrival rate G
- It does not say anything about the retransmission probabilities

To derive a more accurate model of the system, we assume a memoryless model of packet transmission. Normally, in slotted ALOHA a station waits a random integer number of time slots before trying to retransmit. Instead, we assume that a backlogged packet is retransmitted with probability q_r in each successive slot until a successful retransmission occurs.

The behavior of slotted ALOHA can now be represented as a discrete-time Markov chain (Figure 3-11). The state n represents the number of backlogged nodes at the beginning of a given slot. Each of these nodes will transmit a packet in a given slot, independently of each other, with probability q_r. Under the assumption 6 of no buffering, a total of $m–n$ nodes are free to accept newly arriving packets. Since new packets are arriving Poisson-distributed with mean λ/m, the probability of no arrivals is $e^{-\lambda/m}$. Thus, the probability that an unbacklogged node transmits a packet in the given slot is $q_a = 1 - e^{-\lambda/m}$. Let $Q_a(i, n)$ denote the probability that i unbacklogged nodes transmit packets in a given slot, and similarly $Q_r(i, n)$ denote the probability that i backlogged nodes transmit:

$$Q_a(i,n) = \binom{m-n}{i} \cdot (1-q_a)^{m-n-i} \cdot q_a^i \tag{3.4a}$$

$$Q_r(i,n) = \binom{n}{i} \cdot (1-q_r)^{n-i} \cdot q_r^i \tag{3.4b}$$

The transition probabilities P_{ij} in Figure 3-11 can be derived by using $Q_a(i, n)$ and $Q_r(i, n)$, see e.g., [Bertsekas & Gallagher 1992]. Clearly, a transmission will be successful if there is a one new arrival $Q_a(1, n)$ and no backlogged packet retransmissions $Q_r(0, n)$, or if there are no new arrivals $Q_a(0, n)$ and one backlogged packet retransmission $Q_r(1, n)$. The probability of a successful transmission $P_{success}$ may be expressed as follows:

$$P_{success} = Q_a(1,n) \cdot Q_r(0,n) + Q_a(0,n) \cdot Q_r(1,n) \tag{3.5}$$

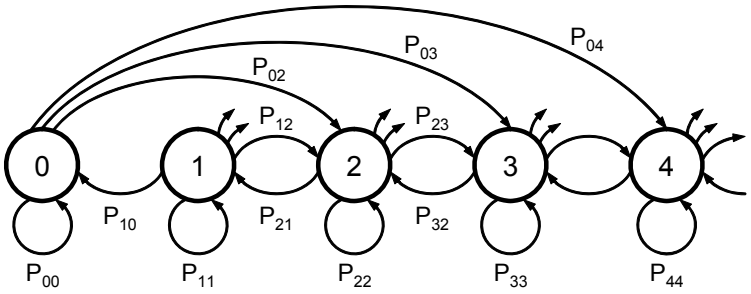


Figure 3-11: Markov chain representation for slotted ALOHA. The states represent the number of backlogged stations, which can decrease by at most one per transmission, but can increase by an arbitrary number. P_{ij} represents the probability of a transition from state *i* to state *j*.

where
$$Q_a(0,n) = (1-q_a)^{m-n} \quad Q_a(1,n) = (m-n)\cdot(1-q_a)^{m-n-1}\cdot q_a$$
$$Q_r(0,n) = (1-q_r)^{n} \quad Q_r(1,n) = n\cdot(1-q_r)^{n-1}\cdot q_r$$

Then,

$$P_{success} = \left[(m-n)\cdot q_a\cdot(1-q_r) + (1-q_a)\cdot n\cdot q_r\right]\cdot(1-q_a)^{m-n-1}\cdot(1-q_r)^{n-1}$$

For small q_a and q_r, we can introduce the following approximations:

$$1-q_a \approx 1;\ 1-q_r \approx 1;\ (1-q_a)^{m-n-1} \approx (1-q_a)^{m-n};\ \text{and,}\ (1-q_r)^{n-1} \approx (1-q_r)^{n}$$

which leads to

$$P_{success} \approx \left[(m-n)\cdot q_a + n\cdot q_r\right]\cdot(1-q_a)^{m-n}\cdot(1-q_r)^{n}$$

Further, it is evident that the transmission attempt rate, G, depends on the number of backlogged packets, n, as $G(n) = (m-n)\cdot q_a + n\cdot q_r$. Then we can write:

$$P_{success} \approx G(n)\cdot(1-q_a)^{m-n}\cdot(1-q_r)^{n}$$

Since q_a and q_r are small, we use the approximation $(1-x)^y \approx e^{-xy}$ for small x, and obtain that $(1-q_a)^{m-n} \approx e^{-q_a\cdot(m-n)}$ and $(1-q_r)^{n} \approx e^{-q_r\cdot n}$.

This yields in:

$$P_{success} \approx G(n)\cdot e^{-(m-n)q_a}\cdot e^{-n\cdot q_r} = G(n)\cdot e^{-(m-n)q_a - n\cdot q_r}$$

Finally,

$$P_{success} \approx G(n)\cdot e^{-G(n)} \tag{3.6}$$

Thus the previous intuitive analysis is essentially correct in that the number of packets in a slot is well approximated as a Poisson random variable, but the parameter $G(n)$ varies with the state. The efficiency diagram is shown in Figure 3-12. A useful parameter in understanding this system behavior is the *drift*, D_n, in state n as the expected change in backlog over one slot time, starting in state n:

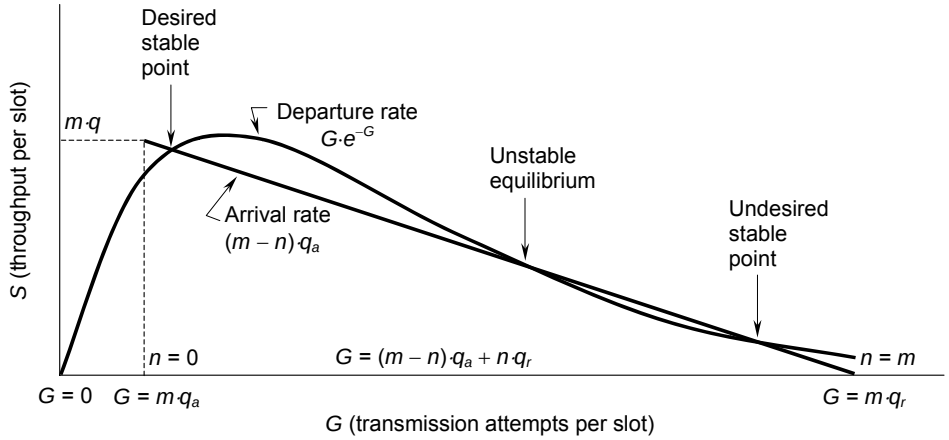


Figure 3-12: Efficiency of slotted ALOHA. The horizontal axis corresponds to both the state n and the attempt rate G, which are related by the linear equation $G = (m - n)q_a + nq_r$ with $q_r > q_a$.

$$\begin{aligned} D_n &= \text{(expected rate of accepted new arrivals)} - \text{(expected number of transmissions in one slot)} \\ &= (m - n)q_a - P_{success} \end{aligned}$$

In Figure 3-12, for n to the left of the unstable equilibrium, D_n is negative and n drifts toward the desired stable point. For n to the right of the unstable point, D_n is positive and n drifts toward the undesirable stable point. In this case, collisions occur in almost all successive slots and the system remains heavily backlogged for a long time. The interested reader is referred to other sources for further analysis of the system, e.g., [Bertsekas & Gallagher 1992; Hayes 1984; Keiser 2002].

3.3 Carrier Sensing Protocols

Unlike ALOHA where stations just talk whenever they need to, in carrier sensing protocols the stations *listen before talk* (or transmit).

In many multiaccess systems, such as local area networks, a station can hear whether other stations are transmitting after a very short propagation and detection delay relative to a packet transmission time. The *detection delay* is the time required for a physical receiver to determine whether or not some other station is currently transmitting. Notice that this implies positively identifying a packet transmission as opposed to random channel noise, so at least a packet preamble must be detected. If stations can detect idle periods quickly, it is reasonable to terminate idle periods promptly upon detection and to allow stations to initiate packet transmissions. This type of strategy, called *carrier sense multiple access* (CSMA), does not necessarily imply the use of a carrier but simply the ability to detect idle periods quickly.

The value of the parameter β (Figure 3-4) has direct effect on a CSMA protocol's performance. First, it is directly proportional to the vulnerable period and, as already mentioned, the longer the vulnerable period, the higher the collision probability. Second, in some variants of CSMA where

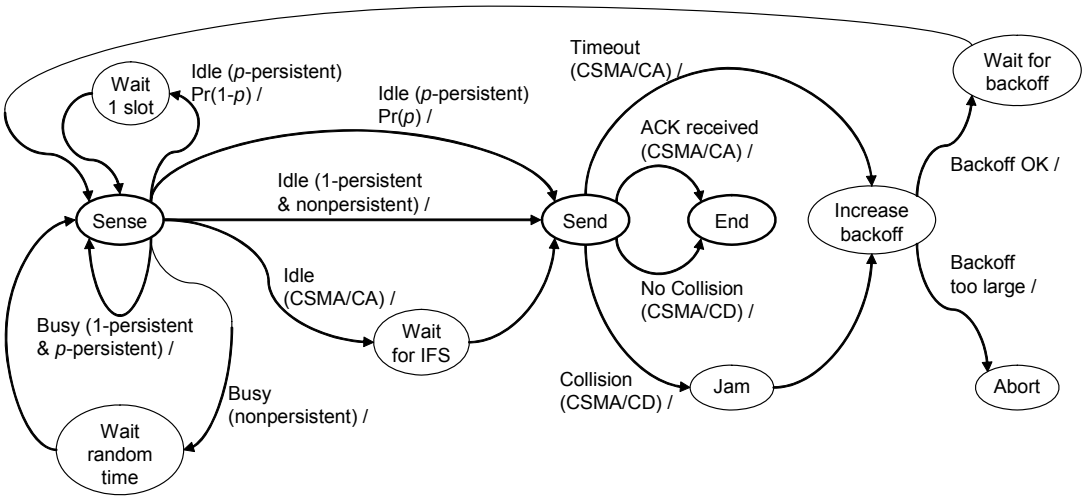


Figure 3-13: State diagram for the CSMA family of protocols. The various CSMA schemes differ in their persistence and how they detect a collision.

the sources listen to the medium simultaneously while transmitting (CSMA/CD, described below), they can detect collision soon after commencing the transmission and the time for collision detection is directly proportional to β. The smaller the detection period, the sooner the collision is detected so they can quickly resolve the contention and resume transmission. Thus, the performance of CSMA degrades with increasing β and also degrades with increasing channel rate and with decreasing packet size.

When using CSMA, a station that has a packet to transmit waits until the channel is silent before transmitting. When the channel is sensed to be idle, a station can take one of three different approaches (depending on the network design) to place a packet onto the channel: *nonpersistent CSMA*, 1-*persistent CSMA*, and *p-persistent CSMA*. The 1-persistent CSMA is a special case of the *p*-persistent scheme, but it is usually considered separately. These protocols differ by the action that a station with a packet to transmit takes after sensing the idle channel. However, when a station notes that a transmission was unsuccessful, in each protocol the scheduling of the packet retransmission is the same. In these reschedulings, the packet is sent again according to a randomly distributed retransmission delay. Table 3-1 summarizes these protocols, and Figure 3-13 gives a state transition diagram.

Table 3-1: Characteristics of the three basic CSMA protocols when the channel is sensed idle or busy. If a transmission was unsuccessful, all three protocols take the same action; see text for details.

CSMA Protocol	Transmission rules
Nonpersistent	If medium is idle, transmit. If medium is busy, wait random amount of time and sense channel again.
1-persistent	If medium is idle, transmit. If medium is busy, *continue sensing* until channel is idle; then transmit immediately.
p-persistent	If medium is idle, transmit with probability *p*. If medium is busy, *continue sensing* until channel is idle;

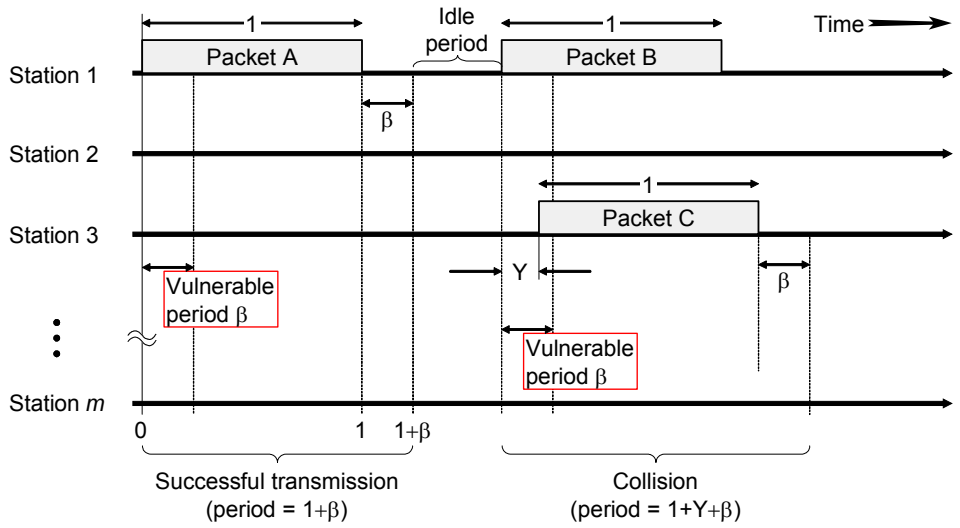


Figure 3-14. Successful and unsuccessful attempts for nonpersistent CSMA. Time is measured in units of the packet transmission time t_{xdat}.

	then transmit with probability p.

The efficiency of CSMA is better than that of ALOHA because of the CSMA's shorter vulnerable period: The stations will not initiate transmission if they sense a transmission already in progress. Notice that nonpersistent CSMA is less greedy than 1-persistent CSMA in the sense that, upon observing a busy channel, it does not continually sense it for the purpose of seizing it immediately upon detecting the end of the previous transmission. Instead, nonpersistent CSMA waits a random period of time and then repeats the procedure. Consequently, this protocol leads to better channel utilization but longer delays than 1-persistent CSMA.

Figure 3-14 illustrates transmission attempts for nonpersistent CSMA. The medium is decided idle if there are no transmissions for time duration β. Once transmission is started, it will be successful if no other station attempts transmission within the vulnerable period, also of duration β. During the vulnerable period the station essentially declares its intention to transmit and it simultaneously transmits. After this period all other station will hear a transmission in progress and will defer their own transmissions. Should another station attempt a transmission within the vulnerable period, a collision takes place.

The length of the vulnerable period determines the performance of a CSMA protocol—the shorter the vulnerable period, the lower the number of collisions, and therefore the better the channel throughput. The duration β of the vulnerable period can be manipulated as discussed in discussion of Figure 3-4 above.

P-persistent CSMA applies to slotted channels and works as follows. Upon sensing an idle channel, it transmits with a probability p. With a probability $q = 1 - p$, it defers until the next slot. If that slot is also idle, it either transmits or defers again, with probabilities p and q. The process is repeated until either the frame has been transmitted or another station has begun transmitting. In the latter case, the unlucky station acts as if there had been a collision (i.e., it waits a random time and tries again). If the station initially senses the channel busy, it waits until the next slot and applies the above algorithm.

3.3.1 Throughput Analysis of Nonpersistent CSMA

Here I provide throughput analysis for nonpersistent CSMA, similar to the one for slotted ALOHA presented above. The analysis for *p*-persistent versions is somewhat more involved and the reader should consult [Kleinrock & Tobagi 1975; Takagi & Kleinrock 1985; Sohraby *et al*. 1987]. This analysis follows the one in [Bertsekas & Gallagher 1992].

The analysis is similar to ordinary slotted ALOHA, with the difference that idle slots have duration β. Also, packet arrival is regarded as backlogged if it arrives while a transmission is in progress. The system can be represented with Markov chain, as in Figure 3-11. Again I assume a memoryless model of packet transmission for the purpose of analysis. This means that retransmission may take place in any subsequent idle slot with probability q_r. The time between state transitions (including self-transitions) is either β, for an idle slot, or $(1 + \beta)$, for a busy slot. The probability that a slot will be idle is the joint probability of no new arrivals and no retransmissions: $P_{idle} = e^{-\lambda \cdot \beta} \cdot (1-q_r)^n$. Conversely, the probability of a busy slot is $P_{busy} = 1 - P_{idle}$.

The expected time between state transitions assuming that the system is in state n is:

$$E[\text{time between state transitions}] = \beta \cdot P_{idle} + (1+\beta) \cdot P_{idle} = 1 + \beta - P_{idle} = 1 + \beta - e^{-\lambda \cdot \beta} \cdot (1-q_r)^n.$$

Thus, the expected number of arrivals between state transitions is simply this time multiplied by the probability of an arrival, λ,

$$E[\text{arrivals}] = \lambda \cdot [1 + \beta - e^{-\lambda \cdot \beta} \cdot (1-q_r)^n]$$

The expected number of departures between state transitions in state n is simply the probability of a successful transmission, Eq. (3.5), rewritten here as:

$$P_{success} = Q_a(1) \cdot Q_r(0,n) + Q_a(0) \cdot Q_r(1,n)$$

$Q_a(\cdot)$ is Poisson with parameter $\lambda \cdot \beta$ and $Q_r(\cdot, n)$ is binomial with parameter q_r. Then,

$$\begin{aligned} P_{success} &= \lambda \cdot \beta \cdot e^{-\lambda \cdot \beta} \cdot (1-q_r)^n + e^{-\lambda \cdot \beta} \cdot n \cdot q_r \cdot (1-q_r)^{n-1} \\ &= \left(\lambda \cdot \beta + \frac{n \cdot q_r}{1-q_r} \right) \cdot e^{-\lambda \cdot \beta} \cdot (1-q_r)^n \end{aligned}$$

It is again useful to define the drift D_n in state n as the expected change in backlog over the period between state transitions, i.e., the drift is the expected number of new arrivals less the expected number of departures between state transitions. Then,

$$D_n = E[\text{arrivals}] - P_{success} = \lambda \cdot \left[1 + \beta - e^{-\lambda \cdot \beta} \cdot (1-q_r)^n \right] - \left(\lambda \cdot \beta + \frac{q_r \cdot n}{1-q_r} \right) \cdot e^{-\lambda \cdot \beta} \cdot (1-q_r)^n$$

The drift characterizes the temporal trend of the system state. If the drift is negative, $D_n < 0$, then the system tends to become less backlogged, which is a desirable outcome. If $D_n > 0$, then the system tends to become more backlogged, which is an undesirable outcome leading to collisions in almost all successive slots. Applying the approximations $(1-q_r)^{n-1} \approx (1-q_r)^n \approx e^{-q_r \cdot n}$ and $1-q_r \approx 1$ for small q_r, we obtain:

$$D_n \approx \lambda \cdot \left(1 + \beta - e^{-\lambda \cdot \beta - q_r \cdot n} \right) - \left(\lambda \cdot \beta + q_r \cdot n \right) \cdot e^{-\lambda \cdot \beta - q_r \cdot n}$$

We define $g(n)$ in state n as the expected number of packet transmissions in the next idle slot (β time units) following transition to state n. This includes the new arrivals and the backlogged retransmissions, $g(n) = \lambda \cdot \beta + q_r \cdot n$. Applying $g(n)$ to the above expression for drift yields:

$$D_n = \lambda \cdot \left(1 + \beta - e^{-g(n)}\right) - g(n) \cdot e^{-g(n)}$$

Clearly, D_n is negative if:

$$\lambda < \frac{g(n) \cdot e^{-g(n)}}{1 + \beta - e^{-g(n)}} \tag{3.7}$$

The numerator is the expected number of departures per state transition and the denominator is the expected duration of a state transition period. Thus, the ratio of (3.7) is also the number of departures per unit time, i.e., the system throughput:

$$E[\text{throughput}] = \frac{g(n) \cdot e^{-g(n)}}{1 + \beta - e^{-g(n)}}$$

This function is shown in Figure 3-15(a). The expected system throughput is maximized at $g(n) = \sqrt{2 \cdot \beta}$. To see this, we take the derivative of the throughput with respect to $g(n)$ and find the zero value. Letting $g = g(n)$, we have:

$$\frac{d}{dg(n)}\left[\frac{g(n) \cdot e^{-g(n)}}{1 + \beta - e^{-g(n)}}\right] = \frac{\left(1 + \beta - e^{-g}\right) \cdot \left(e^{-g} - g \cdot e^{-g}\right) - g \cdot e^{-g} \cdot \left(e^{-g}\right)}{1 + \beta - e^{-g}} = 0$$

We equate the numerator with zero and after few manipulations we have: $1 + \beta - (1 + \beta) \cdot g - e^{-g} = 0$. Applying the approximation $e^{-g} \approx 1 - g + g^2/2$, we have: $1 + \beta - (1 + \beta) \cdot g - (1 - g + g^2/2) = 0$, which leads to $g^2/2 - \beta \cdot g + \beta = 0$. Applying the quadratic formula yields: $g = \beta \cdot \left(\sqrt{1 + 2/\beta} - 1\right)$, which is not exactly what we are looking for. However, assuming that β, the duration of an idle slot as a fraction of data slot, is small, $\beta \ll 1$, we obtain the sought for result: $g(n) = \sqrt{2 \cdot \beta}$.

At $g(n) = \sqrt{2 \cdot \beta}$, the expected value of throughput is $E[\text{throughput}] \approx 1 / \left(1 + \sqrt{2 \cdot \beta}\right)$. This is indicated in Figure 3-15(a). Thus, for small β, the throughput is very close to 1.

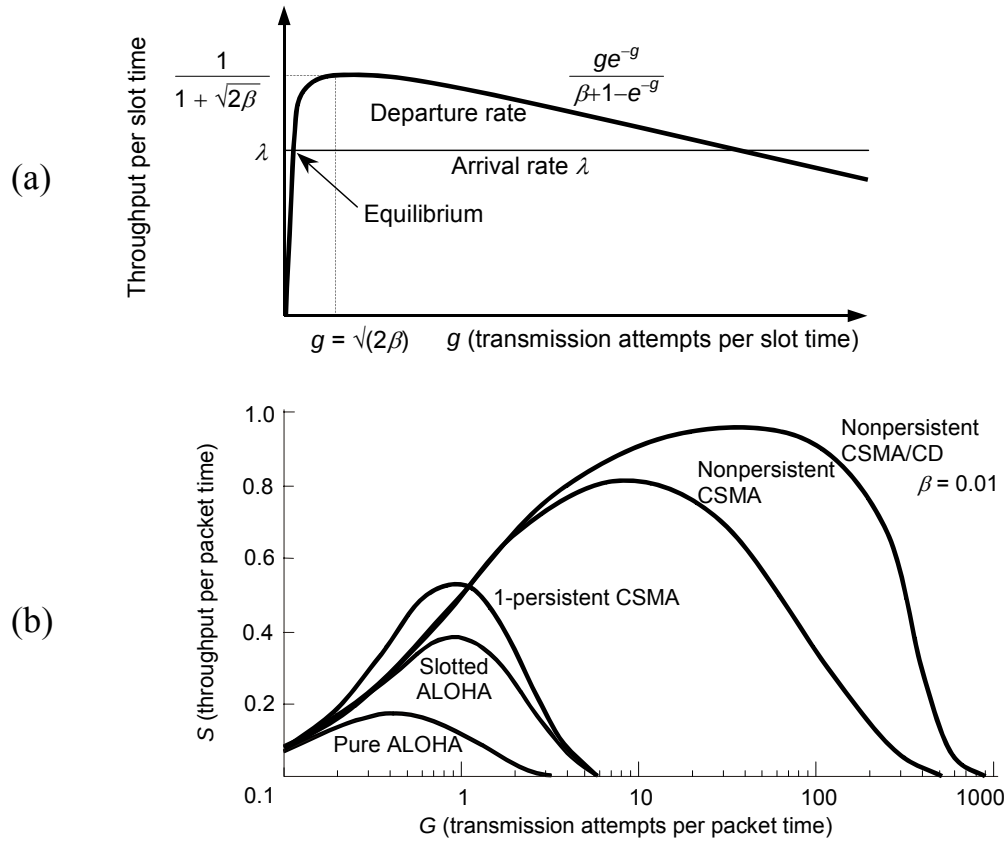


Figure 3-15. (a) Throughput of CSMA protocol as a function of the attempted transmission rate *g* in packets per idle slot. (b) Efficiency comparison of different CSMA MAC protocols. Notice the log scale on the horizontal axis. For CSMA techniques, increasing the number of attempts translates into decreasing the slot size (determined by parameter *β*).

Figure 3-15(b) shows the efficiency of CSMA protocols in terms of the transmission attempts per packet time, which is usually given in the literature. The reader may be puzzled about the relationship between Figure 3-15(a) and Figure 3-15(b)—the difference arises due to different abscissas. Notice that for CSMA techniques, increasing number of attempts per packet time translates into decreasing slot size relative to packet time, which is determined by β parameter. Naturally, if there is more than one transmission attempt per β time there will result a collision. If only one station starts transmission at the beginning of a slot, the others will sense it and will not transmit during the subsequent slots. Thus, the shorter the β, the better utilization is possible; but, as already pointed out β cannot be shorter than what propagation and detection delay of the physical system dictate. Figure 3-15(b) shows the throughput in terms of attempt rate per packet time. Note that G is the attempt rate, not the count of attempts. Hence, the higher the rate, the shorter the time between transmissions. However, once a transmission starts, the other stations will sense busy channel and defer their transmission.

Further analysis of CSMA performance is beyond the scope of this text and the interested reader is referred to, e.g., [Bertsekas & Gallagher 1992; Hayes 1984].

3.3.2 CSMA/CD

Persistent and nonpersistent CSMA protocols are clearly improvement over ALOHA because they ensure that no station begins to transmit when it senses the channel busy. Another improvement is for stations to abort their transmissions as soon as they detect a collision[3]. Quickly terminating damaged packets saves time and bandwidth. This protocol is known as *CSMA with Collision Detection*, or CSMA/CD, which is a variant of 1-persistent CSMA. It works as follows:

1. Wait until the channel is idle.
2. When the channel is idle, transmit immediately *and* listen during the transmission.
3. If collision happens, abort the packet transmission, and then wait for a random delay and go to step 1.

The protocol requires that the transmitter detect the collision before it has stopped transmitting its packet. Thus the transmission time of the smallest packet must be larger than one round-trip propagation time, i.e., 2β. This requirement limits the distance between two computers on the wired Ethernet LAN. The smallest packet is 64 bytes. This 64-byte value is derived from the original 2500-m maximum distance between Ethernet interfaces plus the transit time across up to four repeaters plus the time the electronics takes to detect the collision. The 64 bytes correspond to 51.2 μs, which is larger than the round-trip time across 2500 m (about 18 μs) plus the delays across repeaters and to detect the collision.

It is important to realize that collision detection is an analog process. The station's hardware must listen to the cable while it is transmitting. If what it reads back is different from what it is putting out, it knows that a collision is occurring. The implication is that the signal encoding must allow collisions to be detected (e.g., a collision of two 0-volt signals may well be impossible to detect).

3.3.3 CSMA/CA

In wireless LANs it is not practical to do collision detection because of two main reasons:

1. Implementing a collision detection mechanism would require the implementation of a full duplex radio, capable of transmitting and receiving at once. Unlike wired LANs, where a transmitter can simultaneously monitor the medium for a collision, in many wireless LANs the transmitter's power overwhelms a collocated receiver. The dynamic range of the signals on the medium is very large. This is mainly result of the propagation loss, where the signal drops exponentially from its source (remember Figure 2-14(a)!). Thus, a transmitting station cannot effectively distinguish incoming weak signals from noise and the effects of its own transmission.
2. In a wireless environment we cannot assume that all stations hear each other, which is the basic assumption of the collision detection scheme. Again, due to the propagation loss we have the following problem. The fact that the transmitting station senses the medium free

[3] In networks with wired media, the station compares the signal that it places on the wire with the one observed on the wire to detect collision. If these signals are not the same, a collision has occurred.

does not necessarily mean that the medium is free around the receiver area. (This is the so called "hidden station problem," to be considered below.)

As a result, when a station transmits a packet, it has no idea whether the packet collided with another packet or not until it receives an acknowledgement from the receiver (or times out due to the lack of an acknowledgement). In this situation, collisions have a greater effect on performance than with CSMA/CD, where colliding packets can be quickly detected and aborted. Thus, it makes sense to try to avoid collisions, if possible, and a popular scheme for this is *CSMA/Collision Avoidance*, or CSMA/CA. The IEEE has standardized CSMA/CA as the IEEE 802.11 standard, see Chapter 4 below. CSMA/CA is essentially *p*-persistence, with the twist that when the medium becomes idle, a station must wait for a time period to learn about the fate of the previous transmission before contending for the medium. After a packet was transmitted, the maximum time until a station detects a collision is twice the propagation time of a signal between the stations that are farthest apart plus the detection time. Thus, the station needs at least 2β units. The time interval between packets (frames) required for carrier sense mechanism to determine that the medium is idle and available for transmission is called *interframe space* (IFS). A station gets a higher priority if it is assigned a smaller interframe spacing.

When a station wants to transmit data, it first senses the medium whether it is busy. Two basic rules apply here:

1. If the medium has been idle for longer than an IFS corresponding to its priority level, transmission can begin immediately.
2. If the medium is busy, the station enters the *access deferral* state. The station continuously senses the medium, waiting for it to become idle. When the medium becomes idle, the station first waits for an IFS, then sets a *contention timer* to a time interval randomly selected in the range [0, CW–1], where CW is a predefined contention window length. The station can transmit the packet after this timer expires.

After transmitting a packet, the station waits for the receiver to send an ACK. If no ACK is received, the packet is assumed lost to collision, and the source tries again, choosing a contention timer at random from an interval twice as long as the one before (*binary exponential backoff*). The decrementing counter of the timer guarantees that the station will transmit, unlike a *p*-persistent approach where for every slot the decision of whether or not to transmit is based on a fixed probability p or q_r. Thus regardless of the timer value a station starts at, it always counts down to zero. If the station senses that another station has begun transmission while it was waiting for the expiration of the contention timer, it does not reset its timer, but merely freezes it, and restarts the countdown when the packet completes transmission. In this way, stations that happen to choose a longer timer value get higher priority in the next round of contention.

As it can be seen, CSMA/CA deliberately introduces delay in transmission in order to avoid collision. Avoiding collisions increases the protocol efficiency in terms of the percentage of packets that get successfully transmitted (useful throughput). Notice that efficiency measures only the ratio of the successful transmission to the total number of transmissions. However, it does not specify the delays that result from the deferrals introduced to avoid the collisions. Figure 3-16 shows the qualitative relationship for the average packet delays, depending on the packet arrival rate.

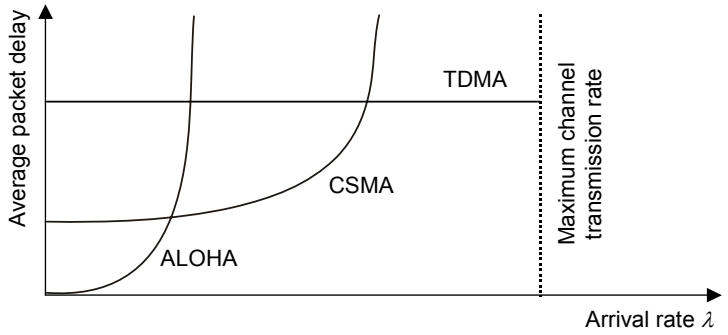


Figure 3-16. Qualitative relationship of average packet delays to packet arrival rate for different MAC protocols.

3.4 Other MAC Protocols

The MAC protocols described above represent only a small selection of different multiaccess communication techniques proposed. I have put emphasis on contention-based MAC protocols. The unlicensed PCS bands spectrum etiquette (Pahlavan and Krishnamurthy, p. 423):

Listen before talk (or transmit) *LBT protocol*

Low transmitter power

Restricted duration of transmission

IEEE 802.11 – Wi-Fi, described in detail in Chapter 4 below.

ETSI – HIPERLAN

IEEE 802.15 – Bluetooth

IEEE 802.16 – Fixed Wireless (Wi-Max)

ZigBee Alliance http://www.zigbee.org/

ZigBee and IEEE 802.15.4 Resource Center http://www.palowireless.com/zigbee/

Other MAC protocol types include reservation-based ones, which are contention-free (collision-free).

3.5 Multiple-access Interference

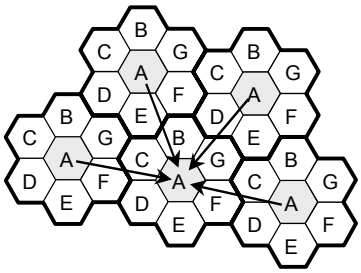


Figure 3-17. Effect of large transmit power on cochannel cells. Letter labels indicate the channel frequencies.

The scarcity of radio spectrum has resulted in frequency reuse and cellular architecture for wireless networks. Signals from mobile stations cause interference among stations in the same or different cells. The objective of a cellular radio system is to allocate frequencies or channels to cells within a cluster so that the distance between interfering cells (cochannel or adjacent channel) is as large as possible. The conflicting requirement of maximum frequency reuse leads to interference. An example channel allocation is shown in Figure 3-17. The symptoms of interference can be simplified as:

- At the physical layer, interference degrades the channel quality and results in increased channel BER
- At the link/MAC layer, interference shows as increased traffic load on the channel, i.e., busy channel and/or packet "collisions"

Although grossly simplified, this classification provides a useful starting point in considering the effects of multiple-access interference (MAI).

3.5.1 Physical Layer Interference

Power control is essential in cellular systems for high system capacity and satisfactory system quality by (*i*) limiting interference among users in different cells using the same frequency channel (i.e., cochannel interference), and (*ii*) alleviating the near-far effect which reduces interference among users in the same or different cells using adjacent frequency channels. [*Near-far effect* is a condition in which a nearby transmitter captures the receiver of the mobile or base station so that the latter is unable to detect the signal of a second transmitter located farther away.] In CDMA systems, power control is required for system operation, as CDMA system capacity is interference limited.

An important performance measure of in cellular system design is *signal-to-interference ratio* (SIR), which can be written as follows:

$$SIR = \frac{P_{desired}}{\sum_i P_{\text{int}erference,i}} \tag{3.8}$$

Here $P_{desired}$ is the signal strength from the desired BS and $P_{interference,i}$ is the signal strength from the *i*th interfering BS. The signal strength falls as (2.9). Suppose there are two base station transmitters BS_1 and BS_2 located in an area with the same transmit power P_T and a mobile station

is at a distance d_1 from the first and d_2 from the second. If the MS is trying to communicate with BS_1, the signal from BS_2 is interference. The SIR for this mobile station is:

$$SIR = \frac{GP_T \Big/ f_c^2 \cdot d_1^{\alpha}}{GP_T \Big/ f_c^2 \cdot d_2^{\alpha}} = \left(\frac{d_2}{d_1} \right)^{\alpha}$$

The larger the ratio d_1/d_2 is, the greater is SIR and the better is the performance. To avoid interference, a mobile station or a BS should operate at the *lowest possible SIR* so that the communication quality is acceptable. This may at first appear counterintuitive, since one would expect that it is important to maintain a high SIR for good communication quality. However, a large transmit power in one frequency channel in one cell results in a large cochannel interference in all the closest cochannel cells that employ the same frequency channel, albeit at a sufficient distance away from the given BS. This reduces the communication quality all around and is not desirable.

Power control procedures determine the power levels and timings for transmission to and from the station. Controlling power level works in such a way that, as the mobile station moves closer to the base station it is using, the system reduces the power in order to reduce interference to other transmissions. Conversely, as the station moves further from the base station, the system may increase the power levels in order to maintain transmission quality. Power control procedures can also promote long battery life by reducing the power radiated at a terminal to the minimum needed to meet transmission quality objectives.

Near-far effect is a condition in which a nearby transmitter captures the receiver of the mobile or base station so that the latter is unable to detect the signal of a second transmitter located farther away. This can be considered both as detrimental and beneficial. Some drawbacks were mentioned above, but a benefit can be in that what otherwise would result in a collision, could end up as a successful transmission.

A useful parameter is SINR, and if SINR exceeds a certain threshold, the receiver will successfully obtain the packet even if some other stations are transmitting simultaneously in the same or neighboring frequency band. Thus, the equation for successful transmission, e.g., (3.5), must be modified to incorporate this possibility.

3.5.2 Link Layer Interference: Hidden and Exposed Stations

Wireless broadcast networks show some phenomena not present in wireline broadcast networks. The air medium is partitioned into broadcast regions, rather than being a single broadcast medium. In 802.11, the partitions are called BSSs. The phenomena described here are due to the facts that: (*i*) not all stations within a partition can necessarily hear each other; and, (*ii*) the broadcast regions can overlap. The former causes the hidden station problem and the latter causes the exposed station problem.

Unlike the wireline broadcast medium, the transitivity of connectivity does not apply. In wireline broadcast networks, such as Ethernet, if station *A* can hear station *B* and station *B* can hear station

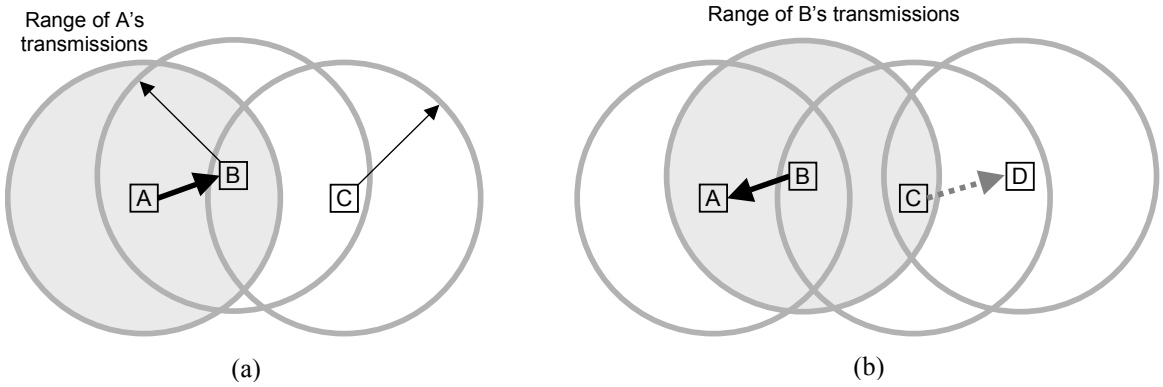


Figure 3-18. (a) Hidden station problem: C cannot hear A's transmissions. (b) Exposed station problem: C defers transmission to D because it hears B's transmission.

C, then station *A* can hear station *C*. This is not always the case in wireless broadcast networks, as seen in Figure 3-18(a). In the *hidden station problem*, station *C* cannot hear station *A*'s transmissions and may mistakenly conclude that the medium is available. If *C* does start transmitting, it will interfere at *B*, wiping out the frame from *A*. Generally, a station *X* is considered to be hidden from another station *Y* in the same receiver's area of coverage if the transmission coverage of the transceivers at *X* and *Y* do not overlap.

Different air partitions can support multiple simultaneous transmissions, which are successful as long as each receiver can hear at most one transmitter at a time. In the *exposed station problem*, station *C* defers transmission to *D* since it hears *B*'s transmission, as illustrated in Figure 3-18(b). If *C* senses the medium, it will hear an ongoing transmission and falsely conclude that it may not send to *D*, when in fact such a transmission would cause bad reception only in the zone between *B* and *C*, where neither of intended receivers is located. Thus, carrier sense is insufficient to detect all transmission on the wireless medium.

Hidden and exposed station problems arise only for CSMA-type protocols. ALOHA, for instance, does not suffer from such problems since it does not perform channel sensing before transmission. Under the hidden stations scenario, the performance of CSMA degenerates to that of ALOHA, since carrier sensing mechanism essentially becomes useless. With exposed stations it becomes worse since carrier sensing prevents the exposed stations from transmission, where ALOHA would not mind the busy channel.

The key insight is that *collisions are located at receiver*. A solution is to use the receiver's medium state to determine transmitter behavior. IEEE 802.11 specifies *Multiple Access with Collision Avoidance for Wireless* (MACAW) for which includes three-frame atomic exchange shown in Figure 3-19(b). The basic idea behind this scheme is for the sender to stimulate the receiver into outputting a short frame, so the stations in the receivers range can detect this transmission and avoid transmitting for the duration of the upcoming (large) data frame. Through this indirection, the sender performs "floor acquisition" so it can speak unobstructed since all other stations will remain silent for the duration of transmission. The rationale is, should a collision happen, it should last only a short period. This objective is achieved by allowing collisions only on RTS packets. Since RTS (and CTS) packets are very short, unlike data packets which can be very long, the collision duration is minimized.

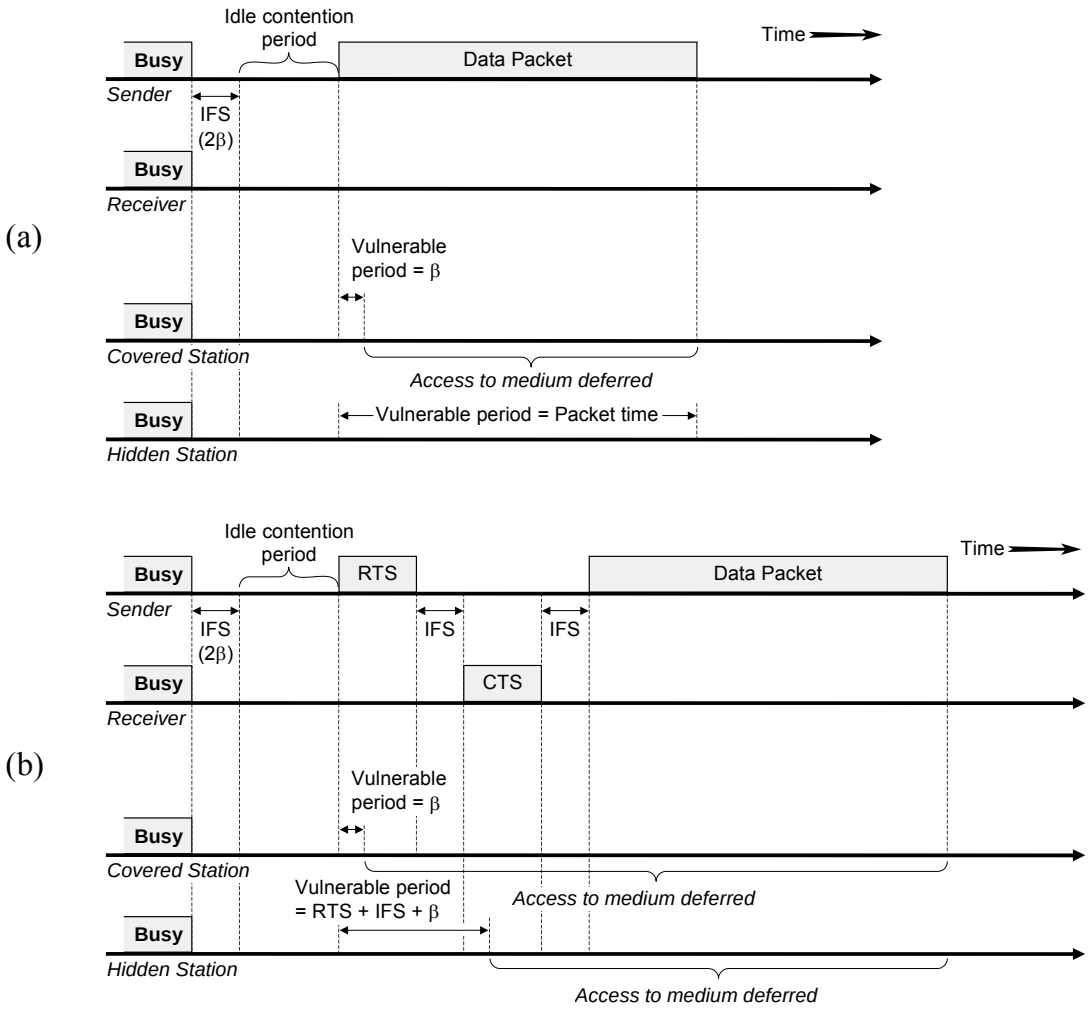


Figure 3-19: (a) Basic CSMA transmission mode. (b) The MACA protocol atomic unit exchange consists of three packets: RTS, CTS, and Data.

The source goes through a regular procedure, as with any other packet (Figure 3-13). It first waits for medium to become idle. Next, it waits for an interframe spacing and an idle contention period. Then it issues a *request to send* (RTS) message to its intended receiver, see Figure 3-19(b). If the RTS succeeds, the receiver returns a *clear to send* (CTS) reply. After receiving the CTS, the source transmits the data frame. If a station overhears an RTS, it waits long enough for a CTS message to be sent by a receiver before it tries to send any data. If a station hears a CTS (which carries the length of the forthcoming data packet in its body), it refrains from transmission long enough for the forthcoming packet transmission to complete. If a station heard RTS but *not* CTS, it assumes that its own transmission will not disturb the parallel transmission, and it proceeds with its own transmission[4].

[4] In the case of IEEE 802.11 RTS/CTS implementation, stations refrain from transmission even if they hear only RTS and not CTS. In other words, IEEE 802.11 does not solve the exposed station problem, as will be seen below.

By comparing Figure 3-19(a) and Figure 3-19(b), we can see that MACAW significantly shortens the vulnerable period of hidden stations, at the expense that the overall atomic unit exchange lasts somewhat longer.

3.6 Summary and Bibliographical Notes

This chapter provides a brief introduction to the issues of multiaccess communications. An in-depth coverage is in [Bertsekas & Gallagher 1992; Hayes 1984]. Although these references are lacking the latest MAC protocols, the general principles remain the same.

Hidden station issues were considered and the MACA RTS/CTS protocol proposed in [Bharghavan *et al*. 1994; Fullmer & Garcia-Luna-Aceves 1997].

Problems

Problem 3.1

Problem 3.2

Assume a wireless broadcast protocol based on a Stop-and-Wait ARQ, with m senders and one receiver contending for the same broadcast medium. Assume that all packets are of equal length. Assume an integer number CW, such that: $0 < CW_{min} \le CW \le CW_{max} < \infty$. Every new packet is transmitted as follows:

1. If the medium is currently busy, wait until it becomes silent. Always observe that the medium is silent for T_W length of time before proceeding.
2. If in Step 1 the medium was at first busy or if this is a retransmission, select a random number $r_n \in \{0, 1, \ldots, CW-1\}$. Listen to the medium and wait for r_n time units, decrementing r_n until it becomes 0. If the medium becomes busy before r_n becomes 0, go to Step 1. If $r_n = 0$ and the medium is idle, transmit immediately.
3. If an acknowledgement is *not* received within a timeout time, set $CW := 2 \times CW$; If CW > CW_{max}, set $CW := CW_{max}$.

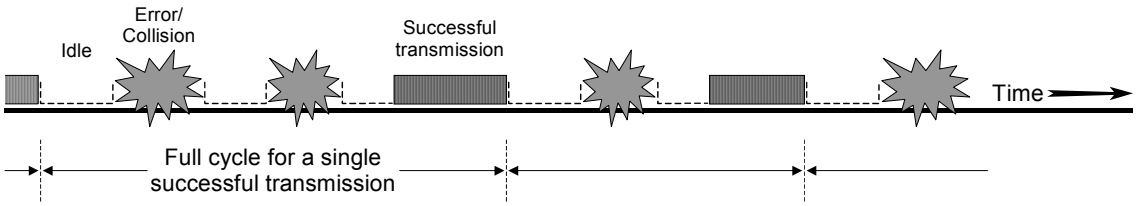


Figure 3-20: Typical sequence of states on a multiaccess channel. As indicated, each successful transmission is preceded by a variable number of idle and collision events.

4. If an acknowledgement is received, set $CW := CW_{min}$.

(a) Derive the collision probability in terms of CW size. What is the upper limit on the number of stations that can be supported for a given CW_{max} so that $Pr(\text{Collision}) < 1$?

(b) How should the timeout times be set? Can you give any estimates on the lower/upper bounds?

(c) Prove the correctness (safety and liveness) of the protocol.

(d) What is the minimum number of bits for the sequence numbers?

Problem 3.4

Derive Eq. (3.1) in Section 3.1.1 above. The diagram in Figure 3-20 may be helpful in the process. [Compare this with Figure 4-21 for the case of IEEE 802.11 wireless networks.] Notice that the channel probabilities are interpreted as frequencies of different states.

Problem 3.5

A number of stations use slotted-Aloha random-access strategy to communicate with a remote controller over a 2400-bps common channel. Each station transmits a 200-bit packet, on the average, once every 2 minutes (assume Poisson arrivals). What should be the slot size? What is the maximum number $m_{\max}$ of such stations that may use the channel with the optimum efficiency? What if the packet lengths are increased to 500 bits? What if stations transmit packets once every 3 minutes, on the average? How do the results change if the line speed is increased to 4800 bps?

Problem 3.6

In a pure-Aloha scenario, what is the probability that at least one collision will take place? Derive the formula in the manner that was done for Eq. (3.3). Given the probability that at least one collision, what is the expected average number of retransmissions? (Show the work.) Can observations about channel activity be used to determine when a station should transmit? (Explain your answer.)

Problem 3.7

For the Markov chain in Figure 3-11, explain why state "0" can transition to any other state except state "1," i.e., why $P_{01} = 0$? What events are captured by self-transitions, i.e., the probabilities P_{ii}?

Problem 3.8

A slotted Aloha system is operating at the optimum utilization (maximum possible throughput). What is the expected number of transmission attempts needed per single packet? What fraction of slots goes empty/idle in this system? What fraction of slots goes idle when the system is operating at a lower-than-optimum utilization?

Problem 3.9

In Figure 3-12, find the value of n for which the average number of backlogged stations after every subsequent time slot grows larger than before it, i.e., $n(t + 1) \geq n(t)$? (Show the work.)

Problem 3.10

Consider a slotted Aloha system with m stations, a unitary slot length, and the channel transmission attempt rate G. Assume that the number of backlogged stations, n, remains constant, on average; or that G does not depend on n. Derive the following probabilities:

(a) A new packet succeeds on the first transmission attempt

(b) A new packet suffers exactly K collisions and then a success

Problem 3.11

In a slotted Aloha system m stations compete for the time slots of a common channel. The slots have duration τ. In any given time slot, every station attempts to transmit with probability p, independent of the others.

(a) What is the probability that a single station will try to transmit in a given slot?

(b) What is the expected time until a single station transmits? (In other words, how long until exactly one station transmits?)

Problem 3.12

Consider a slotted Aloha system with n active stations that operate as follows. Instead of considering new arrivals and retransmissions separately as in Eqs. (3.4), each station transmits with an aggregate probability q_t, which aggregates both new arrivals and retransmissions.

(a) What is the probability of success, P_{success}, for transmission attempts?

(b) For a fixed value of n, what is the optimal value of q_t that maximizes the probability of success?

(c) Now assume that the only thing we know about n is $A \le n \le B$. Which value q_t maximizes P_{success} for the worst $n \in [A, B]$? What is this "worst case optimal" value for q_t if $A = 100$ and $B = 200$?

Problem 3.13

Explain why the contention period in a CSMA protocols has to be 2β, where β is the normalized propagation delay. (Why is duration β not sufficient?)

Problem 3.14

If the minimum packet size for the IEEE 802.3 LAN standard is 64 bytes and the data rate of the network is 10 Mbps then what is the normalized propagation delay (parameter β) for such a network?

Problem 3.15

Figure 3-19 shows the vulnerability periods for base and MACA transmission modes of CSMA/CA. For a given value of parameter β, determine the threshold packet length for which the total packet delay is shorter for MACA than for the base mode, under the hidden stations scenario.

Problem 3.16

Consider a system comprising three CSMA/CA-based wireless stations, *A*, *B*, and *C*; all three within the hearing range of each other and all are backlogged. Assume that IFS expired and the channel is eligible for transmission. List the set of all possible outcomes (sample space) on the broadcast channel, i.e., a single station transmits, two stations transmit, etc. Use Eqs. (3.4) to derive the probabilities for each of those events. The channel can be in one of three possible states: idle, successful, or colliding, as illustrated in Figure 3-3.

(a) Looking from the *network perspective*, derive the probabilities of channel being in idle state, p_{idle} = Pr[no station transmits],
a successful transmission, $p_{success}$ = Pr[single station transmits],
and collision, $p_{collision}$ = Pr[two or more stations transmit].

(b) Looking from a *single station perspective*, say station *A*'s, derive the channel state conditional probabilities
$\tilde{p}_{idle}$ = Pr[no station transmits | *A* doesn't transmit],
$\tilde{p}_{success}$ = Pr[only *A* transmits | at least *A* transmits], and
$\tilde{p}_{collision}$ = … ? (Define this yourself)

Discuss the reasons for the difference between these two perspectives. What is the relationship between these two perspectives?

Problem 3.17

Does CSMA prevent collisions completely? Explain your answer.

Problem 3.18

Describe a scenario in which 1-persistent CSMA performs better than *p*-persistent CSMA.

Problem 3.19

Consider a system of n stations operating under the nonpersistent CSMA protocol. Suppose there are K slots and each contending station independently picks a slot r with probability p_r. We refer to the distribution $p_1, p_2, \ldots, p_K$ as p. What is the probability of successful transmission when all stations select a contention slot using probability distribution p?

Problem 3.20

Suppose two stations are using nonpersistent CSMA with a modified version of the binary exponential backoff algorithm. In the modified algorithm, each station will always wait 0 or 1 time slots with equal probability, regardless of how many collisions have occurred.

(a) What is the probability that contention ends (i.e., one of the stations successfully transmits) on the first round of retransmissions?

(b) What is the probability that contention ends on the second round of retransmissions (i.e., success occurs after one retransmission ends in collision)?

(c) What is the probability that contention ends on the third round of retransmissions?

(d) In general how does this algorithm compare against the nonpersistent CSMA with the normal binary exponential backoff algorithm in terms of performance under different types of load?

Problem 3.21

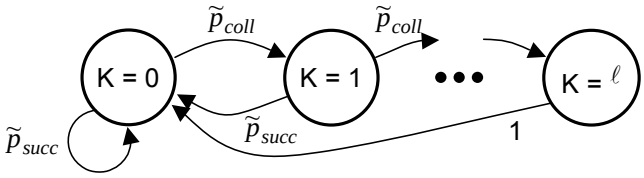
In CSMA/CA, stations will not begin transmitting user data until they have reserved a channel. Explain why is exponential backoff still required in CSMA/CA.

Problem 3.22

Assume a *p*-persistent CSMA/CA system with m stations of which on the average $\overline{n}$ are backlogged. After an IFS, each backlogged station considers transmitting within each subsequent idle slot, and decides to transmit with a *fixed* probability $p = q_r$. Once a transmission takes place, it will result in either a success or a collision. The conditional probabilities for the outcome of a packet transmission are:

$$\widetilde{p}_{\text{success}} = \frac{Q_r(1,n)}{1 - Q_r(0,n)} \qquad \text{and} \qquad \widetilde{p}_{\text{collision}} = 1 - \widetilde{p}_{\text{success}}$$

The Markov chain representation of the random process K counting the number of collisions a packet suffers is shown in the figure, assuming that the station drops the packet after ℓ unsuccessful retry attempts. If the packet suffers ℓ unsuccessful retries, it is dropped and the station moves on to the next packet. [Compare this with Figure 3-11 where the Markov chain counts the number of backlogged stations in the system.]



(Assume that q_r remains fixed regardless of the number of collisions the packet suffered.) Derive the stationary distribution of the Markov chain, the probabilities $b_k = \sum [b_j \cdot \Pr(K = k \mid K = j)]$ of station suffering k collisions for a given packet. For example,

$$b_0 = b_0 \cdot \widetilde{p}_{succ} + b_1 \cdot \widetilde{p}_{succ} + \ldots + b_\ell \cdot 1$$

and generally

$$b_k = \widetilde{p}_{coll} \cdot b_{k-1} = \widetilde{p}_{coll}^{k} \cdot b_0, \quad k = 1, 2, \ldots, \ell$$

We can write a matrix form $\mathbf{b} = \mathbf{P} \cdot \mathbf{b}$, where $\mathbf{P}$ is the transition probability matrix in steady state. Solve for $\mathbf{b}$ and plot it as a graph; assume $q_r = 0.1$ and $\ell = 6$. How does it depend on n and q_r? (Hint: $\sum_k b_k = 1$)

Problem 3.23

Using the parameters from Problem 3.22, determine the probability p_{drop} that station will reach the retry limit ℓ and drop the current packet.

Problem 3.24

On the example of Figure 3-18 provide a step-by-step explanation of how MACA solves both hidden and exposed station problems. Draw also the time diagrams.
Pay particular attention to the exposed terminal case, where B must be unobstructed by C to hear A's CTS and what is D allowed to hear from B↔A conversation before replying with CTS to C's RTS.

Problem 3.25

Suppose two stations are both sending to the same receiver using MACA, and suppose that both have a steady stream of packets to send. If each has an initial timeout of 1 second, describe a possible sequence of actions that leads to one of the stations being shut out.

Problem 3.26

Figure 3-19 shows the vulnerability periods for base and MACA transmission modes of CSMA/CA. For a given value of parameter β, determine the threshold packet length for which the total packet delay is shorter for MACA than for the base mode, under the hidden stations scenario.

Problem 3.27

Show that, if the bit errors are positively correlated, then *PER* for packets of N bits is smaller than $1 - (1 - BER)^N$.

Problem 3.28

Assume the lengths of successive packets are random and that the bit errors are all independent and occur with probability *BER*. Explain how to calculate *PER*.

Problem 3.29

Chapter 4
IEEE 802.11 Wireless LAN

Contents

4.1 Introduction

This chapter gives an example wireless local area network (LAN), which illustrates an implementation of the concepts described above. IEEE 802.11 is a member of the IEEE 802 family, which is a series of specifications for LAN technologies. (See http://standards.ieee.org for online information.) In Europe, an equivalent LAN standard is called HiperLAN and it is being standardized by ETSI (the European Telecommunications Standards Institute). IEEE 802 specifications are focused on the two lowest layers of the OSI model because they incorporate both physical and link layer components. All 802 networks have both a MAC and a Physical (PHY) component. The MAC is a set of rules to determine how to access the broadcast medium and send data, but the details of transmission and reception are left to the PHY. One of the key goals for 802.11 has been to emulate 802.3 LAN for wired networks.

IEEE 802.11 standard was first ratified in 1997. It supported 1 and 2 Mbps transmission rates in the 2.4 GHz band. Two high rate PHY's were ratified in 1999: 802.11a which supports 6 to 54 Mbps in the 5 GHz band, and 802.11b which supports 5.5 and 11 Mbps in the 2.4 GHz band.

4.1.1 Architecture and Services

There are four major components of the WLAN described by IEEE 802.11 (Figure 4-1). The standard focuses on the *mobile station* and the *access point* (AP) and defines a complete management protocol between the mobile station and AP. This management protocol makes it

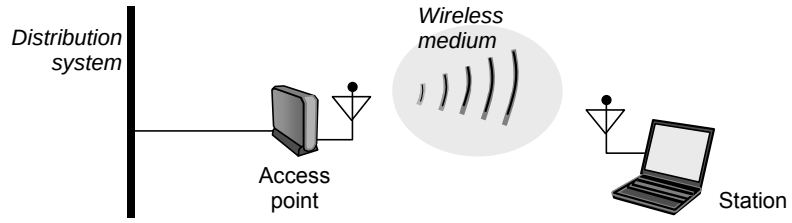


Figure 4-1: Components of 802.11 LANs.

possible for a single IEEE 802.11 WLAN to comprise equipment from different vendors. The components are:

Stations are computing devices with wireless network interfaces. Typically, the stations are battery-powered laptop and handheld computers, but they could also be fixed workstations.

Access points perform the wireless-to-wired bridging function, although they have a number of other functions.

Distribution system (DS) is the mechanism by which APs exchange frames with one another and with wired networks, if any. DS is not necessarily a network, and 802.11 does not specify any particular technology for the DS. In nearly all commercial products, wired Ethernet is used as the backbone network technology.

Wireless medium is used for signal transmission. Several physical layers are defined and the architecture allows multiple physical layers to be developed to support the 802.11 MAC.

The basic building block of an IEEE 802.11 network is the *basic service set* (BSS), which is simply a set of stations that communicate with one another. A BSS does not generally refer to a particular area, due to the uncertainties of electromagnetic propagation. There are two types of BSS, as shown in Figure 4-2. When all of the stations in the BSS are mobile stations and there is no connection to a wired network, the BSS is called an *independent BSS* (IBSS). The IBSS is the entire network and only those stations communicating with each other in the IBSS are part of the LAN. This type of networks is called *ad hoc networks*.

When all of the mobile stations in the BSS communicate with the AP, the BSS is called an *infrastructure BSS* (never called an IBSS). The AP provides both the connection to the wired LAN, if any, and the local relay function for the BSS. Thus, if one mobile station in the BSS must communicate with another mobile station, the packet is sent first to the AP and then from the AP to the other mobile station. This causes communications to consume more transmission capacity

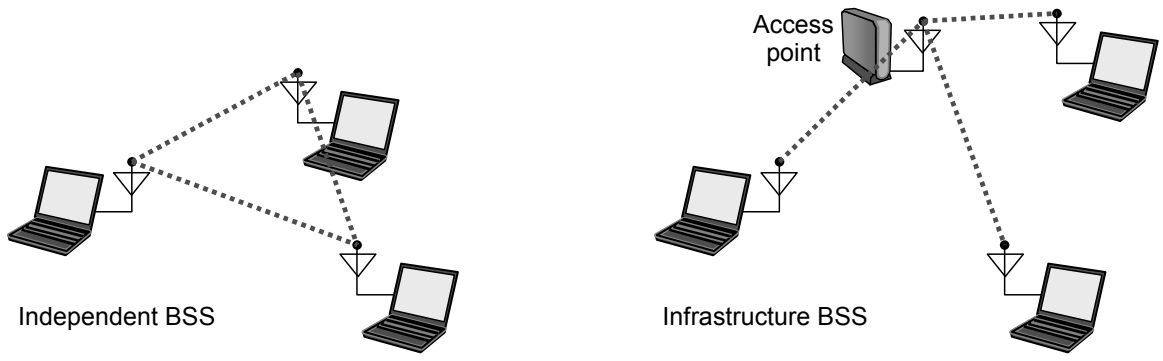


Figure 4-2. Independent and infrastructure basic service sets (BSSs).

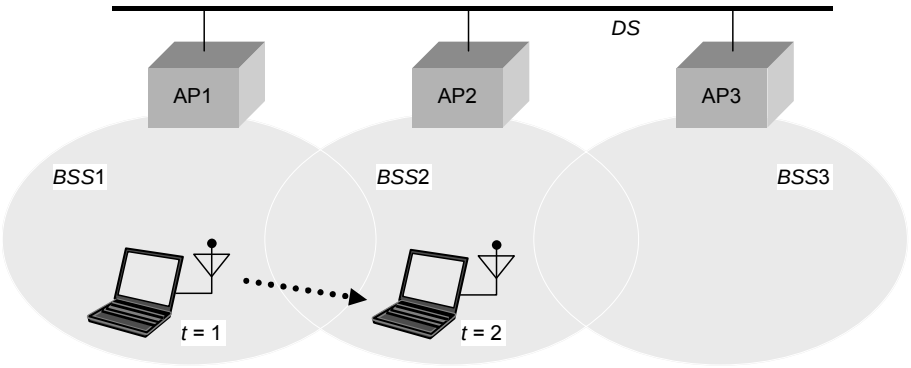


Figure 4-3. Extended service set (ESS) and mobility support.

than in the case where the communications are directly between the source and the destination (as in the IBSS). However, in many cases the benefits provided by the AP outweigh the drawbacks. One of the benefits provided by the AP is that the AP can assist the mobile stations in saving battery power. The mobile stations can operate at a lower power, just to reach the AP. Also, the AP can buffer the packets for a mobile station, when the station enters a power saving mode.

Extended service set (ESS) extends the range of mobility from a single BSS to any arbitrary range, see Figure 4-3. An ESS is a set of infrastructure BSSs, where the APs communicate among themselves to forward traffic from one BSS to another and to facilitate the roaming of mobile stations between the BSSs. The APs perform this communication via the DS. The stations in an ESS see the wireless medium as a single OSI layer-2 connection. ESS is the highest-level abstraction supported by 802.11 networks. Roaming between different ESSs is not supported by IEEE 802.11 and must be supported by a higher-level protocol, e.g., Mobile IP [Perkins 1998].

In terms of network services, 802.11 was intentionally designed to be just another link layer to higher layer protocols. From the user's perspective, 802.11 appears identical to Ethernet; network administrator has many more parameters to set. There are nine services defined by the 802.11 architecture. These services are divided into two groups, station services and distribution services. Table 4-1 summarizes the services. The *station services* are implemented in every 802.11 station, including AP stations. The *distribution services* are provided between BSSs; these services may be implemented in an AP or in another special-purpose device attached to the distribution system.

Table 4-1: IEEE 802.11 network services.

Service	Provider	Description
Distribution	Distribution	Service used by stations to exchange MAC frames when the frame must traverse the DS to get from a station in one BSS to a station in another BSS.
Integration	Distribution	Frame delivery to an IEEE 802 LAN outside the wireless network.
Association	Distribution	Used to establish a logical connection between a mobile station and an AP. This connection is necessary in order for the DS to know where and how to deliver data to the mobile station.
Reassociation	Distribution	Enables an established association to be transferred from one AP to another, allowing a mobile station to move from one BSS to another.
Disassociation	Distribution	Removes the wireless station from the network.
Authentication	Station	Establishes identity prior to establishing association.
Deauthentication	Station	Used to terminate authentication, and by extension, association.

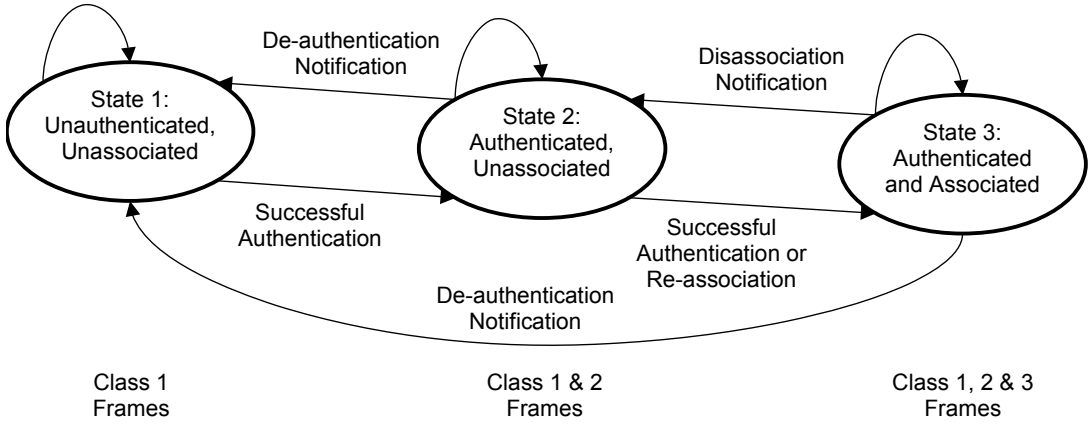


Figure 4-4. Relationship between state variables and services.

Privacy	Station	Provides protection against eavesdropping.
MSDU delivery	Station	Delivers data to the recipient.

MSDU stands for MAC Service Data Unit, which is the block of data passed down from the higher layer to the MAC layer. If the MSDU is too large to be transmitted in a single MAC frame, it may be fragmented and transmitted in a series of MAC frames.

The IEEE 802.11 standard requires that each station maintain two variables that depend on the authentication/deauthentication services and association/reassociation/disassociation services. The two variables are authentication state and association state, see Figure 4-4. The variables are used in a simple state machine that determines the order in which certain services must be invoked and when a station may begin using the data delivery service. The figure also indicates the types of frames that a station is allowed to use in a particular state. The frames of type 1 allow the station to find an IEEE 802.11 WLAN, an ESS, and its APs, to complete the required frame handshake protocols, and to implement the authentication service. In state 2, the additional frame types enable a station in state 2 to implement the association/reassociation/disassociation services. In state 3, all frame types are allowed and the station may use the data delivery service.

When a station wants to access an existing BSS (either after power-up, sleep mode, or just entering the BSS area), the station needs to get the synchronization information from the AP (or from the other stations when in ad-hoc mode). The station can get this information by one of two means:

1. Passive scanning: In this case the station just waits to receive a Beacon Frame from the AP. The beacon frame with synchronization information is sent periodically by the AP.
2. Active scanning: In this case the station tries to find an AP by transmitting Probe Request Frames, and waiting for Probe Response from the AP.

Both methods are valid and either one can be chosen according to the power consumption vs. performance tradeoff. Once the station has found an AP and decided to join its BSS, it goes through the authentication process, where each side proves the knowledge of a given password. When the station is authenticated, it starts the association process which allows it to exchange data frames with the AP.

4.1.2 Beacons

Beacon frames carry information about the BSS to allow a new station to find out everything it needs to match parameters with the BSS and begin communications. Each beacon frame is several slots long. Every station must listen to beacons. In infrastructure BSS, AP sends beacons. In IBSS, every station contends for beacon generation in the beacon window doing as follows:

- Determine a random number *k*;
- Wait for exactly *k* idle slots to pass;
- Transmit a beacon (if no one else has done so).

Beacons are also used for timing synchronization. Time synchronization is needed for frequency hopping and power management. 802.11 timers (clocks) are 64 bits, ticking in microseconds. The accuracy is within ±0.01%, or ±100 ppm. Δ = max tolerable difference between clocks. 802.11's time sync function operates as follows. Each beacon contains a timestamp. On receiving a beacon, STA adopts beacon's timing if T(beacon) > T(STA). The clocks are adjusted only forward, never to a past time.

4.2 Medium Access Control

The basic MAC protocol for IEEE 802.11 is carrier sense multiple access with collision avoidance (CSMA/CA). In this case, the coordination decision is distributed over all the stations. In addition to this, the standard specifies a centralized MAC algorithm, called the *point coordination function* (PCF), which is built on top of the distributed algorithm. PCF provides contention-free service and is primarily intended for configurations in which a number of wireless stations are interconnected with each other and some sort of base station that attaches to a wired backbone. This is especially useful if some of the data is time sensitive or high priority.

Distributed Coordination Function (DCF):

- CSMA/CA used for access control
- When a collision is detected, a random backoff timer is set
- Positive acknowledgements are used (to be explained below)

Point Coordination Function (PCF):

- Built on top of DCF: the point coordinator uses DCF to seize the medium
- A point coordinator determines which station currently has the right to transmit by polling

A wireless network may be configured to use both DCF and PCF simultaneously. For example, a number of stations with time-sensitive traffic are controlled by the point coordinator while remaining stations contends for access using DCF. When the coordinator seizes the medium, it holds it for a certain amount of time by using shorter interframe spaces than those used by DCF.

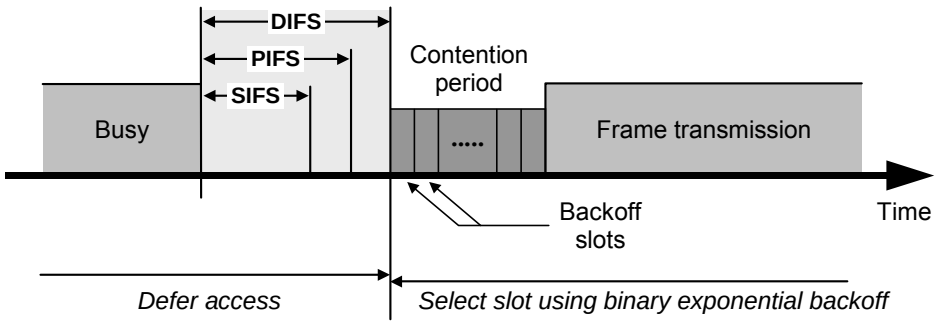


Figure 4-5: IEEE 802.11 interframe spacing relationships. Different length IFSs are used by different priority stations.

Some drawbacks of PCF are:

- Introduces considerable additional complexity
- Needs to choose a coordinator to do the polling (a leader election problem)
- Needs to provide for failure of the coordinator
- A station cannot access the medium unless explicitly polled from the coordinator
- All stations need to hear the poll
- The number of stations in a BSS can scale with no penalty in the performance. However, different BSSs need to be carefully planned not to overlap even when the density of the BSSs is very small. Therefore, it requires coordination between the point coordination functions within peer APs.

4.2.1 Interframe Spaces

Before beginning transmission, a station should determine that the medium is not already carrying a transmission. Interframe space (IFS) is the time interval between frames required for carrier sense mechanism to determine that the medium is idle and available for transmission. The transmission can begin once the IFS has elapsed. To assist with interoperability between different data rates, the interframe space is a fixed amount of time, independent of the physical layer bit rate. There are basic intervals determined by the PHY: the *short interframe space* (SIFS) and the *slot time*, which is slightly longer than SIFS. The slot time is equal to the parameter β, so that a station is always capable of determining if other station has accessed the medium at the beginning of the previous slot. This reduces the collision probability by half. [Slot length is medium dependent; higher-speed physical layers use shorter slots.]

The four different types of IFSs in 802.11 are (see Figure 4-5):

SIFS: Short interframe space is used for the highest priority transmissions, such as control frames, or to separate transmissions belonging to a single dialog (e.g. Frame-fragment–ACK). This value is a fixed value per PHY and is calculated in such a way that the transmitting station will be able to switch back to receive mode and be capable of decoding the incoming packet. For example, for the 802.11 FH PHY this value is set to 28 microseconds.

PIFS: PCF (or priority) interframe space is used by the PCF during contention-free operation. The coordinator uses PIFS when issuing polls and the polled station may transmit after the SIFS has elapsed and preempt any contention-based traffic. PIFS is equal to SIFS plus one slot time.

DIFS: DCF (or distributed) interframe space is the minimum medium idle time for asynchronous frames contending for access. Stations may have immediate access to the medium if it has been free for a period longer than the DIFS. DIFS is equal to SIFS plus two slot times.

EIFS: Extended interframe space (not illustrated in Figure 4-5) is much longer than any of the other intervals. It is used by any station that has received a frame containing errors that it could not understand. This station cannot detect the duration information and set its NAV for the Virtual Carrier Sense (defined below). EIFS ensures that the station is prevented from colliding with a future packet belonging to the current dialog. In other words, EIFS allows the ongoing exchanges to complete correctly before this station is allowed to transmit.

The idea behind different IFSs is to create different priority levels for different types of traffic. Then, high-priority traffic does not have to wait as long after the medium has become idle. If there is any high-priority traffic, it grabs the medium before lower-priority frames have a chance to try.

The values of some important 802.11b system parameters are shown in Table 4-2. The values shown are for the 1Mbps channel bit rate and some of them are different for other bit rates.

Table 4-2: IEEE 802.11b system parameters. (PHY preamble serves for the receiver to distinguish silence from transmission periods and detect the beginning of a new packet.)

Parameter	Value for 1Mbps channel bit rate
Slot time	20 μsec
SIFS	10 μsec
DIFS	50 μsec (DIFS = SIFS + 2 × Slot time)
EIFS	SIFS + PHY_preamble + PHY_header + ACK + DIFS = 364 μsec
CW_{min}	32
CW_{max}	1024
PHY_preamble	144 bits (144 μsec)
PHY_header	48 bits (48 μsec)
MAC data header	28 bytes = 224 bits
ACK	14 bytes + PHY_preamble + PHY_header = 304 bits (304 μsec)
RTS	20 bytes + PHY_preamble + PHY_header = 352 bits (352 μsec)
CTS	14 bytes + PHY_preamble + PHY_header = 304 bits (304 μsec)
MTU*	Adjustable, up to 2296 bytes

(*) The Maximum Transmission Unit (MTU) size specifies the maximum size of a physical packet created by a transmitting device.

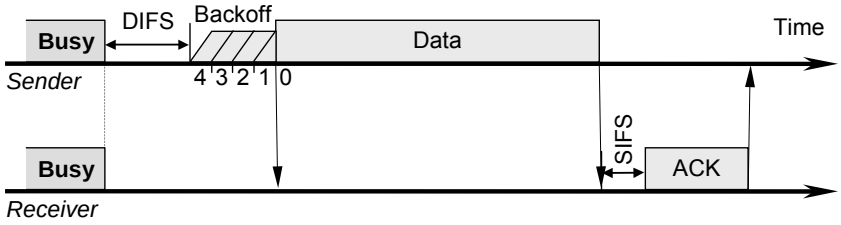


Figure 4-6: IEEE 802.11 basic transmission mode is a simple stop-and-wait ARQ. Notice the backoff slot countdown during the contention period.

4.2.2 Virtual Carrier Sensing and Network Allocation Vector

Carrier sensing is used to determine if the medium is available. In addition to the physical carrier sensing, 802.11 employs also *virtual carrier sensing* functions. If either one of them indicates that the medium is busy, the station defers the transmission.

It is difficult and expensive to build physical carrier sensing hardware for RF-based media, because transceivers can transmit and receive simultaneously only if they incorporate expensive electronics. Furthermore, unlike the wireline broadcast medium, the transitivity of connectivity does not apply. In wireline broadcast networks, such as Ethernet, if station A can hear station B and station B can hear station C, then station A can hear station C. This is not always the case in wireless broadcast networks, as will be explained below in Section 4.2.5 on hidden stations.

Virtual carrier sensing is provided by the *network allocation vector* (NAV). Most 802.11 frames carry a duration field, which can be used to reserve the medium for a fixed time period. Stations set the NAV to the time (in microseconds) for which they expect to use the medium, including any frames necessary to complete the current operation. Other stations that receive the frame set their *NAV timer*, which is a timer that indicates the amount of time the medium will be reserved, and count down from the NAV to 0. The NAV is kept current through duration values that are transmitted in all frames. When the NAV is nonzero, the virtual carrier sensing function indicates that the medium is busy; when the NAV reaches 0, the virtual carrier sensing function indicates that the medium is idle. By examining the NAV, a station may avoid transmitting even when the medium appears idle according to the physical carrier sense. The use of NAV will become apparent in the examples below.

4.2.3 ARQ and Atomic Operations

The 802.11 ARQ (automatic repeat request) retransmission strategy is essentially the simplest one: *stop-and-wait* (see Figure 4-6). This protocol does not initiate transmission of the next packet before ensuring that the current packet is correctly received. Notice that even the units of the atomic transmission (Data and ACKnowledgement) are separated by SIFS, which is intended to give the transmitting station a short break so it will be able to switch back to receive mode and be capable of decoding the incoming (in this case ACK) packet. The state diagram for a single packet transmission is shown in Figure 4-7.

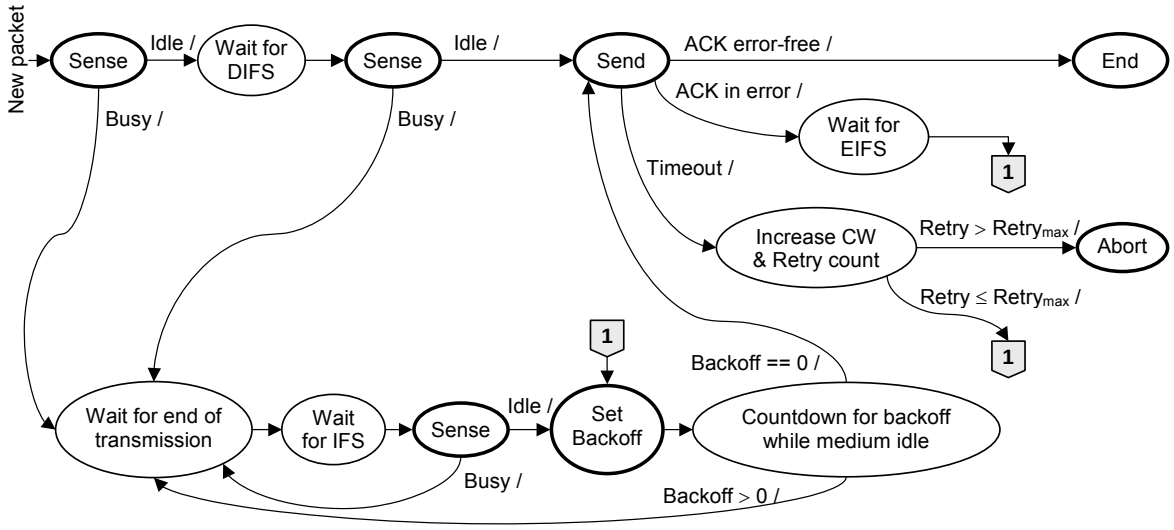


Figure 4-7: Sender's state diagram of basic packet transmission for 802.11 MAC protocol. In the state "Set Backoff," if the backoff counter is zero it is set to a number ∈ {0, …, *CW*–1}.

Wireless link is significantly more unreliable than a wired one. Noise, interference, and other propagation effects result in the loss of a significant number of frames. Even with error-correction codes, a number of MAC frames may not successfully be received. This situation can be dealt with by reliability mechanisms at a higher layer, such as TCP. However, timers used for retransmission at higher layers (which control paths comprising many links) are typically on the order of seconds. It is therefore more efficient to deal with errors at the MAC level. For this purpose, 802.11 includes a frame exchange protocol. When a station receives a data frame from another station it returns an acknowledgement (ACK) frame to the source station. This exchange is treated as an *atomic unit*, not to be interrupted by a transmission from any other station. If the source does not receive an ACK within a short period of time, either because its data frame was damaged or because the returning ACK was damaged, the source retransmits the frame.

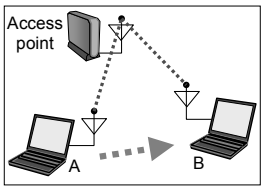
The basic data transfer mechanism in IEEE 802.11 involves an exchange of two frames. As we will see in Section 4.2.5 below, there could be other atomic exchanges which may involve more frames.

Atomic operations start like regular transmissions: they must wait for the DIFS before they can begin. However, the second and any subsequent steps in an atomic operation take place using the SIFS, rather than the DIFS. This means that the second (and subsequent) parts of an atomic operation will grab the medium before another type of frame can be transmitted. By using SIFS and the NAV, stations can seize the medium for as long as necessary.

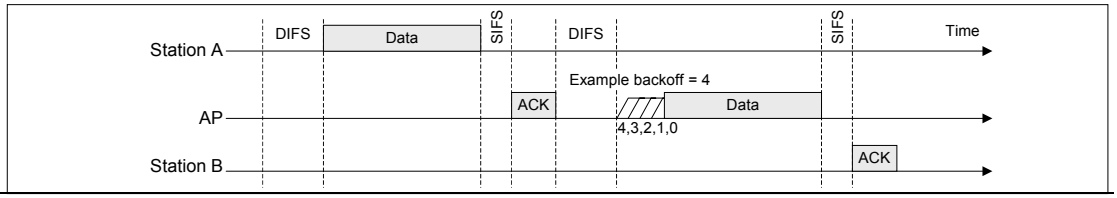
Stop-and-wait ARQ requires maintaining a timer: After the data packet is sent, the timer is set to limit the amount of time of waiting in case the data packet is lost to error or collision. If no acknowledgement is received within the timeout time, the sender retries the transmission, as indicated in Figure 4-7, but this time with a double contention window size, *CW*.

Example 4.1 Packet Transmission in Infrastructure BSS

Suppose we have an 802.11 Infrastructure BSS with an access point and two mobile STAs, *A* and *B*. Consider a scenario where *A* has a single packet to send to *B* and show precise time diagram for each of the three nodes from the start to the completion of the packet transmission. For each station, select a reasonable number of backoff slots (if any) to count down before the station commences its transmission. Assume that all nodes are using the basic transmission mode and the channel is error free.



Since this is an Infrastructure BSS, all stations must communicate through the AP. The time diagram is as follows:



The timing of successful frame transmissions is shown in Figure 4-8(a). If the channel is idle upon the packet arrival, the station transmits immediately, without backoff. Note that, in these figures, a crossed block represents a loss or erroneous reception of the corresponding frame. Figure 4-8(b) shows the case where an ACK frame is received in error, i.e., received with an incorrect frame check sequence (FCS). The transmitter re-contends for the medium to retransmit the frame after an EIFS interval. This is also indicated in the state diagram in Figure 4-7. On the other hand, if no ACK frame is received within a timeout interval, due possibly to an erroneous reception at the receiver of the preceding data frame, as shown in Figure 4-8(c), the transmitter contends again for the medium to retransmit the frame after an ACK timeout. (Notice that the ACK timeout is much shorter than the EIFS interval; in fact, ACK_timeout = $t_{SIFS} + t_{ACK} + t_{slot.}$)

Note that the 802.11 ARQ is stop-and-wait with a slight variation in that the number of retransmission attempts is limited. The reason for using the stop-and-wait ARQ is that the propagation constant β is very small, see Section 3.3. This is due to the short propagation time, which in turn is due to short distances (typically less than 500 m) at which 802.11 operates. Transmission time is the key contributor to the delay, rather than the propagation time. Thus, it is

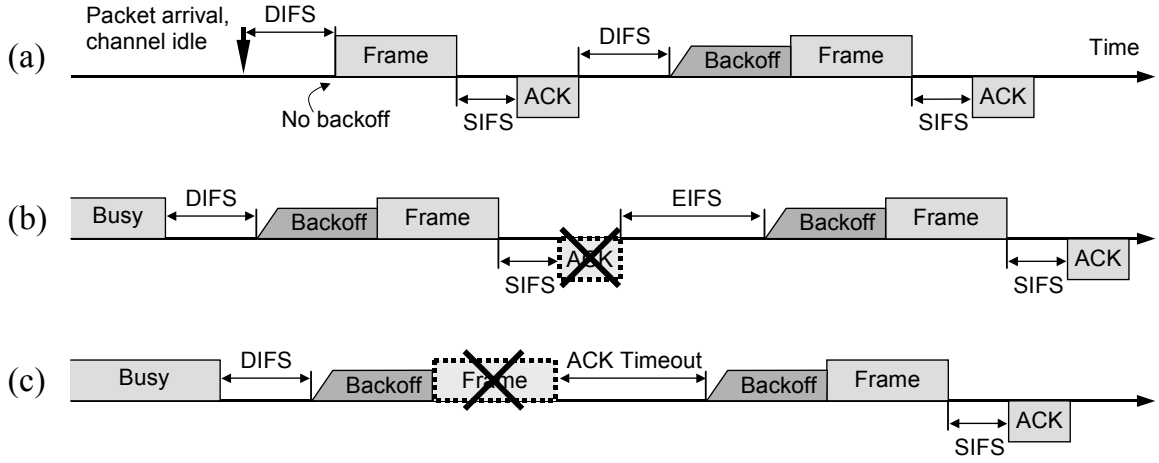


Figure 4-8: Timing diagrams. (a) Timing of successful frame transmissions under the DCF. (b) Frame retransmission due to ACK failure. (c) Frame retransmission due to an erroneous data frame reception.

more efficient to wait for an acknowledgement for each packet than to have to retransmit more than one packet should a go-back-*n* ARQ have been used.

Another variation from the ordinary stop-and-wait is that the standard allows the station to transmit to a different address between retransmissions of a given frame or fragment. This is particularly useful when an AP has several outstanding packets to different mobile stations and one of them does not respond.

Retransmissions with the DCF

Error recovery is the responsibility of the station sending the frame. Positive acknowledgements are the only indication of success. Senders expect acknowledgements for each transmitted frame and are responsible for retrying the transmission until it is successful. When a frame is not received or received in error, a retransmission occurs until the transmission is successful, or until the relevant limit is reached, whichever occurs first.

The IEEE 802.11 standard requires that a data frame is discarded by the transmitter's MAC after a certain number of unsuccessful transmission attempts. Each frame or frame fragment awaiting transmission has a single retry count associated with it, depending on the frame length. Frames that are shorter than the `dot11RTSThreshold` attribute (defined in Section 4.2.5 below) are considered to be short; frames longer than the threshold are long. Depending on the length of the frame, it is associated with a short (`dot11ShortRetryLimit`) or long retry counter (`dot11LongRetryLimit`). The frame's retry count begins at 0 and is incremented every time the frame retransmission fails. The default values of `dot11ShortRetryLimit` and `dot11LongRetryLimit` are 7 and 4, respectively.

Using the counters the MAC may determine that it is no longer worthwhile to continue attempting to transmit a particular frame. When the MAC makes that determination, it may cancel the frame's transmission and discard the frame. If a frame is cancelled, the MAC indicates this to the upper protocol layer, through the MAC service interface.

4.2.4 Backoff Procedure with the DCF

802.11 slightly modifies the CSMA/CA access deferral algorithm, Section 3.3. When the higher-layer protocol passes a packet to MAC for transmission, the station first senses the channel. If the medium is found idle, the station first waits for DIFS; if the medium remains idle, the station transmits immediately. Conversely, if the medium is at first found busy, the station's access deferral must include the *backoff procedure*. The station first waits for the medium to become idle for DIFS (if the last frame is received correctly) or EIFS (if the last frame is *not* received correctly). After this, it sets the backoff timer to a random integer number of slots that is drawn from a uniform distribution over [0, *CW*], where *CW* is the number of slots in the *contention window* or *backoff window*.

When several stations are attempting to transmit, the station that picks the lowest random number, i.e., the first slot, wins. IEEE 802.11 defines the retry counter limit, so the packet is discarded after a certain number ℓ of retransmission attempts. It is common to assume ℓ=7, as stipulated by the standard. The *backoff stage k* is the round of re-transmission $k \in \{0, 1, \ldots, \ell\}$, which determines the contention window size. Each time the retry counter *k* increases, the contention

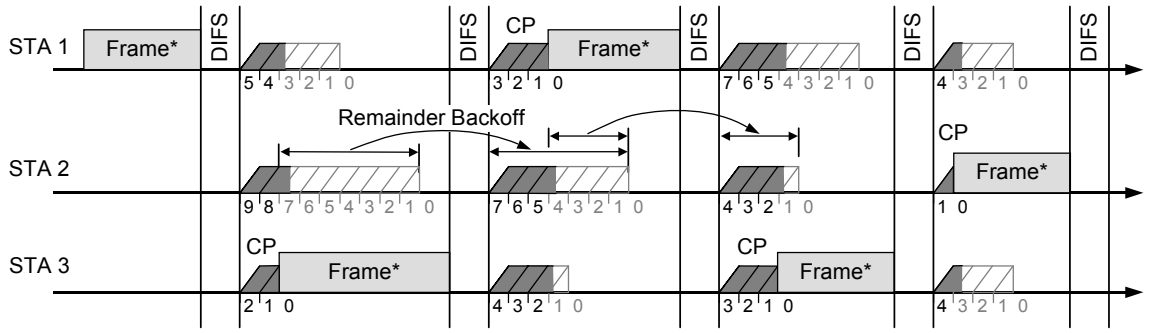


Figure 4-9: The backoff mechanism of the 802.11 MAC. The Frame* transmission time includes the RTS/CTS exchange and the MAC layer ACK. CP: Contention period.

window CW_k doubles its size between CW_{min} and CW_{max}. The maximum size of the contention window is determined by the physical layer. For example, for the DSSS physical layer CW_{min} = 32 and CW_{max} = 1024. When the contention window reaches its maximum size, it remains constant for the backoff stages $k > \log_2(CW_{max}/CW_{min}) = 5 \leq \ell$ until it can be reset. The contention window is reset to its minimum size when frames are transmitted successfully, or the associated retry counter is reached, and the frame is discarded.

Figure 4-9 illustrates the working of the backoff mechanism. With multiple stations contending for the channel, once the channel is sensed idle for a DIFS, each station with a packet to transmit decrements its backoff timer. The station whose timer expires first begins transmission and the remaining stations freeze their countdown and defer their transmission. Once the current station finishes transmission, the process repeats again. The stations that were preempted during the countdown start decrementing their timer from where they left off rather than selecting a new countdown period. In this way, the preempted stations have slight priority compared to others.

4.2.5 Hidden and Exposed Stations

IEEE 802.11 DCF protocol is based on the MACAW protocol for solving the problem of hidden and exposed stations, reviewed in Section 3.5. The key differences arise due to using a stop-and-wait ARQ and virtual carrier sensing.

This RTS/CTS exchange partially solves the hidden station problem but the exposed node problem remains unaddressed. The 4-way handshake of the RTS/CTS/DATA/ACK exchange of the 802.11 DCF protocol (Figure 4-10) requires that the roles of sender and receiver are interchanged several times between pairs of communicating nodes, so neighbors of both these nodes must remain silent *during the entire exchange*. This is achieved by invoking the virtual carrier sense mechanism of 802.11, i.e., by having the neighboring nodes set their Network Allocation Vector (NAV) values from the Duration field specified in either the RTS or CTS packets they hear. By using the NAV, the stations ensure that atomic operations are not interrupted. The NAV time duration is carried in the frame headers on the RTS, CTS, data and ACK frames.

It should be noted that virtual carrier sensing using NAV and hidden station solving using RTS/CTS are important improvements but not essential parts of the 802.11 MAC. The essential part is CSMA/CA. The 802.11 MAC can function without RTS/CTS. The only difference is that

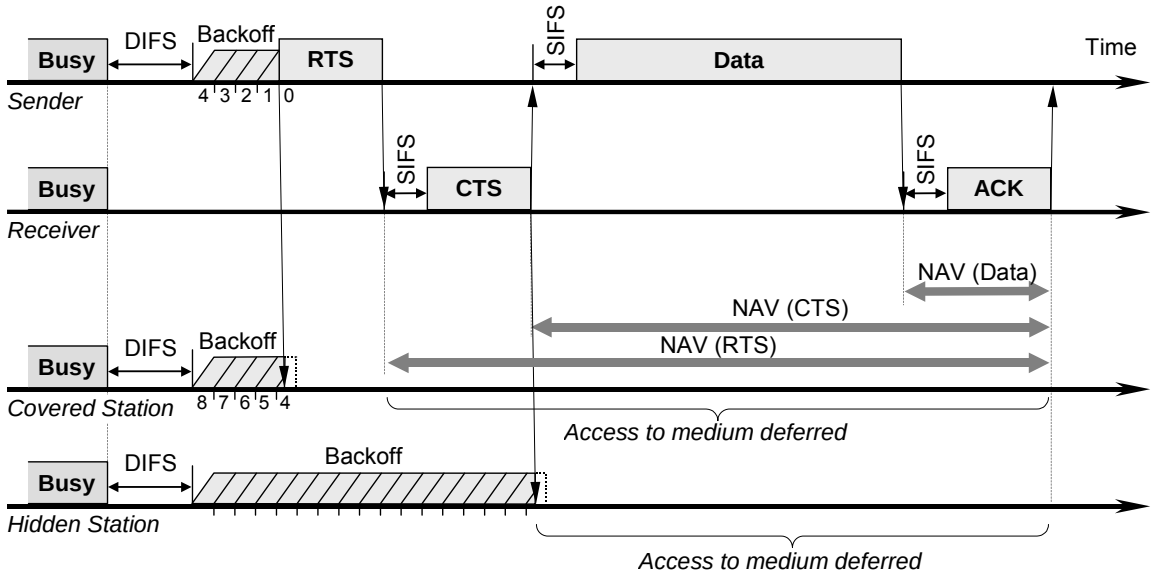


Figure 4-10. The 802.11 protocol atomic unit exchange in RTS/CTS transmission mode consists of four frames: RTS, CTS, Data, and ACK.

the efficiency is reduced, since the hidden stations cause collisions and the exposed stations defer transmissions. RTS/CTS is just an improvement, rather than an essential component. It is a mechanism to improve *efficiency*, in terms of the fraction of transmitted frames that escape collisions, but the price paid is in increased *delays*. If the expected traffic is relatively low, the performance of the protocol in terms of reducing the delays will be improved by turning off RTS/CTS.

The RTS/CTS mechanism is very efficient in terms of system performance, especially when large packets are considered, as reduces the number of the frames involved in the contention process. In fact, in the assumption of perfect channel sensing by every station, collision may occur only when two (or more) packets are transmitted within the same slot time. If both transmitting stations employ RTS/CTS, collision occurs only on the RTS frames, and it is early detected by the transmitting stations by the lack of CTS responses.

RTS/CTS can be *disabled* by an attribute in the management information base (MIB). The value of the `dot11RTSThreshold` attribute defines the length of a frame that is required to be preceded by the RTS and CTS frames. All frames of a length greater than the threshold will be sent with the RTS/CTS four-way frame exchange. Frames of a length less than or equal to the threshold will not be preceded by the RTS/CTS. This allows a network designer to tune the operation of the IEEE 802.11 WLAN for the particular environment in which it is deployed. If the number of transmissions is relatively low and all stations can hear each other, the RTS/CTS just consumes bandwidth and introduces delays for no measurable gain. In this case, the threshold should be set high to disable the RTS/CTS and this is the default setting. Conversely, if the number of transmissions is relatively high and not all stations can hear each other, the threshold may be set lower, causing long frames to use RTS/CTS before transmitting the data. A value to which the threshold should be set is arrived at by comparing the bandwidth lost to the additional overhead of the protocol to the bandwidth lost from transmissions being corrupted by hidden stations. A typical value for the threshold is 128.

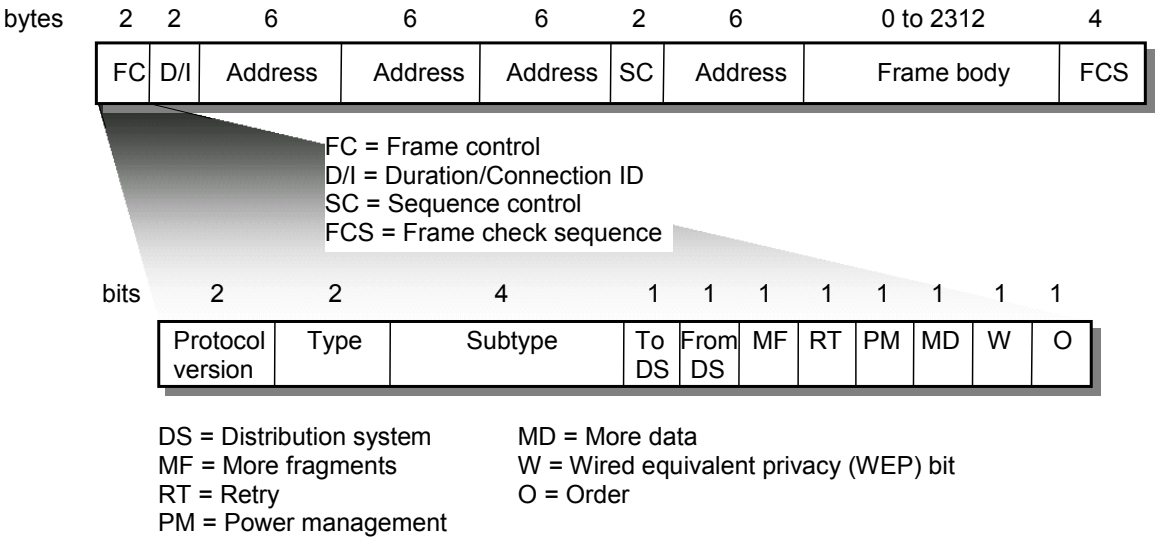


Figure 4-11. Generic IEEE 802.11 MAC frame format.

4.2.6 Frame Structure

Figure 4-11 shows the 802.11 frame format. This general format is used for all data and control frames, but not all fields are used in all types of frames. It is beyond the scope of this text to elaborate on all the 802.11 frame types. I just comment on several interesting aspects.

There can be up to four address fields in an 802.11 frame. When all four fields are present, the address types include source, destination, transmitting station, and receiving station. One of the fields could also be the BSS identifier, which is used in the probe request and response frames, used when mobile stations scan an area for existing 802.11 networks.

Duration/connection ID field is used as a duration field to set the NAV. It indicates the time (in microseconds) the channel will be allocated for successful transmission of a MAC frame. In some control frames, this field contains an association, or connection, identifier.

There are three groups of frame types:

2. *Control frames* assist in reliable delivery of data frames. There are six control frame subtypes: power-save poll (PS-Poll), request to send (RTS), clear to send (CTS), acknowledgement, contention-free (CF)-end, and CF-end + CF-ack.
3. *Data frames* carry data between the stations. There are eight data frame subtypes, organized in two groups: data, data + CF-ack, data + CF-poll, data + CF-ack + CF-poll. The remaining four subtypes do not carry any user data.
4. *Management frames* are used to manage communications between stations and APs. The following subtypes are included: association request, association response, reassociation request, reassociation response, probe request, probe response, beacon, announcement traffic indication message, dissociation, authentication, and deauthentication.

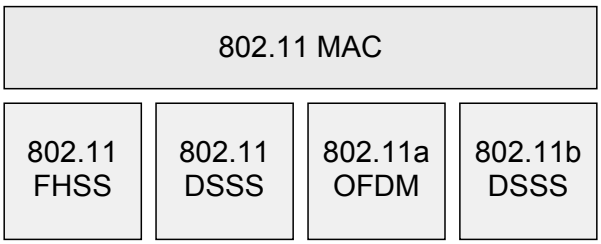


Figure 4-12. 802.11 protocol architecture—MAC layer comes atop different PHY layers.

4.3 Physical Layer and Rate Adaptation

There are several types of IEEE 802.11 physical layers, as illustrated in Figure 4-12.

4.3.1 Physical Signals

The wireless physical layer is split into two parts, called the PLCP (Physical Layer Convergence Protocol) and the PMD (Physical Medium Dependent) sublayer. The PMD takes care of the wireless encoding explained above. The PLCP presents a common interface for higher-level drivers to write to and provides carrier sense and CCA (*Clear Channel Assessment*), which is the signal that the MAC layer needs so it can determine whether the medium is currently in use (see Figure 4-13).

The PLCP consists of a 144-bit preamble that is used for synchronization to determine radio gain and to establish CCA. The preamble comprises 128 bits of synchronization (scrambled 1 bits), followed by a 16-bit field consisting of the pattern 1111001110100000. This sequence is used to mark the start of every frame and is called the SFD (Start Frame Delimiter).

The next 48 bits are collectively known as the PLCP header. The header contains four fields: signal, service, length and HEC (header error check). The signal field indicates how fast the payload will be transmitted (1, 2, 5.5 or 11 Mbps). The service field is reserved for future use. The length field indicates the length of the ensuing payload, and the HEC is a 16-bit CRC of the 48-bit header.

To further complicate the issue (and degrade performance) in a wireless environment, the PLCP is always transmitted at 1 Mbps. Thus, 24 bytes of each packet are sent at 1 Mbps. The PLCP

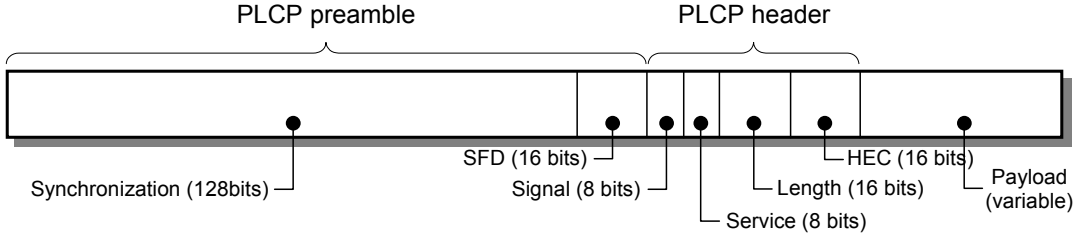


Figure 4-13: IEEE 802.11 PHY frame using DSSS.

introduces 24 bytes of overhead into each wireless Ethernet packet before we even start talking about where the packet is going. Ethernet introduces only 8 bytes of data. Because the 192-bit header payload is transmitted at 1 Mbps, 802.11b is at best only 85 % efficient at the physical layer. In other words, of the raw 11 Mbps channel capacity, the useful payload is at most at $0.85 \times 11 = 9.35$ Mbps.

4.3.2 Transmission Rate Adaptation

It should be clear from Chapter 2 that wireless channel conditions are time-varying and received signal-to-noise ratio changes dynamically. Multi-rate radios are capable of transmitting at several rates, using different modulation schemes. IEEE 802.11 physical layers (PHYs) support multiple transmission rates. The wireless radio generates a 2.4-GHz carrier wave (2.4 to 2.483 GHz) and modulates that wave using a variety of techniques.

IEEE 802.11 PHY automatically chooses the modulation scheme as a function of channel conditions. Different modulation schemes provide a trade-off between throughput and coverage range. For 1-Mbps transmission, BPSK is used (one phase shift for each bit). To accomplish 2-Mbps transmission, QPSK is used. Remember that QPSK encodes 2 bits of information in the same space as BPSK encodes 1. The trade-off is that you must increase power or decrease range to maintain signal quality. Because the FCC regulates output power of portable radios to 1 watt EIRP (equivalent isotropically radiated power), range is the only remaining factor that can change. Thus, on 802.11 devices, as you move away from the radio, the radio adapts and uses a less complex (and slower) encoding mechanism to send data.

To provide the higher PHY rates of 5.5 and 11 Mbps, the IEEE 802.11b defines a Complementary Code Keying (CCK) modulation scheme. CCK is a variation on M-ary Orthogonal Keying Modulation that uses I/Q modulation architecture with complex symbol structures. It is based on a complex set of 64 eight-bit Walsh/Hadamard functions known as Complementary Codes. For the 5.5 Mbps rate, 4 bits are encoded per word, while for the 11 Mbps rate, 8 bits are encoded per word. Both PHY rates use QPSK as the modulation technique and signal at 1.375 MSps. The spreading maintains the same chipping rate and spectrum shape as the original 802.11 DSSS, hence, occupying the same channel bandwidth.

The PHY rate to be used for a particular frame transmission is solely determined by the transmitting station. The transmission rate should be chosen in an adaptive manner since the wireless channel condition varies over time due to such factors as station mobility, time-varying interference, and location-dependent errors.

For example, typical ranges of IEEE 802.11b are:

- 30-50 m at 11 Mbps
- 40-60 m at 5.5 Mbps
- 80-120 m at 2 Mbps

Figure 4-14(a) shows the Bit Error Rate (BER) curves vs. SNR provided by the chipset manufacturer Intersil for the IEEE 802.11b PHY modes [Intersil 2000]. [The SNR is measured at the antenna of the receiver, before decoding the spread signal.] Compare these curves to those derived theoretically (Figure 2-9).

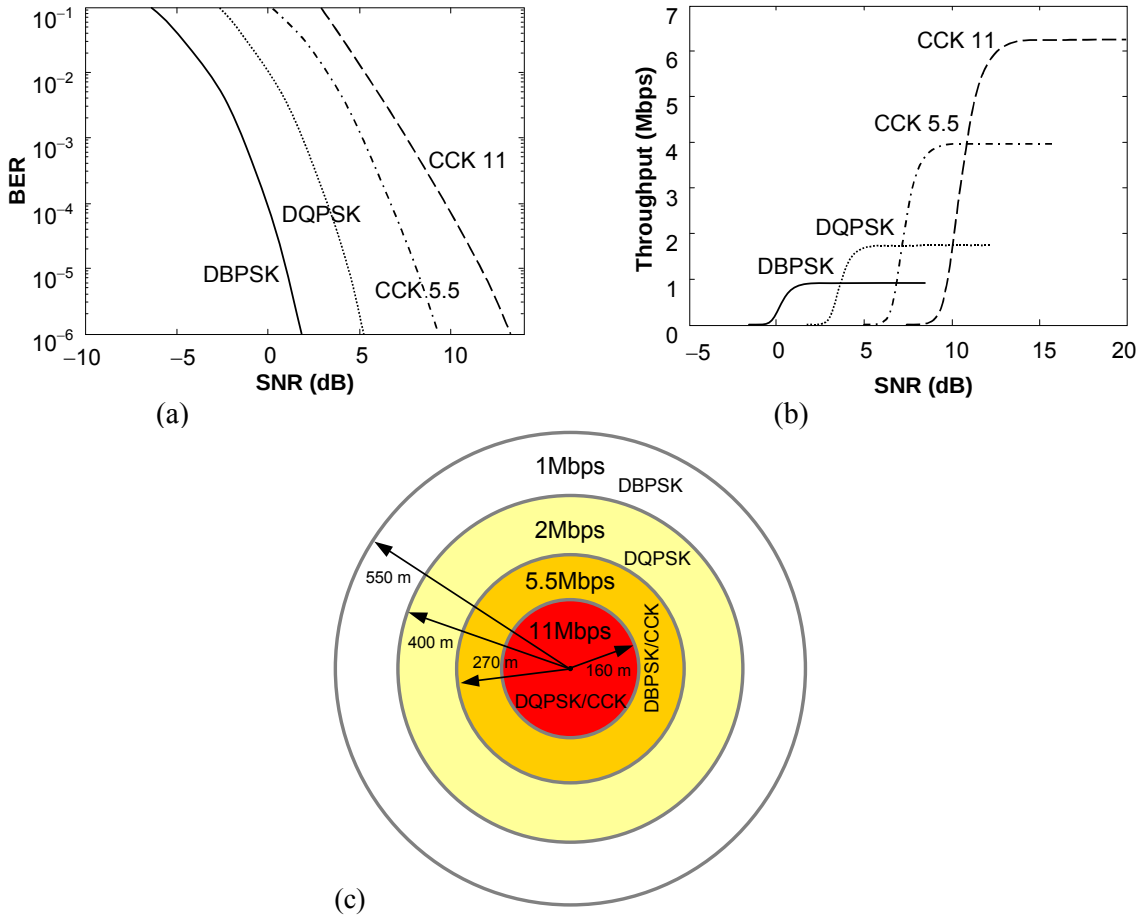


Figure 4-14: (a) Empirical BER vs. SNR curves provided by Intersil for its HFA3861B chip. (b) Simulated throughput vs. SNR. (c) The used modulation scheme depends on the node's distance to the access point, in the center (open space or outdoors).

The throughput of an erroneous channel can be computed as in Section 4.5 below. Instead of the probability of collision, we substitute the packet error probability, Eq. (2.10).

Simulations obtain a throughput as a function of the SNR, for every PHY mode, as shown in Figure 4-14(b). One can imagine that this ideal performance can be achieved only if the SNR at the receiver is known to the transmitting STA in advance.

Figure 4-14(c) shows how available link capacity may vary depending on the distance a user is from an access point (based on typical distances measured at the Lucent Technologies ORiNICO Laboratories for environments where there are no obstructions between antennas [ORiNICO 802.11b PC Card Manual, Lucent Technologies, August 2000.]). In fact, the variance in nominal link capacity is likely to be even greater for the 802.11a (6Mbps-54Mbps) and 802.11g (> 11Mbps) versions of the Wi-Fi standard.

Assuming nodes are randomly situated about an access point in accordance with a 2-dimensional uniform distribution, there will typically be a large fraction of users that are in a region of

coverage that provides a low transmission data rate. This results in an inefficient use of scarce spectrum resources.

If physical layer chooses the modulation scheme transparent to MAC, then MAC cannot know the time duration required for the transfer. Thus, PHY must involve MAC protocol in deciding the modulation scheme. Some implementations use a sender-based scheme for this purpose. [Kamerman & Monteban 1997] uses sender-based "autorate fallback." Probing mechanisms. Sender decreases bit rate after X consecutive transmission attempts fail. Sender increases bit rate after Y consecutive transmission attempt succeed. Advantage: Can be implemented at the sender, without making any changes to the 802.11 standard specification. Disadvantage: Probing mechanism does not accurately detect channel state; Channel state detected more accurately at the receiver; Performance can suffer since the sender will periodically try to send at a rate higher than optimal. Also, when channel conditions improve, the rate is not increased immediately.

Receiver-based schemes can perform better, e.g., receiver-based autorate MAC protocol [Holland *et al*. 2001], see example in Figure 4-15. Sender sends RTS containing its best rate estimate. Receiver chooses best rate for the conditions and sends it in the CTS. Sender transmits DATA packet at the new rate. Information in data packet header implicitly updates the NAV vector in nodes that heard the old rate (node *A* in Figure 4-15).

Review also [del Prado & Choi 2002].

4.4 Power-saving Mechanisms

The users of mobile devices are frequently aware of time limits on the device use imposed by the size of batteries. A major part of the battery power is consumed during transmission of signals. To extend battery life, mobile devices employ power-saving mechanisms which make the mobile station enter a suspended or semisuspended mode of operation with limited capabilities. This is, however, done in cooperation with the network, so as to *not* disrupt normal communications or provide the user with a perception of such a disruption even if there was one.

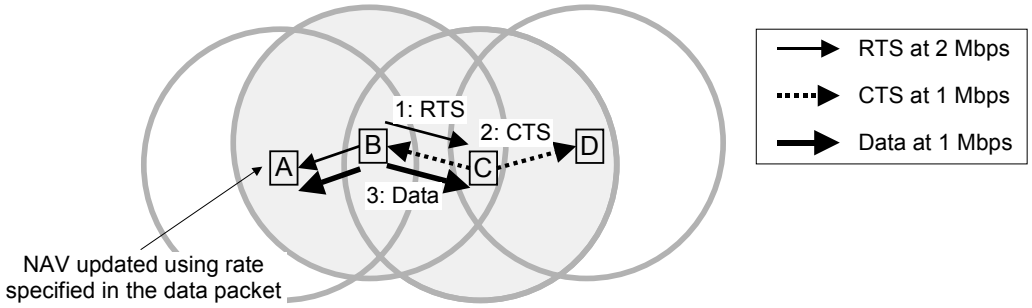


Figure 4-15. Example for receiver-based autorate MAC protocol. Node B sends data to C, but nodes A and D overhear their communication due to wireless broadcast.

This type of power control is usually not used at cellular phones for voice communication. Discontinuous is good for data traffic, since voice traffic has stringent QoS (latency, jitter) requirements.

Here I cover some aspects of power saving specific to the IEEE 802.11 wireless local area networks.

Motivation: battery energy drained at a slower rate if used in an ON/OFF mode than in a continuous ON mode, for the same total duration of ON mode.

Main idea: convert the data traffic to ON-OFF traffic.

How long can we set the entire length of ON-OFF duration?

By default, wireless LANs use CAM (Constant Access Mode) to constantly listen to the network and get the data they need. When power utilization is an issue, however, the workstations and access points can be configured for PAM (Polled Access Mode). With this, the clients on the network wake up on a regular period and listen for a special packet called a TIM (traffic information map) from the access point. In between TIMs, the client shuts off its radio and thus conserves power. All the devices on the network share the same wake-up period, as they must all wake up at exactly the same time to hear the TIM from the access point.

The TIM informs certain clients that they have data waiting at the access point. A client card stays awake when the TIM indicates it has messages buffered at the access point until those messages are transferred, and then the card goes to sleep again. The access point buffers the data for each card until it receives a poll request from the destination station. Once the data is exchanged, the station goes back into power-saving mode until the next TIM is transmitted. In our lab tests, we've found that PAM mode can save power by as much as 1,000 percent, depending on the volume of traffic on your network.

The access point indicates the presence of broadcast traffic with a DTIM (delivery traffic information map) packet. The DTIM timer is always a multiple of the TIM timer and is often adjustable at the access point. Setting this value high cuts down on the amount of time the station must stay awake checking for broadcast traffic. However, a higher DTIM timer means that the radio will stay on longer to receive DTIM traffic when it does come up in the time cycle.

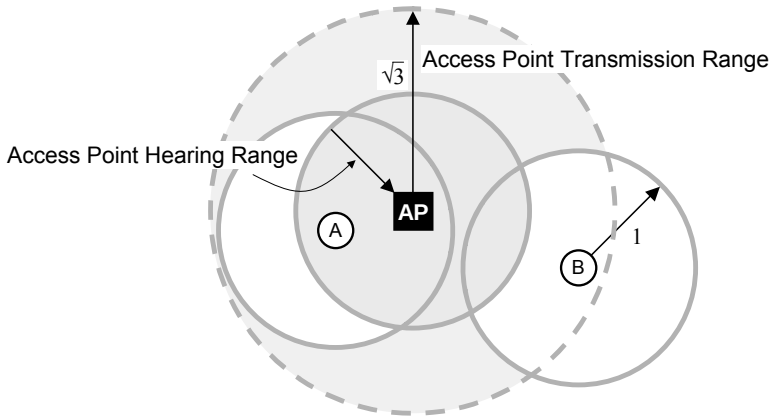


Figure 4-16. Asymmetry between the transmission ranges. The 3:1 ratio between the transmit power of an AP and a STA translates to a roughly $\sqrt{3}$ longer transmission range.

What 802.11 specifies is tools to control ON-OFF periods, but does not specify the actual policy. That is part of a different 802.11 standard. Application knows best what kind of traffic it works with and during what periods it can enter power saving mode.

How do the stations negotiate the PS strategy before the start?

What is the length of the IFS period after TIM, before PS-Poll?

What is the frequency of beacons?

TIM has priority to DTIM, but what if we want to broadcast something at high priority?

Explain that in CW, after the station with lowest random number transmits, other competing stations are shut right after one slot.

Does PS-Poll contain sequence number?

Does data frame contain sequence number?

In infrastructureless, the transmitter sends ATIM and does not wait for anything (?), just sends the data frame or RTS frame.

See also O'Hara book on how association is liked with PS mode.

4.4.1 Effect on Capacity

Tampering with the battery power to save the energy has negative impact on the link transmission capacity. How different power saving strategies affect channel capacity?

It is interesting to notice that there is a 3:1 ratio between the transmit power of an AP and a STA, although the transmission powers of some commercially available sets can be adjusted. There is an additional 3:1 ratio or more when external antenna is used at the AP. This is illustrated in Figure 4-16. Although the AP can hear roughly $\sqrt{3}$ more distant stations than a STA can, the

STA cannot talk back (as shown for STA B in Figure 4-16), so this does not increase the effective communication range. One justification for this is an expected asymmetry of upstream/downstream traffic: a 50:1 ratio is expected between the amount of traffic downloaded vs. the amount of traffic uploaded. Since higher power translates to a higher bit rate, this accommodates the traffic asymmetry.

4.5 Performance Analysis

Let us assume that there are a total of m nodes of which n are backlogged. We assume that all nodes with a packet to transmit have the same priority to use the channel. The analysis will be performed for the steady state—we assume that the average number of backlogged nodes remains constant. Although at one time a node may enter an empty state from backlogged state, there will be another node transitioning in the opposite way, so that the system is maintained in a balanced state, with constant $E[n] > 0$.

As a consequence of non-zero n, the channel is working in a saturated traffic condition. That is, since at anytime there is at least one backlogged node, the channel is always "busy"—from the channel point of view packets always arrives regardless from which node. (Note that saturated channel does not imply saturated stations.)

In the analysis below I consider neither the initial queuing at the node nor the arrival pattern for arrivals to the channel (i.e., application traffic profile). In other words, a newly arriving packet always finds an empty queue at its own node. This can be justified by assuming real-time traffic with bounded delay per packet, so the stale packets get purged from the queue. This prevents queue buildup and so n is not function of the collision probability.

We assume that transmission is error-free (noise-free channel) and consider only packet collisions. The analysis assumes the RTS/CTS transmission mode, but can be easily modified for the basic transmission mode. I first consider the wireless LAN case where all stations are within each other's transmission range; then I extend the analysis to include hidden stations. An example transmission sequence is shown in Figure 4-17.

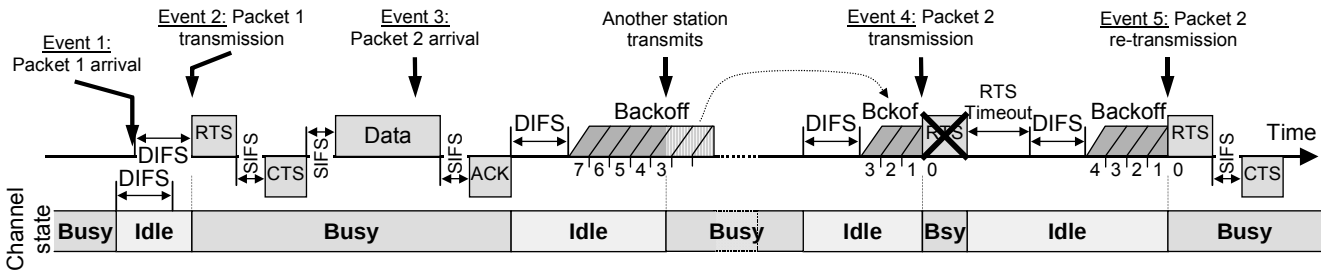


Figure 4-17. Illustration of the node events and the corresponding channel states. Note that the node is preempted during the second contention period. If a packet arrives to an idle channel (as is the case for packet 1), it waits only DIFS time, i.e., there is no backoff period.

4.5.1 Channel Event Probabilities

Assume a unit time during which a station can transmit at most one packet. Our system is (approximately) memoryless in terms of offered traffic load—the fact that the previous time slot was idle does not change the probability that the next slot will be idle. We can consider the system as consisting of n independent Bernoulli trials. The probability that i out of n backlogged nodes transmit a frame with probability q_u during the unit period u is:

$$Q_u(i, n) = \binom{n}{i} \cdot q_u^i \cdot (1 - q_u)^{n-i} \tag{4.1}$$

For a backlogged station counting down, the time units are contention slots t_{slot}. Then, for a station counting down the probability of an idle slot is the conditional probability that no station at all transmits given that the station itself does not transmit:

$$p_{idle} = \frac{Q_c(0, n)}{(1 - q_c)} = \frac{\binom{n}{0} \cdot (1 - q_c)^n}{(1 - q_c)} = (1 - q_c)^{n-1} \tag{4.2a}$$

The probability of a successful transmission during the preemption in a contention slot is the conditional probability that a single station other than the one considered transmits:

$$p_{succ} = \frac{(1 - q_c) \cdot Q_c(1, n)}{(1 - q_c)} = (n - 1) \cdot q_c \cdot (1 - q_c)^{n-2} \tag{4.2b}$$

And, the probability of a collision during the preemption in a contention slot is:

$$p_{coll} = 1 - p_{idle} - p_{succ} \tag{4.2c}$$

Given that a node transmitted, the transmission will result in one of two events: success or collision. The conditional probability of a successful transmission given that the node transmitted is the probability that no other node transmits during its vulnerable period (Figure 4-18):

$$\tilde{p}_{succ} = (1 - q_v)^{n-1} \tag{4.3a}$$

For the case when the vulnerable period $t_{vp} = t_{slot}$ (no hidden stations), the transmission probability $q_v = q_c$. The probability of a collision in the vulnerable period is:

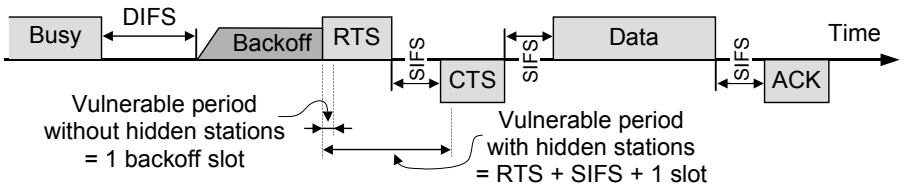


Figure 4-18: Definition of the vulnerable period for the cases with and without hidden stations.

$$\widetilde{p}_{coll} = 1 - \widetilde{p}_{succ} \tag{4.3b}$$

So far this analysis is similar to that in Sections 3.2 and 3.3. However, the probability q_r that a node transmits, Eq. (3.4b), is an independent variable that is controlled to achieve desired performance. Conversely, the probability q_c here is a dependent variable that is derived from other protocol parameters as follows.

The node's state is defined by two random variables: the backoff stage K, which takes values $\{0, \ldots, \ell\}$, and the channel contention counter C, which for $K=k$ takes values $\{0, \ldots, CW_k-1\}$, The state transition diagram is represented as the Markov chain model in Figure 4-19. We introduce the backoff stage $K = -1$ in order to account for the fact that the station does not enter backoff countdown if a new packet arrives during an idle channel state. In addition, we truncate the chain, so the packet is dropped after ℓ retransmissions. The transition probabilities $P\{K_{t+1}=k, C_{t+1}=i \mid K_t=l, C_t=j\}$ are as follows:

$$P\{-1, 0 \mid k, 0\} = \widetilde{p}_{succ} \cdot p_{idle}, \qquad -1 \le k \le \ell-1 \tag{4.4a}$$

$$P\{-1, 0 \mid \ell, 0\} = p_{idle} \tag{4.4b}$$

$$P\{0, i \mid -1, 0\} = \frac{\widetilde{p}_{succ} \cdot p_{busy} + \widetilde{p}_{coll}}{CW_0}, \qquad 0 \le i \le CW_0 - 1 \tag{4.4c}$$

$$P\{k, i \mid k, i\} = p_{busy}, \qquad 0 \le k \le \ell \text{ and } 1 \le i \le CW_k - 1 \tag{4.4d}$$

$$P\{k, i \mid k, i+1\} = p_{idle}, \qquad 0 \le k \le \ell \text{ and } 1 \le i \le CW_k - 2 \tag{4.4e}$$

$$P\{0, i \mid k, 0\} = \frac{\widetilde{p}_{succ} \cdot p_{busy}}{CW_0}, \qquad 0 \le k \le \ell-1 \text{ and } 0 \le i \le CW_0 - 1 \tag{4.4f}$$

$$P\{0, i \mid \ell, 0\} = \frac{p_{busy}}{CW_0}, \quad 0 \le i \le CW_0 - 1 \tag{4.4g}$$

$$P\{k, i \mid k-1, 0\} = \frac{\widetilde{p}_{coll}}{CW_k}, \; 0 \le k \le \ell \text{ and } 0 \le i \le CW_k - 1 \tag{4.4h}$$

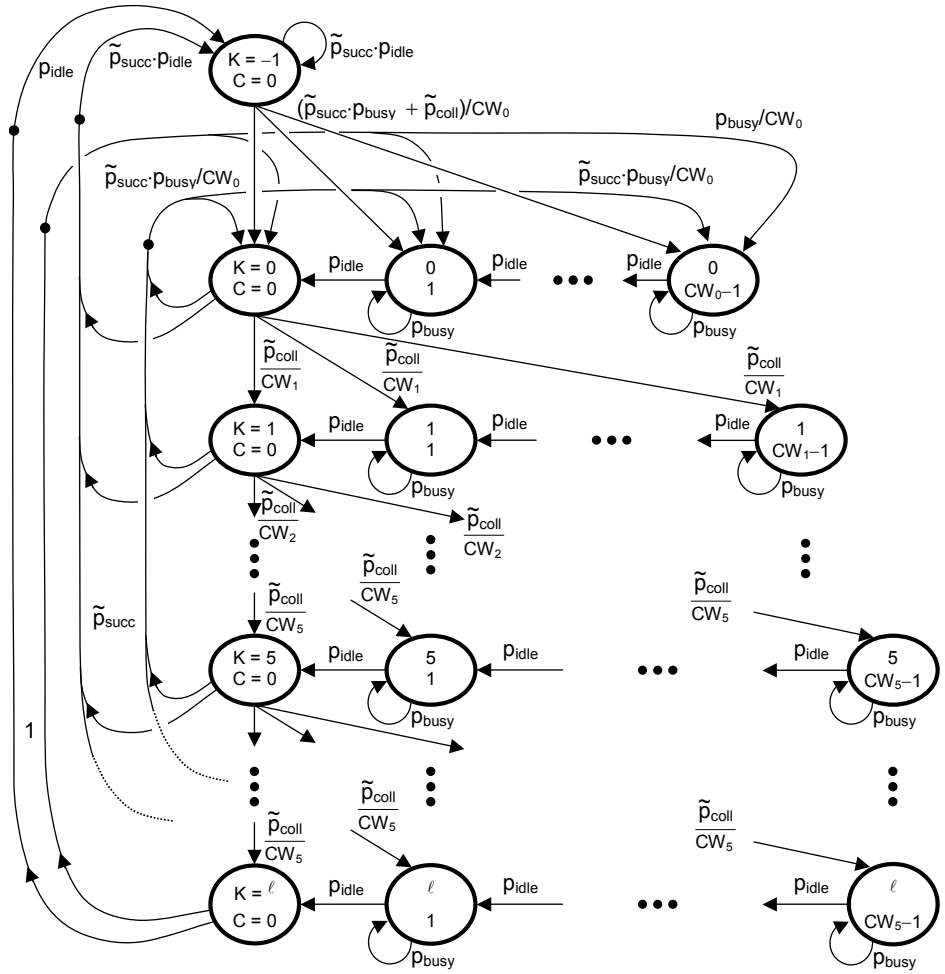


Figure 4-19. Markov chain for the derivation of the transmission probability q_c. Notice that unlike Figure 3-11 which counts the number of backlogged stations in the system, this Markov chain counts the number of collisions (vertical axis) and the number of remaining countdown slots (horizontal axis), both for a current packet.

For example, Eq. (4.4a) means that the node will transition to the state $\{-1, 0\}$ from any state $\{k, 0\}$ if the previous packet transmission is successful and the new packet arrives while the channel is idle. Let $b_{k,i}$ denote the probability that the node is in the state $\{K_t=k, C_t=i\}$. Then $\sum_{k=-1}^{\ell}\sum_{i=0}^{CW_k} b_{k,i} = 1$. The state probabilities can be determined from the stationary distribution of the Markov chain:

$$b_{k,i} = \sum_{l=-1}^{\ell}\sum_{j=0}^{CW_l} b_{l,j} \cdot P\{k,i|l,j\}, \qquad 0 \le k \le \ell \text{ and } 0 \le i \le CW_k - 1 \tag{4.5}$$

The details are omitted for brevity and can be found in references cited in Section 4.7 below. A node transmits if its backoff counter reaches zero; thus:

$$q_c = \sum_{k=-1}^{\ell} b_{k,0} = b_{-1,0} + \frac{1-(\tilde{p}_{coll})^{\ell+1}}{\tilde{p}_{succ}} \cdot b_{0,0} = \frac{1 - p_{busy} \cdot \tilde{p}_{coll}^{\ell+1} + p_{busy} \cdot \tilde{p}_{coll}^{\ell+2} - \tilde{p}_{coll}^{\ell+2}}{\tilde{p}_{succ} \cdot p_{idle}} \cdot b_{-1,0}$$

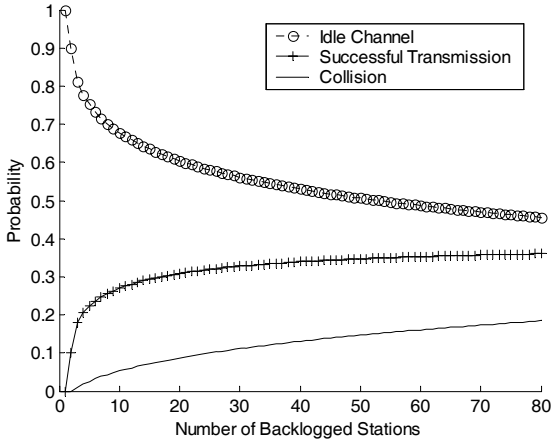


Figure 4-20. The dependence of the channel state probabilities on the number of backlogged nodes *n* seen from a node's viewpoint.

After solving for $b_{-1,0}$, we obtain:

$$q_c = \frac{2 \cdot p_i \cdot (1-2\tilde{p}_c) \cdot \left(1 - p_b \cdot \tilde{p}_c^{\ell+1} + p_b \cdot \tilde{p}_c^{\ell+2} - \tilde{p}_c^{\ell+2}\right)}{2p_i^2(1-2\tilde{p}_c)\tilde{p}_s + CW_0\tilde{p}_s(\tilde{p}_s p_b + \tilde{p}_c)\left[1-(2\tilde{p}_c)^6\right] + (1-2\tilde{p}_c)(\tilde{p}_s p_b + \tilde{p}_c)\left(1-\tilde{p}_c^{\ell+1}\right) + CW_{\max}\tilde{p}_c^6(1-2\tilde{p}_c)\left(1-\tilde{p}_c^{\ell-5}\right)(\tilde{p}_s p_b + \tilde{p}_c)}$$

(The notation is temporarily abbreviated for space reasons, e.g., $p_i = p_{idle}$, $\tilde{p}_s = \tilde{p}_{succ}$, etc.) In the case ℓ=5, the equation simplifies to:

$$q_c = \frac{2 \cdot p_i \cdot (1-2\tilde{p}_c) \cdot \left(1 - p_b \cdot \tilde{p}_c^{6} + p_b \cdot \tilde{p}_c^{7} - \tilde{p}_c^{7}\right)}{2 \cdot p_i^2 \cdot (1-2\tilde{p}_c) \cdot \tilde{p}_s + CW_0 \cdot \tilde{p}_s(\tilde{p}_s \cdot p_b + \tilde{p}_c) \cdot \left[1-(2\tilde{p}_c)^6\right] + (1-2\tilde{p}_c) \cdot (\tilde{p}_s \cdot p_b + \tilde{p}_c) \cdot \left(1-\tilde{p}_c^6\right)} \tag{4.6}$$

Figure 4-20 shows the dependence of different channel probabilities on the number of backlogged nodes from a single node's perspective.

4.5.2 Throughput and Delay Performance

For the sake of convenience, I first derive the overall average performance of the network. The average packet delay per node and throughput per node are then derived by noting that the aggregate network throughput must be shared among *n* backlogged nodes. This requires that the channel event probabilities be redefined from the channel viewpoint, rather than from an individual station's viewpoint as above. Therefore,

$$p'_{idle} = Q_c(0, n) = (1 - q_c)^n = (1 - q_c) \cdot p_{idle}$$

$$p'_{succ} = Q_c(1, n) = (1 - q_c) \cdot p_{succ} + q_c \cdot \tilde{p}_{succ}$$

$$p'_{coll} = 1 - p'_{idle} - p'_{succ}$$

The channel probability p'_{succ} is reflects the fact that channel success happens if the station under consideration does not transmit (it is in backoff) and another station succeeds or our station transmits and it succeeds.

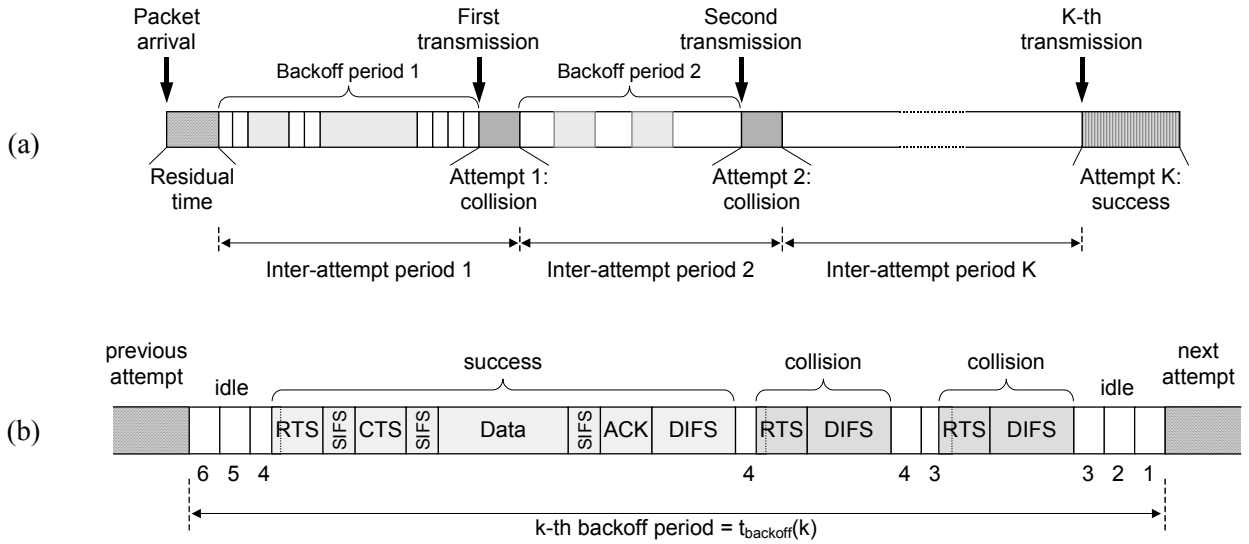


Figure 4-21: Illustration of the packet transmission process with multiple attempts. (a) The inter-attempt period (IAP) includes the backoff period and the attempted transmission time. Note that the first IAP does not include the residual time. (b) A station may be preempted multiple times during the backoff period, even within the same slot, see e.g., slot 4.

The packet transmission process is illustrated in Figure 4-21. The throughput and delay parameters are determined by applying Eq. (3.1). We define the following quantities:

- σ_{succ} ≡ Average number of contention slots between consecutive successful packet transmissions:

  $$\sigma_{succ} = 1 / p'_{succ}$$

  These slots may be idle or busy, and the busy slots may contain a successful packet transmission or a collision.

- The average numbers of channel events (idle, successful packet transmission, collision) every σ_{succ} contention slots are:

  $$n_{idle} = \sigma_{succ} \cdot p'_{idle} = \frac{p'_{idle}}{p'_{succ}}$$

  $$n_{succ} = 1$$

  $$n_{coll} = \sigma_{succ} \cdot p'_{coll} = \frac{p'_{coll}}{p'_{succ}}$$

- The channel time constants for different events (idle, success, and collision) are defined in terms of the number of backoff slots as follows:

  $$\tau_{idle} = 1$$
  $$\tau_{succ} = (\text{RTS} + \text{SIFS} + \text{CTS} + \text{SIFS} + E[\text{DATA}] + \text{SIFS} + \text{ACK} + \text{DIFS}) / t_{slot}$$
  $$\tau_{coll} = (\text{RTS} + 2\times\text{SIFS} + \text{CTS} + 2\times t_{slot} + \text{DIFS}) / t_{slot}$$

The collision time deserves a comment. The literature usually assumes the collision time equal to RTS + DIFS, see, e.g., [Bianchi 2000]. According to the 802.11 standard, the collision time is different depending on whether or not the node participates in collision. The waiting time for the nodes that caused collision is the CTS_Timeout (or ACK_Timeout in the basic transmission mode), which makes the total time equal to RTS + 2×SIFS + CTS + 2×t_{slot} + DIFS according to

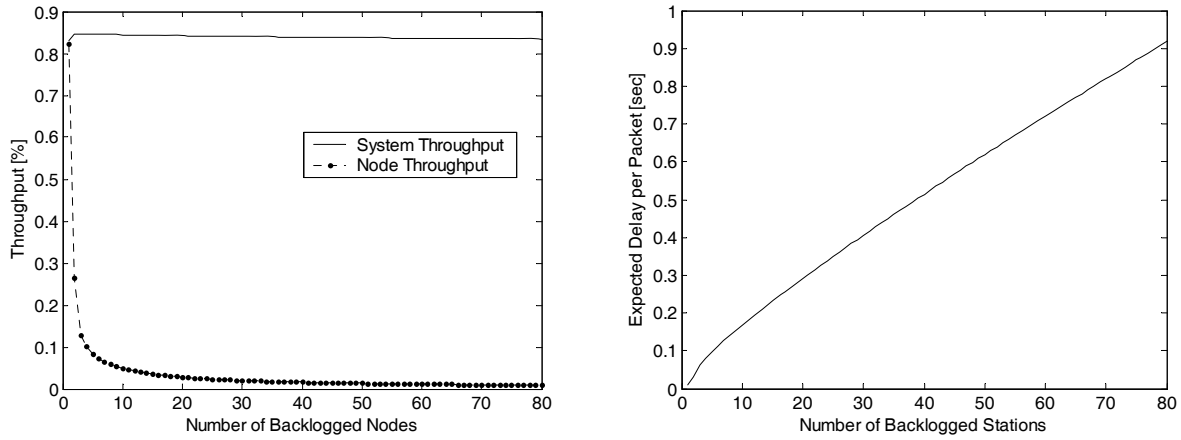


Figure 4-22. Percentage throughput (left) and expected packet delay (right) depending on the number of backlogged nodes.

the 802.11 standard [37], p. 79. Conversely, the waiting time for the *non-colliding* nodes is equal to the duration of busy channel plus EIFS, since these nodes do *not* correctly receive the RTS frame. EIFS is defined in Table 4-2 above and it is equal to the CTS_Timeout (or ACK_Timeout); thus τ_{coll} is defined as above. Then, the total number of contention slots per successful packet transmission is:

$$\tau = \tau_{succ} + n_{coll} \cdot \tau_{coll} + n_{idle} \tag{4.7}$$

(Compare this to Eq. (3.1).) The average total network throughput is:

$$\Gamma = \frac{1}{\tau \cdot t_{slot}} \text{ [packets/second]} \tag{4.8}$$

The average throughput per backlogged node is:

$$\gamma = \Gamma / n \text{ [packets/(second·node)]} \tag{4.9}$$

The expected delay per packet is:

$$W = 1 / \gamma = n / \Gamma = n \cdot \tau \text{ [seconds]} \tag{4.10}$$

This rough analysis is somewhat inaccurate since the packet's data transmission time (when at last successfully transmitted) is not part of its delay time. We also ignore the delay accrued due to dropped packets when a station reaches a retry limit. More importantly, as already noted in the discussion of Eq. (3.1), the analysis assumes that the collision probability remains the same, regardless of how many collisions a packet has already suffered.

Figure 4-22 shows the dependence of performance parameters on the number of backlogged nodes.

4.6 IEEE 802.11 Family of Standards

IEEE 802.11 is a family of different standards. Here I briefly review different standards and the interested reader should check http://standards.ieee.org/wireless/ for detailed information.

802.11c: support for 802.11 frames

802.11d: new support for 802.11 frames

802.11e: QoS enhancement in MAC

802.11f: Inter Access Point Protocol

802.11g: 2.4 GHz extension to 22 Mbps

802.11h: channel selection and power control

802.11i: security enhancement in MAC

802.11j: 5 GHz globalization

802.11n: WWiSE, speeds up to 540 Mbps

Why multiple overlapping protocols – discuss the relation to Bluetooth.

Fewer features vs. more features, depending on the application: security, QoS, power control, etc.

802.11a and 802.11b

802.11a standard was approved in 1999, but the products based on it appeared only in 2001. It works at 5GHz unlicensed bands and achieves data rates as high as 54 Mbit/sec. Thus, the standard can support many broadband applications, letting wireless users access the most demanding applications. More specifically, the object of this service is to provide wireless data links supporting rates up to 54 Mbps to link workstations, laptops, printers, and personal digital assistants to a network access node without the expense of cabling. The key points are high capacity and short range (order of tens of meters).

To dramatically increase throughput, 802.11a had to solve a major challenge of indoor radio frequency: delay spread in the current 2.4-GHz, single-carrier, delay-spread system. The 802.11a standard cleverly solves this through a modulation technique called Coded Orthogonal Frequency Division Multiplexing (COFDM). See more information here: http://www.nwfusion.com/news/tech/2001/0129tech.html.

802.11b is a high rate enhancement of the original 802.11 standard. It supports up to 11 Mbps in the 2.4-GHz unlicensed spectrum band. It was ratified in 1999 and the products based on it appeared in 1999. 802.11b specifies a high-rate direct-sequence layer (HR/DSSS).

Current (November 2002) prices for 802.11b adapters range from $45 to $140; for 802.11b access points, from about $385 to $1,050, depending on features. Network components complying with 802.11a or 802.11b standards are not interoperable.

In the U.S., 802.11a offers eight non-overlapping channels vs. three channels shared by 802.11b and 802.11g (see below). More channels make it easier to configure an 802.11a network to avoid interference. In dense installations, extra channels can make 802.11a networks up to 14 times faster than 802.11b networks.

The 802.11-based WLAN standards are promoted by the Wireless Ethernet Compatibility Alliance (WECA: http://www.wi-fi.org/) and branded Wi-Fi (802.11b) and Wi-Fi5 (802.11a).

802.11g

802.11g standard is designed as a higher-bandwidth (54M bit/sec) successor to 802.11b, or Wi-Fi standard, which tops out at 11M bit/sec. The IEEE has yet to approve the standard, and the ratification is expected to take place by March. An 802.11g access point will support 802.11b and 802.11g clients. Similarly, a laptop with an 802.11g card will be able to access existing 802.11b access points as well as new 802.11g access points. That is because wireless LANs based on 802.11g will use the same 2.4-GHz band that 802.11b uses. The higher data rate, translating into actual throughput of about 17M to 19M bit/sec, would give users more bandwidth for an array of multimedia and other data-intensive applications.

The result is a wireless LAN that has the longer range of 802.11b and the higher throughput of 802.11a. In addition, any existing 802.11b adapter card will be able to work with an 802.11g access point, which is not possible with 802.11a adapters.

Theoretically, 802.11g incorporates most of the good qualities of the other two standards. But there are also some potential issues. One issue is the number of channels that the various wireless LAN standards support. 802.11b and 802.11g use three channels; 802.11a uses eight. In practical terms, that means more 802.11a access points can be deployed in a given area, to support considerably more users, than is possible with 802.11b or 802.11g.

Still another drawback is that using the 2.4-GHz band could be as much of a drawback as a benefit for 802.11g. 802.11g does not do anything to mitigate the interference problem experienced by all wireless LANs at 2.4 GHz. Baby monitors, cordless phones and Bluetooth devices can all interfere with 802.11g, though Orthogonal Frequency Division Multiplexing used in 802.11g is more interference-immune than other modulation schemes. Table 4-3 compares the features of the three standards.

Table 4-3: Comparison of IEEE 802.11 a, b, and g wireless LAN standards. (Adapted from: http://www.nwfusion.com/techinsider/2002/0520wlan/0520feat1.html)

Standard	802.11a	802.11b	802.11g
Frequency range	5.1–5.8 GHz	2.4–2.485 GHz	2.4–2.485 GHz
Number of channels	×		
Interference	×		
Bandwidth (Data rate)	× (up to 54 Mbps)	(up to 11 Mbps)	× (up to 54 Mbps)
Power consumption		×	×
Range/penetration		×	×
Upgrade/compatibility			×
Price		×	×

× indicates superior technology or feature

For more information on 802.11g, check here: http://www.nwfusion.com/news/2002/1111comdex.html.

802.11h

The U.S. 802.11a standard will be adapted in Europe to conform to that region's regulations; that variant will be called 802.11h. The two 5GHz 802.11 standards are nearly identical, except that 802.11h adds TPC (Transmit Power Control) which limits the PC card from emitting more radio signal than is needed, and DFS (Dynamic Frequency Selection), which lets the device listen to what is happening in the airspace before picking a channel. TPC and DFS are European requirements.

In Europe, vendors united in the HiperLAN2 Global Forum (http://www.hiperlan2.com/), and the European Telecommunications Standardization Institute (ETSI) are pushing the HiperLAN2 standard, which is competing with 802.11h. The HiperLAN2 and 802.11 standards have nearly identical physical layers, but are very different at the MAC layer. The products are not interoperable. 802.11 is a true wireless Ethernet, while HiperLAN2 on a technical level is more like wireless ATM (Asynchronous Transfer Mode).

IEEE 802.11i

WEP (Wired Equivalent Security) algorithm was suggested in the IEEE standard 802.11-1999 to provide security equivalent with that of a wired Ethernet. The WEP algorithm should insure confidentiality and integrity of the frames on the wireless network. A cyclic redundancy check is used to compute an integrity check value, which is concatenated on the message before encrypting with the stream cipher RC4. RC4 is a symmetric cipher, i.e., the same key encrypts and decrypts the data. The key used is a per-packet key which is obtained by concatenating the Initialization Vector (IV) with the user key. Because of export regulations the standard specifies 64-bit keys where 24 bits are the known IV, but many vendors have also implemented 128-bit keys where 24 bits are the IV.

From a cryptographical point of view WEP is totally broken, there is no security features left. Authentication, access control, replay prevention, message modification detection, message privacy and key protection can be circumvented by a dedicated attacker. It is still recommended by most vendors and security experts to have WEP enabled, it will keep the average computer user off your network. However, there are hacker tools available on the Internet that easily can be used to break all the security features. See here for more details about the attacks.

Attacks on WEP: http://www.ii.uib.no/~moen/wireless/WEP/attacks.shtml

The IEEE 802.11i committee has defined the Temporal Key Integrity Protocol (TKIP) as an interim standard, compatible with existing wireless networks, and designed to provide "good enough" security, pending a stronger standard. TKIP has been tested intensively, but has had a shorter testing period than usual for a critical security standard.

IEEE 802.1x: This standard, supported by Windows XP, defines a framework for MAC-level authentication. Unfortunately, two University of Maryland researchers recently noted serious flaws in client-side security for 802.1x. (See online: http://www.cs.umd.edu/~waa/1x.pdf)

See also http://www.ii.uib.no/~moen/wireless/WPA/802_1x.shtml

IEEE 802.11n

WWiSE, speeds up to 540 Mbps

See: http://www.cnn.com/2004/TECH/internet/08/13/tech.wireless.reut/index.html

4.7 Summary and Bibliographical Notes

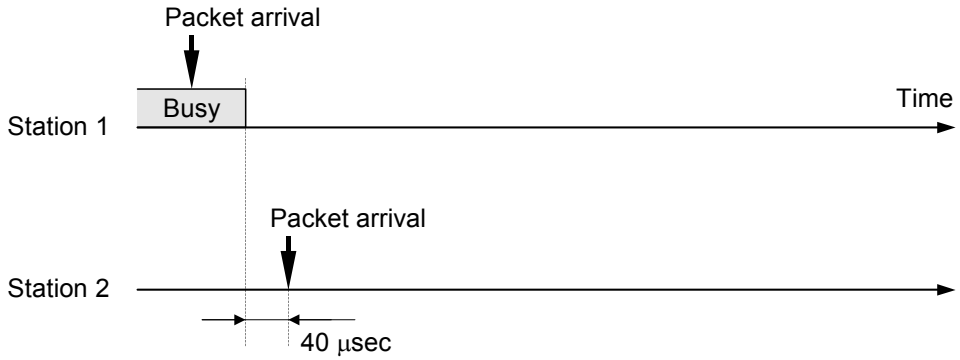
This chapter reviews the IEEE 802.11 wireless LAN standard, which is currently one of the most popular commercially available wireless networks. Books that cover IEEE 802.11 are [Gast 2002; O'Hara & Petrick 1999].

The performance analysis of IEEE 802.11 in Section 4.5 follows the method presented originally in [Bianchi 2000] and refined in [Wu *et al*. 2002; Ziouva & Antonakopoulos 2002].

Problems

Problem 4.1

A timing diagram for two IEEE 802.11b mobile stations is shown in the figure, the packet arrival times are indicated by the bold arrows.



Complete the timing diagrams for transmission of the two newly arrived packets. Assume the basic access mechanism.

Problem 4.2

Consider an independent BSS (IBSS) with two mobile STAs, *A* and *B*, where each station has a single packet to send to the other one. Draw precise time diagram for each station from the start to the completion of the packet transmission. For each station, select different packet arrival time and a reasonable number of backoff slots to count down before the station commences its transmission so that no collisions occur during the entire session. Assume that both station are using the basic transmission mode and only the first data frame transmitted is received in error (due to channel noise, not collision).

Problem 4.3

Explain why we cannot use the propagation time, t_{prop}, for the backoff slot length in IEEE 802.11 and similar CSMA/CA protocols. Why it has to be the parameter β, Eq. (3.2)? In other words, why the stations cannot in every backoff slot just observe whether the first bit of a new packet is arriving? Search the literature to determine what parameters are used in Eq. (3.2) to obtain the slot length of 20 μsec (see Table 4-2).

Problem 4.4

For an 802.11 BSS using the basic transmission mode, assume that the probability of data packet error in forward direction is p_f and the probability of ACK packet error in reverse direction is p_r. Determine is the expected average number of transmissions and retransmissions? (Assume that the number of retries is not limited, i.e., $\ell = \infty$)

Problem 4.5

Draw the timing diagrams as in Figure 4-8 of the IEEE 802.11 RTS/CTS protocol for different cases of frame transmissions.

Problem 4.6

The link layer protocols usually do not support frame fragmentation, but 802.11 does. Explain why?

Problem 4.7

IEEE 802.11 is based on CSMA/CA which expects positive acknowledgement for every frame (a kind of stop-and-wait ARQ). If the ACK does not arrive, it assumes that the frame is lost to a collision. Discuss what kind of performance problems this can cause when the communicating stations are at the fringe area of each other's transmission range.

Problem 4.8

Search the web for the definition of IEEE 802.11 parameter RSSI (Received Signal Strength Indicator). Wireless card vendors provide application programming interface (API) for reading the RSSI and noise level for each received packet. For example, Cisco has Aironet Client Utility (ACU). For Linux, this information is available from http://www.hpl.hp.com/personal/Jean_Tourrilhes/Linux/Linux.Wireless.Extensions.html. Search the

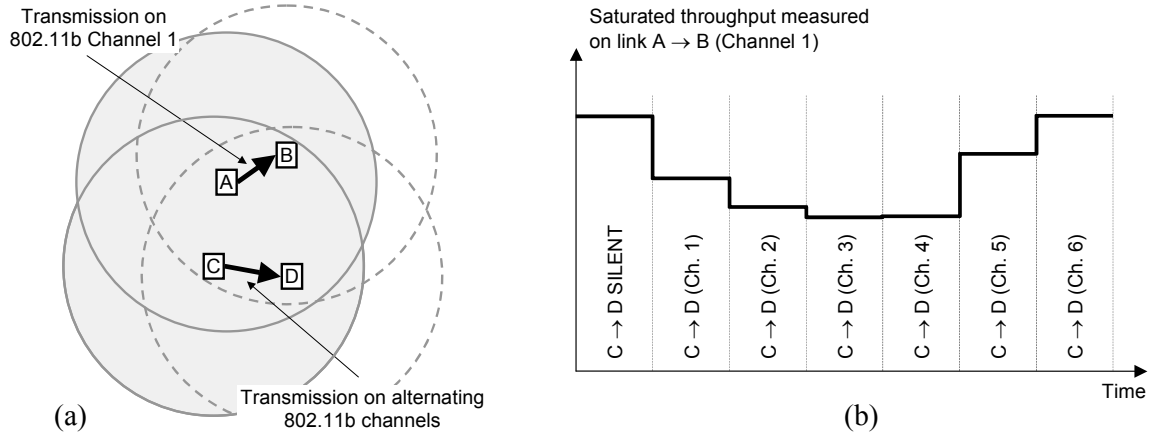


Figure 4-23. Illustration for Problem 4.9. (a) Concurrent communication sessions (A→B and C→D) interfere with each other. (b) Saturated throughput observed for session A→B as session C→D employs different channels.

wireless chipset reference manuals and other sources and describe how are RSSI and noise level measured in 802.11. What is the duration for the measurement of each?

Problem 4.9

Suppose you perform the following experiment with an IEEE 802.11b network. (Assume *no* auto rate adjustment feature is involved.) There are two parallel communication sessions: station A sends to B and C sends to D, both at the maximum possible rate. As illustrated in Figure 4-23(a), all stations are within each other's transmission range. Assume there are no other channel impairments except the mutual interference from the parallel sessions. Now suppose that stations A and B use only Channel 1 for transmissions and stations C and D change channels from 1 to 6, starting with 1. The throughput measured at B would be as shown Figure 4-23(b).

(a) Explain why the throughput at first decreases and then increases. In particular, explain why the achieved throughput is worse when the concurrent transmissions are in neighboring channels (A→B in Channel 1 C→D in Channel 2) then when they both share Channel 1.

(b) Note that, when C→D changes from silent to transmission in Channel 1, the throughput does not drop by 50 %. What could be reason(s) for this?

[Hint: Remember that 802.11b channels overlap. Search the Web for literature on how *clear channel assessment* (CCA) mechanism works. See also Ref. [37], Section 15.4.8.4, page 223.]

Problem 4.10

Using Eq. (4.5), compute the probability that a station is in the backoff stage k as $b_k = \sum_i b_{ki}$. Compare the numeric values obtained in this manner to those obtained in Problem 3.22 in Chapter 3.

Problem 4.11

Assume that t_{proc} and t_{prop} are the packet processing time at the end-hosts and the packet propagation time across the 802.11 channel, respectively. Modify Eq. (4.7) to include these.

Problem 4.12

Given the knowledge of current/predicted signal strength/ bandwidth, devise a policy for selecting the stations to which to talk vs. those to which not to talk because of the low bandwidth it is likely the conversation will be wasted to frequent retransmissions.
[Hint: Search the Web for literature on opportunistic transmissions.]

Problem 4.13

Determine the probability p_{drop} that an 802.11 station will reach the retry limit ℓ and drop the current packet. (Compare this with Problem 3.23 in Chapter 3.)

(a) What is the delay t_{drop} experienced due to a dropped packet? [Remember that Eq. (4.10) does not include the delay due to dropped packets.]

(b) Assuming that an upper-layer *reliable* stop-and-wait protocol is operating on the station and it must retransmit the dropped packets (remember that 802.11 is semi-reliable), what is the average delay for a successful packet transmission?

Assume an Infrastructure Basic Service Set (BSS) with an access point (AP) that can hear and talk to the n backlogged stations in its area of coverage, and all nodes (including AP) have the same transmission range. For an arbitrary station "A" there are $n_h \le n-2$ backlogged stations that are "hidden" to it (excluding itself and the AP, assuming that AP is also backlogged) and $n_c = n-n_h-1$ covered nodes (excluding itself), see Figure 4-24(a). In the Infrastructure BSS scenario, any station (including the AP) can transmit, but the sink of any station's traffic must be the AP. The key problem is that different stations observe different channel state, as illustrated in Figure 4-24. Suppose a backlogged station A is ready to transmit to the AP and currently undergoes the backoff countdown. Other backlogged stations, e.g., B and C, may transmit during or at the end of A's countdown.

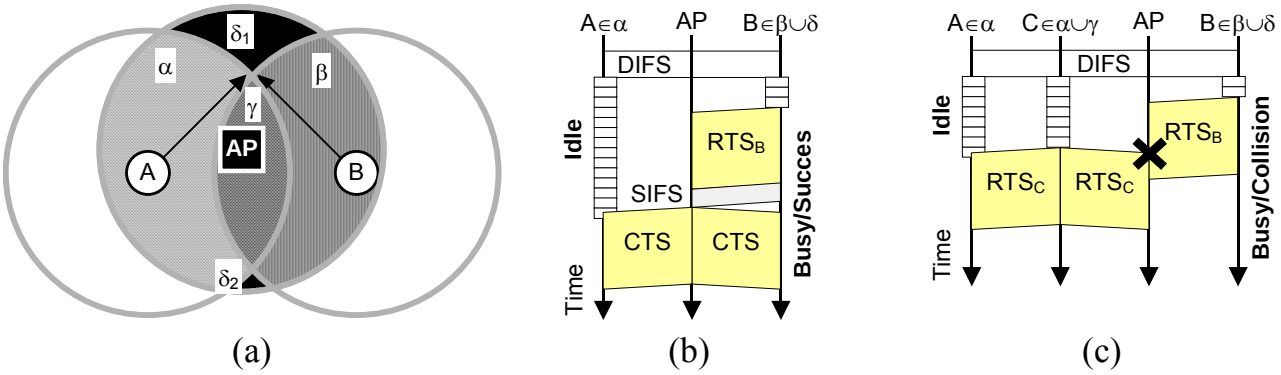
(a) Catalogue all the possible channel states observed by the station A and the AP under different events in a tabular form as follows:

Number of covered STAs that transmit	Number of hidden STAs that transmit	Channel state observed by A	Channel state observed by the AP

(b) A given station will observe a successful transmission if and only if a single station transmits and no other stations transmit during that station's vulnerability period. The success state probability is:

$$p_{succ}(n_h) = Q_c(1, n_c) \cdot Q_v(0, n_h) + Q_v(0, n_c) \cdot Q_c(1, n_h) \qquad (\#)$$

Assuming the uniform distribution of all nodes as well as the backlogged ones and the same coverage radius for all stations, what are the values of n_c and n_h for an "average" station? (Hint: Check [Ni *et al*., 1999].)



(a) (b) (c)

Figure 4-24. Hidden station problem in 802.11 (Problem 4.13). (a) The regions of coverage relative to a station A under consideration. (b) Example timing diagram when a hidden station B transmits first: At first A observes idle channel state, then observes busy/success state and suspends the countdown. (c) Example when a hidden station B transmits first followed by a visible station C: At first A observes idle channel state, then observes busy/collision state.

(c) Write the equation for the probability of collision, $p_{coll}(n_h)$, similar to (#) above.

(d) Draw a qualitative chart showing how the probabilities $p_{succ}(n_h)$ and $p_{coll}(n_h)$ depend on the distance of a node under consideration from the AP.

(e) Draw a qualitative chart showing how the fractions of times for which a node under consideration observes different channel states, depending on its distance from the AP.

Problem 4.14

Discuss possible consequences of multipath reflections on the way IEEE 802.11 receiver measures noise level. Assume different distances between the sender and the receiver, e.g., 10 m, 100 m, and 300 m.

[Hint: The radio signal strength becomes attenuated with the traveled distance. Assume that it becomes too low after traveling about 500 m distance, and calculate the greatest possible phase delay of the multipath signal. If sufficiently delayed, this signal is measured as noise.]

Problem 4.15

Given the knowledge of current/predicted signal strength/ bandwidth, devise a policy for selecting the stations to which to talk vs. those to which not to talk because of the low bandwidth it is likely the conversation will be wasted to frequent retransmissions.

[Hint: Search the Web for literature on opportunistic transmissions.]

Problem 4.16

Chapter 5
Ad Hoc Networks

Contents

5.1 Introduction

Ad hoc networks are infrastructureless wireless networks in which nodes act as relay for packets generated by and addressed to different nodes. An Ad hoc network is self-organizing and adaptive. This means that a formed network can be de-formed on-the-fly without the need for any system administration. The term "ad hoc" tends to imply "can take different forms" and "can be mobile, standalone, or networked." A mobile ad hoc network is commonly referred to as MANET.

When discussing routing protocols, "host" is usually term that reserved for communication endpoints and "node" is used for intermediary computing nodes that relay packets on their way to destination. In wired networks with fixed infrastructure, these are usually not mixed, whereas in ad hoc networks it is common that computing nodes assume both roles. Therefore, in the rest of this chapter I use interchangeably "host" and "node."

In a multihop wireless ad hoc network, mobile nodes cooperate

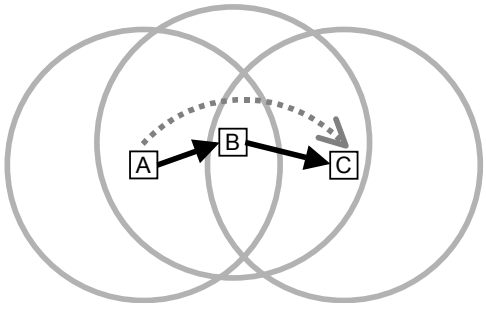


Figure 5-1: Example mobile ad-hoc network: *A* communicates with *C* via *B*.

to form a network without the help of any infrastructure such as access points or base stations. The mobile nodes, instead, forward packets for each other, allowing nodes beyond direct wireless transmission range of each other to communicate over possibly multihop routes through a number of forwarding peer mobile nodes. The mobility of the nodes and the fundamentally limited capacity of the wireless channel, together with wireless transmission effects such as attenuation, multipath propagation, and interference, combine to create significant challenges for network protocols operating in an ad hoc network.

In ad hoc network, a distributed routing protocol is employed to determine the hop-by-hop path that a packet follows between source and destination.

5.2 Routing Algorithms

The main purpose of routing is to bring packets efficiently to their destination. The capacity of the resulting end-to-end path directly depends on the efficiency of the routing protocol employed.

Since the MANET environment presupposes the possibility of multiple hop packet forwarding, there is an implicit requirement on network nodes to participate in some sort of routing algorithm. Most of the current research in MANETs is on efficient and scaleable routing algorithms. The mobility of network nodes, the variability of wireless link quality and the lack of hierarchical structure in the physical topology of the network makes routing in MANETs arguably much more difficult than in wired networks.

Although there have been dozens of new routing protocols proposed for MANETs, the majority of these protocols actually rely on fundamental techniques that have been studied rigorously in the wired environment. However, each protocol typically employs a new heuristic to improve or optimize a legacy protocol for the purposes of routing in the mobile wireless environment. In fact, there are a few mechanisms that have received recent interest primarily because of their possible application to MANETs. There are two main classes of routing protocols:

- Proactive
    - Continuously update reachability information in the network
    - When a route is needed, it is immediately available
    - DSDV by Perkins and Bhagwat (SIGCOMM 94)
    - Destination Sequenced Distance vector
- Reactive
    - Routing discovery is initiated only when needed
    - Route maintenance is needed to provide information about invalid routes
    - DSR by Johnson and Maltz
    - AODV by Perkins and Royer

- Hybrid
    - Zone routing protocol (ZRP)

Centralized vs. localized solution:

Nodes in *centralized* solution need to know full network information to make decision; mobility or changes in activity status (power control) cause huge communication overhead to maintain the network information.

Nodes in *localized* algorithm require only local knowledge (direct neighbors, 2-hop neighbors) to make decisions. Majority of published solutions are centralized, compared with other centralized solutions.

Next, a brief survey of various mechanisms is given.

5.2.1 Dynamic Source Routing (DSR) Protocol

Source routing means that the sender must know in advance the complete sequence of hops to be used as the route to the destination. DSR is an *on-demand* (or *reactive*) ad hoc network routing protocol, i.e., it is activated only when the need arises rather than operating continuously in background by sending periodic route updates. DSR divides the routing problem in two parts: *Route Discovery* and *Route Maintenance*, both of which operate entirely on-demand. In Route Discovery, a node actively searches through the network to find a route to an intended destination node. While using a route to send packets to the destination, Route Maintenance is the process by which the sending node determines if the route has broken, for example because two nodes along the route have moved out of wireless transmission range of each other.

An example is illustrated in Figure 5-1, where host *C* needs to establish a communication session with host *H*. A node that has a packet to send to a destination (*C* in our example) searches its Route Cache for a route to that destination. If no cached route is found, node *C* initiates Route Discovery by broadcasting a ROUTE REQUEST (RREQ) packet containing the destination node address (known as the *target* of the Route Discovery), a list (initially empty) of nodes traversed by this RREQ, and a *request identifier* from this source node. The request identifier, the address of this source node (known as the *initiator* of the Route Discovery), and the destination address together uniquely identify this Route Discovery attempt.

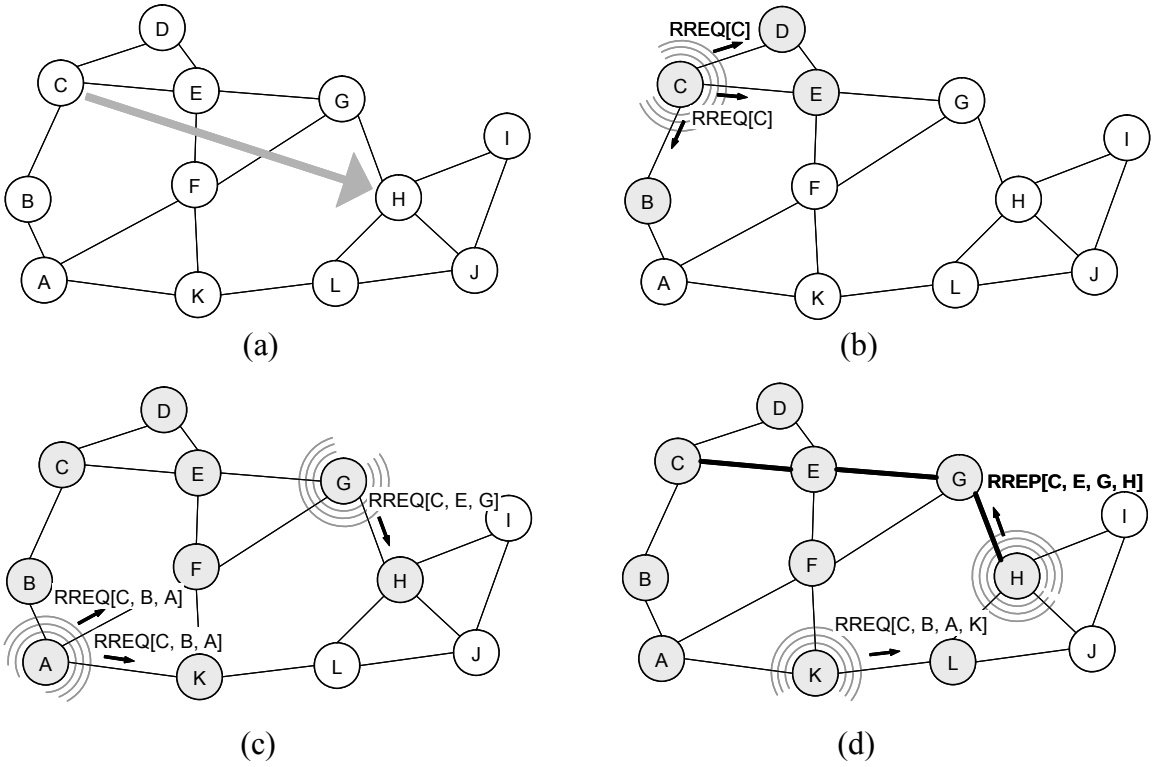


Figure 5-2: Route discovery in DSR: node C seeks to communicate to node H. Gray shaded nodes already received RREQ. The path in bold in (c) indicates the route selected by H for RREP. See text for details. (Note: the step where B and E broadcast RREQ is not shown.)

A node receiving a ROUTE REQUEST checks to see if it has previously forwarded a RREQ from this Discovery by examining the IP Source Address, destination address, and request identifier. For example, in Figure 5-2(b), nodes *B*, *E*, and *D* are the first to receive RREQ and they re-broadcast it to their neighbors. If the recipient of RREQ has recently seen this identifier, or if its own address is already present in the list in RREQ of nodes traversed by this RREQ, the node silently drops the packet. Otherwise, it appends its address to the node list and re-broadcasts the REQUEST. When a RREQ reaches the destination node, *H* in our example, this node returns a ROUTE REPLY (RREP) to the initiator of the ROUTE REQUEST. If an intermediary node receives a RREQ for a destination for which it caches the route in its Route Cache, it can send RREP back to the source without further propagating RREQ. The RREP contains a copy of the node list from the RREQ, and can be delivered to the initiator node by reversing the node list, by using a route back to the initiator from its own Route Cache, or "piggybacking" the RREP on a new ROUTE REQUEST targeting the original initiator. This path is indicated with bold lines in Figure 5-2(d). When the initiator of the request (node *C*) receives the ROUTE REPLY, it adds the newly acquired route to its Route Cache for future use.

In Route Maintenance mode, an intermediary node forwarding a packet for a source attempts to verify that the packet successfully reached the next hop in the route. A node can make this confirmation using a hop-to-hop acknowledgement at the link layer (such as is provided in IEEE 802.11 protocol), a passive acknowledgement (i.e., listen for that node sending packet to its next hop), or by explicitly requesting network- or higher-layer acknowledgement. Transmitting node

can also solicit ACK from next-hop node. A packet is possibly retransmitted if it is sent over an unreliable MAC, although it should not be retransmitted if retransmission has already been attempted at the MAC layer. If a packet is not acknowledged, the forwarding node assumes that the next-hop destination is unreachable over this link, and sends a ROUTE ERROR to the source of the packet, indicating the broken link. A node receiving a ROUTE ERROR removes that link from its Route Cache.

In the basic version of DSR, every packet carries the entire route in the header of the packet, but some recent enhancements to DSR use implicit source routing to avoid this overhead. Instead, after the first packet containing a full source route has been sent along the route to the destination, subsequent packets need only contain a flow identifier to represent the route, and nodes along the route maintain flow state to remember the next hop to be used along this route based on the address of the sender and the flow identifier; one flow identifier can designate the default flow for this source and destination, in which case even the flow identifier is not represented in a packet.

A number of optimizations to the basic DSR protocol have been proposed [Perkins 2001, Chapter 5]. One example of such an optimization is *packet salvaging*. When a node forwarding a packet fails to receive acknowledgement from the next-hop destination, as described above, in addition to sending a ROUTE ERROR back to the source of the packet, the node may attempt to use an alternate route to the destination, if it knows of one. Specifically, the node searches its Route Cache for a route to the destination; if it finds one, then it salvages the packet by replacing the existing source route for the packet with the new route from its Route Cache. To prevent the possibility of infinite looping of a packet, each source route includes a *salvage count*, indicating how many times the packet has been salvaged in this way. Packets with salvage count larger than some predetermined value cannot be salvaged again.

In summary, DSR is able to adapt quickly to dynamic network topology but it has large overhead in data packets. The protocol does not assume bi-directional links.

5.2.2 Ad Hoc On-Demand Distance-Vector (AODV) Protocol

DSR includes source routes in packet headers and large headers can degrade performance, particularly when data contents of a packet are small. AODV attempts to improve on DSR by maintaining *routing tables* at the nodes, so that data packets do *not* have to contain routes. AODV retains the desirable feature of DSR that routes are maintained only between nodes which need to communicate.

ROUTE REQUEST packets are forwarded in a manner similar to DSR. When a node re-broadcasts a ROUTE REQUEST, it sets up a reverse path pointing towards the source. AODV assumes symmetric (bi-directional) links. When the intended destination receives a RREQ, it replies by sending a ROUTE REPLY. RREP travels along the reverse path set-up when RREQ is forwarded.

An intermediate node (not the destination) may also send a RREP, provided that it knows a more recent path than the one previously known to sender S. To determine whether the path known to an intermediate node is more recent, destination sequence numbers are used. The likelihood that an intermediate node will send a RREP when using AODV is not as high as in DSR. A new

RREQ by node S for a destination is assigned a higher destination sequence number. An intermediate node, which knows a route but with a smaller sequence number, *cannot send* RREP.

A routing table entry maintaining a reverse path is purged after a timeout interval. Timeout should be long enough to allow RREP to come back. A routing table entry maintaining a forward path is purged if not used for an active_route_timeout interval. If no is data being sent using a particular routing table entry, that entry will be deleted from the routing table (even if the route may actually still be valid).

In summary, routes in AODV need not be included in packet headers. Nodes maintain routing tables containing entries only for routes that are in active use. At most one next-hop per destination is maintained at each node, whereas DSR may maintain several routes for a single destination. Lastly, unused routes expire even if topology does not change.

5.2.3 End-to-End Path Capacity

Routing requires control messages or signaling, which is overhead and decreases channel capacity for useful payload.

5.2.4 Mobility

Communication between arbitrary hosts in a MANET requires routing over multiple-hop wireless paths. The main difficulty arises because without a fixed infrastructure these paths consist of wireless links whose end-points are likely to be moving independently of one another. Consequently, node mobility causes frequent failure and activation of links, leading to increased network congestion while the network routing algorithm reacts to topology changes.

5.3 Capacity

[Gupta & Kumar 2000]

Assume n nodes in area A transmitting at W bits/sec using a fixed range (distance between a random pair of nodes is $O(\sqrt{n})$. Bit-distance product that can be transported by the network per second is $\Theta\left(W \cdot \sqrt{A \cdot n}\right)$. Throughput per node is $\Theta\left(W \cdot \sqrt{n}\right)$.

5.3.1 Capacity of Mobile Ad Hoc Networks

[Grossglauser & Tse 2001]

Assume random motion, so any two nodes become neighbors once in a while. Each node assumed sender for one session, and destination for another session. Relay packets through at most one other node. Packet go from S to D directly, when S and D are neighbors, or from S to a

relay and the relay to D, when each pair becomes neighbor respectively. Then, throughput of each session is $O(1)$, independent of n.

[Li *et al*. 2001]

5.3.2 Measured Scaling Law

Measured in static networks [Gupta & Kumar 2001]. Throughput declines worse with n than theoretically predicted. Due to limitations of existing MAC protocols. Unable to exploit "parallelism" in channel access.

How to design MAC and routing protocols to approach theoretical capacity is an open research problem.

5.4 MAC Protocols

In discussing MAC in Chapter 3, I assumed a single broadcast region, where the source and the destination can hear each other, i.e., the end-to-end communication path consists of a single link or hop. In MANETs that is not always the case; very often the source and destination communicate over one or more intermediary nodes, which can hear both of them. Therefore, MAC for MANETS has the following characteristics (see Figure 5-1):

1. B needs to transmit acknowledgement to A, and this causes B to delay the $B \rightarrow C$ transmission.
2. A has to be quiet while B transmits to C.

Nodes can use a passive acknowledgement in which this node overhears the wireless transmission of the next hop node forwarding the packet on to the following hop along the source route (due to wireless broadcast).

5.5 User Mobility

To provide communication services to a larger number of users using the limited radio spectrum, wireless systems are designed based on the cellular concept for frequency reuse. Cellular providers expand their capacity by shrinking the cells and using the same frequency simultaneously in different parts of town. By dividing a large service area into small non-overlapping (in an ideal case) cells and letting each base station communicate with all the mobile stations in the cell with low transmitter power, radio frequency spectrum can be reused in different cells subject to transmission quality satisfaction. As a mobile station moves from cell to

cell, the serving base station changes. The process in which a mobile station switches its serving base station while crossing the cell boundary is referred to as *handoff management*. The process that tracks the user's movement, supports user roaming on a large scale, and delivers calls to the user at its current location is referred to as *location management*. As a result, the operation of a wireless network requires proper *mobility management* functions, which include both handoff management and location management.

Need for user tracking

Location management is the process of identifying the physical location of the user so that data directed to that user can be routed to the user's location (cell). *Routing* consists of setting up a route through the network over which data directed to a particular user is sent, and dynamically reconfiguring the route as the user location changes. In cellular systems location management and routing are coordinated by the base stations or the central mobile telephone switching office (MTSO), whereas on the Internet these functions are handled by the Mobile Internetworking Routing Protocol (Mobile IP). Mobile IP does not support real-time handoff of a mobile station between different networks: it is designed mainly for stationary users who occasionally move their computers from one network to another.

5.6 Summary and Bibliographical Notes

The chapter briefly introduces mobile ad hoc networks. A collection of articles on mobile ad hoc networks, particularly the routing protocol aspect, is available in [Perkins 2001]. [Murthy & Manoj 2004] provide a comprehensive overview of ad hoc networks.

Problems

Problem 5.1

x

Chapter 6
Technologies and Future Trends

Contents

6.1 Introduction

By far the most successful application of wireless networking has been the cellular telephone system. The current generation of cellular systems are all digital. In addition to voice communication, these systems provide e-mail, voice mail, and paging services. It is still uncertain whether there will be a large demand for all wireless applications. Companies are investing heavily to build multimedia wireless systems, yet the only highly profitable wireless application so far is voice.

Proliferation of Wi-Fi hotspots, see the reports on municipal wireless and broadband projects at: http://www.muniwireless.com/reports/.

Network Architecture - The Reality: The Mass Market is the Critical Issue

● Picture a map of the U.S. with dots representing the major cities. Now connect the dots with fiber.

Industry has been there, and done that (and continues to do so!)

● Picture flying into the SF Bay Area at night. Picture each light as a building or vehicle that needs to be connected by fiber or broadband wireless respectively. Now design a cost effective network for 20% market share.

That's a challenge!

Weinstein [2002] argues that 3G will never take off; it will simply be overtaken by IEEE 802.11 deployments. [See recent Network World announcement] See also [Lehr & McKnight 2003]

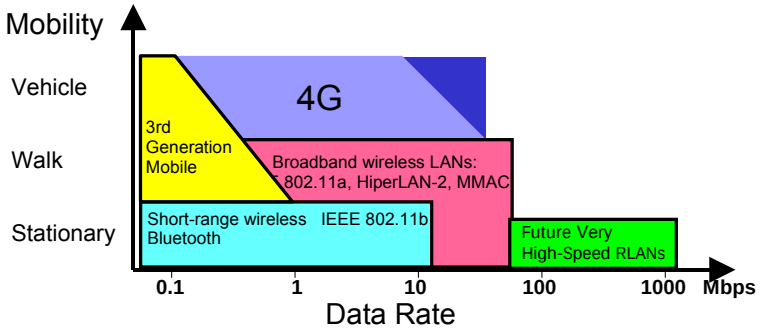


Figure 6-1: Wireless data rates.

New research problems:

- Continuous connectivity for a mobile host

- Seamless movement between networks

Many new wireless ventures are listed at WirelessVentures annual private-equity forum on mobile computing and data communications (http://wireless-ventures.com/).

6.2 Heterogeneous Wireless Networks

Wireless applications have different requirements which make it difficult to build single wireless system that can satisfy them all. Figure 6-1 shows the data rates supported by different wireless technologies for different ranges of mobility. Naturally, the rates shown are optimal and apply only near the center of the cell (near the base station) and assuming an unobstructed line-of-sight communication. Data rates obtained at the cell periphery may differ by order of magnitude.

3G Cellular Radio Systems

UMTS, IMT-2000, …

Broadband Wireless Local Area Networks (WLANs)

IEEE 802.11a, ETSI HiperLAN-2, ARIB MMAC, …

Short-Range Wireless Personal Area Networks (PANs)

Bluetooth, IEEE 802.15, HomeRF, …

WANs (cellular, satellite)

LANs (IEEE 802.11/Wi-Fi, HiperLAN)

PANs (Bluetooth)

IETF MANET – Ad Hoc

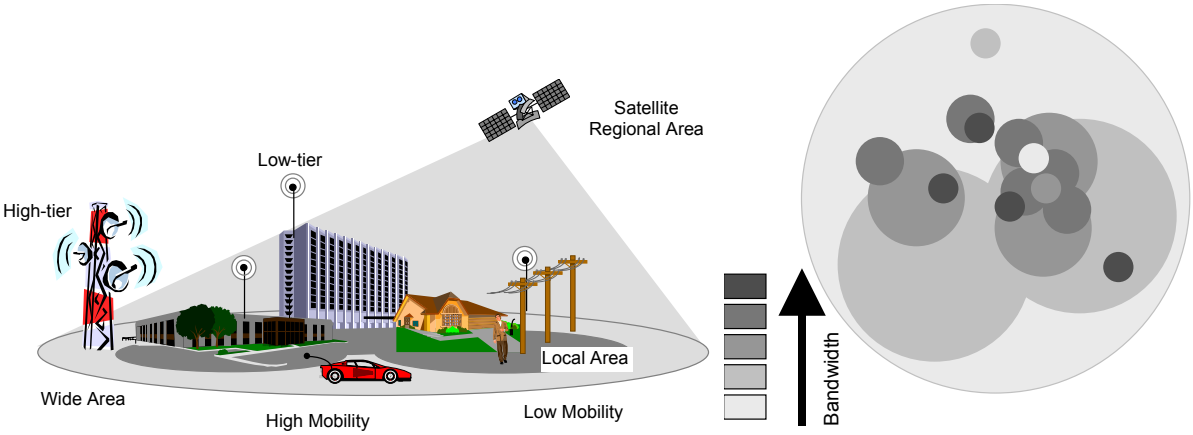


Figure 6-2: Cellular hierarchy and heterogeneous wireless data networks.

These networks provide coverage regions that range from pico-cells, on the order of meters, provided by technologies such as Bluetooth, to cells which spread over entire geographic regions, provided by satellite networks. (See Walrand & Varaiya, page 310).

Seamless mobility across diverse overlay networks

"vertical" hand-offs

software "agents" for heterogeneity management

IP as the common denominator?

How to achieve seamless mobility between these?

6.2.1 3G Cellular Services

First generation cellular networks were analog. Carriers upgraded networks to second generation digital cellular technologies to add capacity without adding spectrum. TDMA, GSM, and CDMA are air interfaces used to transmit signals between handsets and antennas at base stations. Their throughput is 14.4 kbps on TDMA and CDMA networks and 9.6 kbps on GSM networks. As cellular service became very popular, carriers wanted to increase capacity for voice traffic and offer higher speed cellular service for data and multimedia capability. They also wanted to create a standard so that subscribers could use their mobile telephone number and handset worldwide.

The International Telecommunications Union (http://www.itu.int/) started an effort called IMT-2000 (International Mobile Telephone) to define one advanced digital standard for high-speed data on cellular networks. Unfortunately, because of political pressure from manufacturers the ITU eventually endorsed two third generation technologies (Table 6-1). Ericsson and Nokia back one, W-CDMA. Qualcomm supports the other, cdma2000. Qualcomm receives royalties on all W-CDMA and cdma2000 service and it also supplies chips for cdma2000 handsets.

Table 6-1: Third generation (3G) cellular services.

Service	Other Designations for the Service	Comments
W-CDMA Wideband code division multiple access	Also known as Universal Mobile Telecommunications System (UMTS). European Telecommunications Standards Institute (ETSI) is reviewing proposals for a UMTS standard	Ericsson and Nokia supports this 3G technology for higher speed data. Most GSM networks have stated they will evolve their networks to W-CDMA
cdma2000	An upgrade from CDMA IS-95A and IS-95B service*	Qualcomm supports this 3G technology. Most CDMA providers and some providers in developing countries such as China and India have stated they will use cdma2000
The following are three different cdma2000 platforms		
1xRTT First generation candidate radio transmission technologies	IS-95C 1xMC (first generation multi-carrier)	First "generation" of cdma2000 service that doubles voice capacity and increases data speeds
1xHDR First generation high data rate	1xEV-DO (first generation evolution data only)	Higher data speeds but no increase in capacity for voice traffic
3xMC Three multi-carrier		Bonds three channels together to achieve higher data speeds

* IS-95A is the CDMA technology used in most of the world. IS-95B is an upgrade that provides packet data service. Most carriers are migrating from IS-95A to 1xRTT.

6.2.2 Hybrid Networks

Hybrid networks of infrastructure-based and MANETs.

Considering Figure 6-3, it is apparent that spectrum efficiency (i.e., useful capacity) is maximized when a communicating pair of nodes are close to one another. Therefore, it should be possible to employ multiple short distance transmissions to maximize throughput. As shown in Figure 6-3(a), a node situated in a remote part of the access point's area of coverage may alternately obtain Internet access via a multi-hop path. If the capacity-range relationship of Figure 4-14 is applied, the multi-hop path of Figure 6-3(a) provides two benefits: both the application throughput gain and the network efficiency gain have been improved. That is, a transmission rate of 11Mbps over 4-hop yields a capacity efficiency of 2.75Mbps. (This is obtained by dividing the 11Mbps bit rate by 4 packet transmissions of which 3 are relays, since they cannot send simultaneously.) This is superior to the 1Mbps capacity efficiency that is achieved over the direct link.

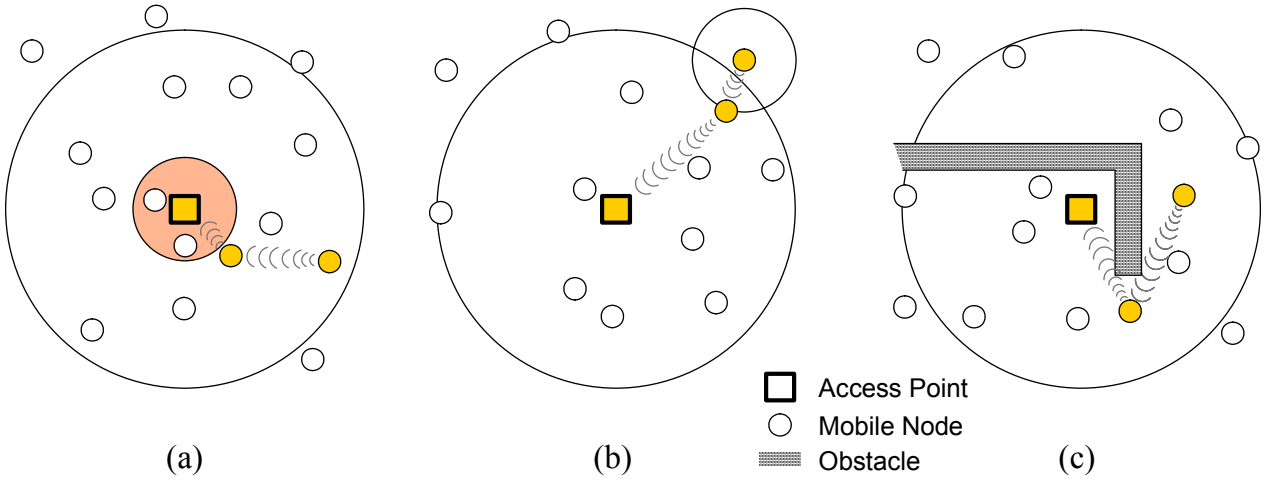


Figure 6-3: Example benefits of multiple-hop wireless access: (a) increases capacity; (b) increases the access range; and (c) helps communicate in the presence of obstructions.

Considering the other cases in Figure 6-3, it is clear that many nodes would benefit, in terms of availability and reliability of Internet access, if multiple wireless hops were permissible. Figure 6-3(b) depicts how this increases availability of wireless Internet access to nodes situated beyond the AP coverage and provides reliability of Internet access for nodes that may drift in and out of coverage range. Multiple intermediary hops can conceivably extend the AP connectivity to several times the nominal range of AP transmission coverage.

Lastly, the MANET paradigm also helps in scenarios where obstacles or noise degrade or eliminate the connectivity between a user and an AP, even though their spatial distance may be relatively small, as in Figure 6-3(c).

Hybrid ad hoc wireless networks examples:

- Sensor networks
- Cellular multi-hop networks
- Mesh/rooftop networks: wireless fast Internet access
- Vehicular networks on roads

6.3 Sensor Networks

Wireless sensor networks are ad hoc networks whose nodes are equipped with sensors and a wireless transceiver, designed to monitor events and deliver information on events to "sinks"—gateways to other networks or capable of processing such information. Sensor network characteristics:

- Very large number of battery-powered nodes
- Simple, inexpensive devices, with limited memory and computing capabilities
- Short range radio, low data rate of the communication channel, low traffic
- Special type of traffic: event driven, messages rather than connections; from and to sinks

Sensor networks appear to be, at least short-term, the most viable application of ad hoc networks. Embedded in homes, offices, cars, and factories, thousands of tiny computerized sentries could track inventories, perform object and person localization, monitor electricity use, and even detect ground vibrations and toxic gases to provide early warning of earthquakes and chemical spills. By designing ultralow-power, small size hardware and smart networking software for sensors makes possible to deploy them anywhere, anytime. Sensor networks are already being used to automate and report meter readings, detect carbon monoxide and turn on ventilation fans in parking garages, and track temperature-sensitive perishables like food, drink, and medicines in transport. Sensor(s) can be mounted in the back of the shipping container, measuring temperature and humidity every few minute. Upon arrival at the loading dock, the nodes send their tracking history to a fixed node, which forwards it to the receiving company's computer to store and process a detailed record. Traffic applications envision having sensors in streetlights alert customers to weather or traffic problems via cell phones.

Key requirements can be summarized as:

- Monitoring geographic region
- Unattended operation
- Long lifetime

Therefore, key to practical sensor networks is energy efficiency: nodes may need to work for years without access to fresh batteries. Sources are used such as solar and wind energy as well as piezoelectric generators in vibrating environments (say, a rattling heating duct or a bumpy truck ride).

6.3.1 Data Gathering and Aggregation

A common paradigm is such that all nodes periodically produce relevant information about their vicinity. Data is conveyed to an information sink for further processing.

Data might be processed as it is routed to the information sink to reduce the amount of traffic. Different sensor nodes partially monitor the same spatial region. Hence, we can exploit spatial and/or temporal data correlation. Issue: coding scheme.

Thus, the objective is to find a routing scheme and a coding scheme to deliver data packets from all nodes to the sink such that the overall energy consumption is minimal.

6.3.2 Commercial Sensor Networks

Millennial Net, Inc. (http://www.millennial.net/) is an MIT spinoff.

Crossbow Technology, Inc. (http://www.xbow.com/) is a UC Berkeley spinoff which builds battery-powered wireless sensors called "motes" that keep getting smaller and smarter. The motes are based on the "Smart Dust" concept and utilizing UC Berkeley's TinyOS. By early 2005, they expect to have a version about the size of a bottle cap that can sense chemicals or the presence of vehicles and perhaps even take photographs. Eventually smart dust could be as small as a grain of sand and operate for years at a time without new batteries.

Another UC Berkeley spinoff, Dust Networks (http://www.dustnetworks.com/), develops and markets the SmartMesh networking system. If several handfuls of the motes sensors are dropped in a nine-square-mile area, they can communicate to one another and then transmit their collective assessment to a main computer, or even a PalmPilot. The company is selling dust as a way to monitor almost anything: the efficiency of refrigerators, the strength of a military convoy, and the activities of people.

ZigBee Alliance

http://www.zigbee.org/

ZigBee and IEEE 802.15.4 Resource Center

http://www.palowireless.com/zigbee/

Ember -- Embedded Wireless Networking

http://www.ember.com/

6.4 Mobile Ad-hoc Networks

6.4.1 Why MANETs?

MANETs impose great complexity and many researchers feel they are of only theoretical interest. A natural question would be, why not provide the spontaneity of ad hoc networking through single hop wireless networking (as in a giant 802.11 LAN) or deploy base station technology on a massive scale? Such solutions would obviate the need for complex multi-hop routing protocols for mobile nodes.

The justification for continued intensive research of multi-hop mobile ad hoc network routing is based on three well-documented performance advantages that the multi-hop paradigm has over the single-hop solution. First, there is the feature of *adaptability* (or *survivability*), which is crucial for battlefield scenarios [Lauer *et al*. 1984]. By deploying a multi-hop packet forwarding network, packets can be routed around obstructions or areas captured by enemy units. Second, there is the advantage of *spatial reuse* [Kleinrock & Silvester 1987]. If between base stations there is a large number of nodes for which peer-to-peer communications are prevalent, then packet forwarding over multiple hops via small radii transmissions will exploit spatial reuse and maximize throughput. Third, depending on transceiver power specifications, packet forwarding via multiple small radii transmissions may *require less energy* than a single large radius transmission for peer-to-peer communications [*]. The energy savings afforded by multi-hop packet forwarding would help conserve battery resources of mobile nodes.

[*]M. B. Srivastava, "Energy efficiency in mobile computing and networking," MobiCom Tutorial, August 2000.

For example, power is consumed while sending, receiving, and overhearing. For a network of n fully connected nodes, if e_t is power consumed for transmission and e_r is power consumed for reception, then instead of $e_t + e_r$ units of power consumption, $e_t + (n-1) \cdot e_r$ units is consumed by the network as a whole.

Although there is currently little consumer demand for multi-hop ad hoc networking, these three advantages indicate that the multi-hop paradigm has the *potential* to greatly improve ad hoc networking performance. To exploit these advantages, however, research of mobile wireless communication systems and mobile multi-hop ad hoc network routing protocols must continue to advance.

[Tschudin *et al*. 2003] about the prospects of real-world MANET applications.

ITS-Radio Services = DSRC (in the UNII range, 5.85-5.925 GHz)

The use of non-voice radio techniques to transfer data over short distances:

– Between roadside and mobile radio units

– Between mobile units

– Between portable and mobile units

to perform operations related to the improvement of traffic flow, traffic safety and other intelligent transportation service applications in a variety of public and commercial applications.

– Control (Routing) is allowed between beacons

What is "Short Range"?

– Short range is 100 m to 1000 m for most ITS applications

DSRC systems may also transmit status and instructional messages related to the units involved.

Internet to the vehicle (car, vanpool, bus, and train) is critical for ITS to really happen. DSRC is the right technology.

(www.ttc.or.jp/e/link/gsc6/contents/RAST_49-ppt.pdf)

The installation and maintenance of a backhaul network for access points is a key factor in the cost of cell-based wireless networks and will likely remain so in the foreseeable future [Webb 2001, pp. 93]. The commercial sector has already recognized the MANET paradigm as a less-costly wireless Internet access approach and has achieved some limited deployment.

The Metricom Ricochet network was a "pole-top" packet radio network deployed in the mid to late 1990s [Chesire & Baker 1996; Tang & Baker 1999]. Mobile nodes accessed the network via fixed-position pole-top repeater nodes some of which also functioned effectively as Internet access points. Only repeater nodes were permitted to function as packet forwarders, this therefore being a very narrow implementation of the MANET paradigm. Although the concept of fixed-position relays is potentially valuable since they can rely on electric utility network for energy, such deployment may be expensive or impossible due to restrictive real estate privileges.

MeshNetworks, Inc. [MeshNetworks 2003] goes beyond the Ricochet architecture by allowing mobile nodes to forward packets. MeshNetworks offers commercially available MANETs for metropolitan and office environ-ments, the latter being called MeshLAN. The work proposed here assumes a network architecture similar to MeshLAN. There are no advertised QoS and power management features of their products.

Rafe Needleman, PacketHop's Networkless Network (www.PacketHop.com)

http://www.alwayson-network.com/comments.php?id=3330_0_8_0_C

6.5 Community Networks

Also known as Rooftop Networks, see example in Figure 6-4. Characteristics:

- Fixed/stationary rather than mobile
- Very large number of nodes
- Multi-hop wireless

[Schuler 1996; Flickenger 2001]

Currently envisioned applications include:

- Inexpensive (shared) broadband Internet
- Ubiquitous access (one "true" network)
- Medical and emergency response
- Gaming

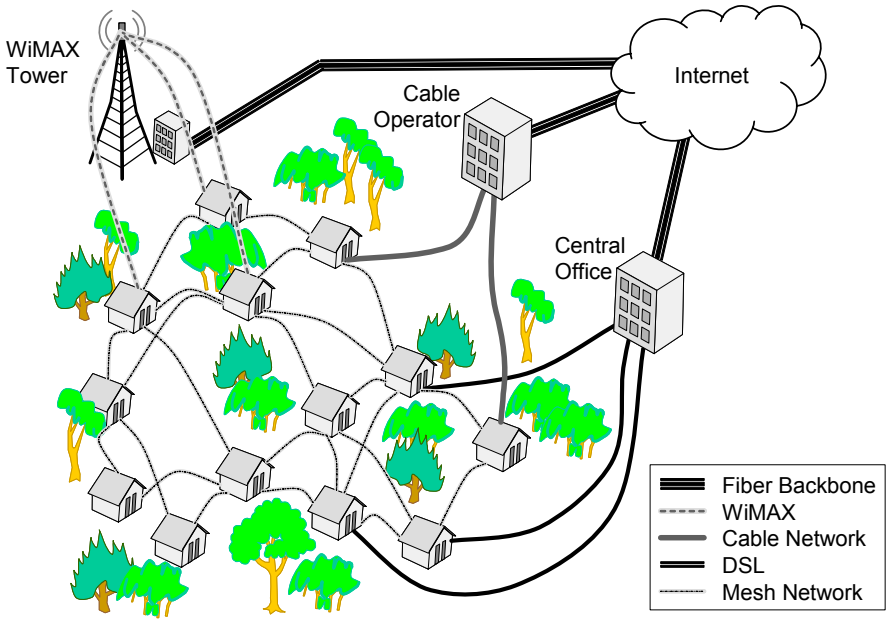


Figure 6-4: Example of a community wireless mesh network.

- Neighborhood video surveillance
- Shared resources
  - Media repository (neighborhood TiVo)
  - Distributed backup

Seattle Wireless

http://www.seattlewireless.net/

Seattle Wireless is a not-for-profit effort to develop a wireless broadband community network in Seattle. Our use of inexpensive wireless technology gets around the artificial bottlenecks imposed by the local telco which prevent true, inexpensive, Metro Area Networking. They are using widely available, standards-based RF technology operating in license free frequency, to create a free, locally owned wireless backbone. This is a MetropolitanAreaNetwork (not just a "wireless LAN" in your home or business) and a community-owned, distributed system (NYASPTWYOMB - not yet another service provider to whom you owe monthly bills). The wireless technology used by the members of the Seattle Wireless network creates the first telecommunications infrastructure that is not only inexpensive, but widely available and easily used so that it is now truly possible for a network to grow from the grass roots of our community based upon on a wonderful combination of self interest and community spirit.

MIT Roofnet

http://www.pdos.lcs.mit.edu/roofnet/

MIT Roofnet is an experimental rooftop wireless network testbed for the Grid Ad-Hoc Networking Project in development at MIT Laboratory for Computer Science's Parallel and Distributed Operating Systems group. The goal of our project is to build a production-quality self-organizing network capable of providing Internet service while researching scalable routing protocols.

6.6 Cognitive (Software) Radio

As the density and variety of wireless devices continues to increase, technology leaders are increasingly voicing concerns about the currently static spectrum management. There is a demand for free markets for spectrum, more unlicensed bands, new services, etc. Static spectrum allocation (Figure 2-1) is a slow and expensive process that cannot keep up with technology advances. There is already anecdotal evidence of wireless LAN spectrum congestion, and with the proliferation of wireless devices, such as home gadgets, sensors, and pervasive computing, this is bound to become much worse. With proliferation, there will also be increasing need to have interoperable radios that can form cooperating networks across multiple PHYs (80.11a,b,g,n, UWB, 802.16, 4G, etc.). There are different approaches to tackle the spectrum problem, including

- Dynamic centralized allocation methods
- Distributed coordination with spectrum etiquette—the employment of simple coordination rules
- Incentive based cooperation—packets are sent including access tokens with pricing determined by congestion
- Cognitive radio bands—agile/smart radios capable of adaptive strategies for interference avoidance
- Some researches also advocate an open access approach—technology is expected to evolve towards co-existence

The term "cognitive radio" is used to denote new generation of adaptive wireless devices capable of dynamic spectrum coordination[5]. Such radios will learn and autonomously perform "cognitive" functions, thanks to a form of intelligence that comes from their ability to be defined and upgraded using software. Software-defined radio (SDR) is the foundation upon which the cognitive radio will be built. Baseline capability includes spectrum scanning and frequency agility; fast adaptation of transmitted signal to fit into changing radio environment; capable of higher-layer spectrum etiquette or negotiation protocols; may also participate in ad-hoc networks

[5] The terms cognitive radio and software-defined radio were coined by Joseph Mitola, chief scientist at the Defense Advanced Research Projects Agency. More info at: http://ourworld.compuserve.com/homepages/jmitola/

formed with other cognitive radios; interoperability with multiple radio technologies based on SDR capabilities.

In an SDR, generic hardware provides the programmable silicon (or general-purpose processor) and other parts needed to transmit and receive radio signals on any appropriate frequencies. This means that one basic device could have the ability to transmit and receive signals in multiple frequency bands, using multiple link-layer protocols. For example, a single radio might be able to transmit and receive on frequencies currently regulated for cell phone communications, digital downloads (such as e-mail), and even marine communications, such as those used to receive permission to enter or leave the locks on a well-traveled river.

Cognitive-radio technology enables the radio itself to learn, allowing it to perform such cognitive functions as identifying and using empty spectrum to communicate more efficiently. Cognitive radios will sense and adapt their behavior according to the environment in which they operate. Once you have a device in which the software implements the protocols programmed for it, the radio can become smart and alert and can negotiate with its environment. In one potential commercial application, a cognitive radio learns about various services of interest to its user by being aware of its user's activities. It knows how to find those services and knows the likelihood that some services will be of interest to its user (the subscriber) in the immediate area.

For example, a cognitive radio could be aware of a Bluetooth network and what is available and of interest to its user within the Bluetooth service zone. It could also be aware of what is available in wireless-LAN range, cell phone range and so on.

Hardware modules for an SDR include antennas, analog front ends (which convert radio frequency, or RF, signals into intermediate signals), and matching digital baseband modules. It's possible that some or all of the hardware parts could be precertified by the FCC for use in radios with modular designs. A modular approach and generic hardware base will significantly reduce the burden on manufacturers who want to enable different feature sets for different radio environments.

6.7 Summary and Bibliographical Notes

Solutions to Selected Problems

Problem 2.1 — Solution

Problem 2.2 — Solution

Input signal, $S_{in} = S_{out} / G = 0.4$ mW / 40 = 10 μW

$(S/N)_{in} = S_{in} / N_{in} = 10\ \mu$W / 20 pW = 500,000 ≈ 57 dB

Problem 2.3 — Solution

Problem 2.4 — Solution

Problem 2.5 — Solution

Problem 2.6 — Solution

Problem 2.7 — Solution

Problem 2.8 — Solution

We know that $P_e = Q\left(\sqrt{2 \cdot E_b / N_0}\right)$ and the requirement is that BER should be better than 10^{-5}. From Figure 2-9, we can read that this requirement implies that $E_b / N_0 \geq 10$ dB. Since we also have rms, $\sigma = 100$ mV

$(V/\sigma)^2 = 2 \cdot E_b / N_0 \geq 20$ dB

From here it follows that the minimum voltage separation between the voltage levels of each symbol is 0.9 V.

Problem 2.9 — Solution

Problem 2.10 — Solution

Problem 2.11 — Solution

Transmission rate $D = R \times n$, where n is the number of bits per symbol. Then, $M = 2^n$ and $n = \log_2 M$. After substituting n, we have $D = R \times \log_2 M$.

Problem 2.12 — Solution

Problem 3.1 — Solution

Problem 3.2 — Solution

— Solution

Problem 3.4 — Solution

Problem 3.5 — Solution

With $C = 2400$ bps channel and $L = 200$ bits for packet size, the slot size is:

$$t_{slot} = L / C = 200/2400 = 83.33 \text{ msec}$$

Use the formula for maximum throughput $S_{max} = (m_{max} \cdot \lambda) / \mu$.

Channel capacity (service rate) for 2400 bps channel is:

$$\mu = C / L = 2400/200 = 12 \text{ packets/sec}$$

$$\lambda = 1 \text{ packet} / 120 \text{ sec}$$

Maximum throughput $S_{max} = (m_{max} \cdot \lambda) / \mu = 1/e \cong 0.36 \Rightarrow m_{max} \cong 518$.

For 500-bit packet, $\mu = 2400/500 = 4.8$ packets/sec, $m_{max} \cong 207$.

For 200-bit packet, but λ = 1 packet / 3 min, $m_{\max} \cong 777$.

For 4800-bps channel, 200-bit packet and λ = 1 packet / 120 sec, $m_{\max} \cong 1036$.

Problem 3.6 — Solution

Assume that the average number of transmissions per slot on the channel is G.

The probability of at least one collision is a probability of two or more arrivals in two slots (for pure Aloha), which is:

$$P_{coll} = 1 - e^{-2G}$$

The average number of retransmissions is

$$\overline{K} = G \cdot P_{coll} = G \cdot \left(1 - e^{-2G}\right)$$

Another way to solve the problem is as follows. The average number of retransmissions is the difference between the total number of transmission attempts on the channel, G, and the effective throughput, S,

$$\overline{K} = G - S = G - G \cdot e^{-2G} = G \cdot \left(1 - e^{-2G}\right)$$

The probability of a collision is the average number of retransmissions per total arrival rate on the channel

$$P_{coll} = \frac{\overline{K}}{G} = 1 - e^{-2G}$$

Problem 3.7 — Solution

The state "0" can transition to any other state except state "1," i.e., $P_{01} = 0$, because it takes at least two transmissions to collide. The probabilities P_{ii} represent the random event of whether or not a station will transmit, and also if multiple backlogged stations transmit and collide. If one or more stations from the unbacklogged pool are involved in the collision, the Markov chain transitions into a higher state.

Problem 3.8 — Solution

The expected number of transmissions is e^G.

Problem 3.9 — Solution

Problem 3.10 — Solution

(a) Probability of success is e^{-G}.

(b) See the solution for Problem 3.8. The probability of exactly K collisions and then a success is

$$(1 - e^{-G})^k \cdot e^{-G}$$

Problem 3.11 — Solution

Problem 3.12 — Solution

(a) Probability of success

$$P_{\text{success}} = \binom{n}{1} \cdot (1 - q_t)^{n-1} \cdot q_t = n \cdot (1 - q_t)^{n-1} \cdot q_t$$

(b)

$$\frac{dP_{success}}{dq_t} = n \cdot (1 - q_t)^{n-1} - n \cdot (n-1) \cdot (1 - q_t)^{n-2} \cdot q_t = 0$$

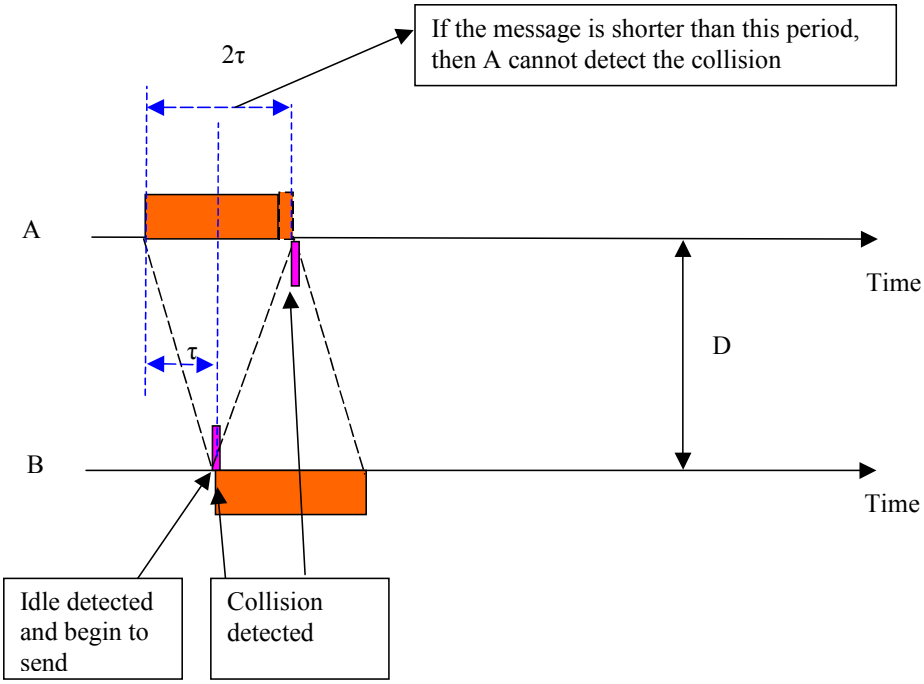
$$\Rightarrow \qquad q_t = \frac{1}{n}$$

(c)

"Worst case optimal" value for q_t is $q_t = 1/200 = 0.005$

Problem 3.13 — Solution

Assume CSMA/CD. The problem that needs to be solved is that a node that started transmission, continues to transmit bits long enough to be able to detect that another node concurrently started a transmission. If t is the time it takes for a signal to reach the farthest node in a CSMA/CD network, the contention period should be equal to $2t$. Assume A starts transmission at T_0. In the worst case, the farthest station begins its transmission just before $T_0 + t$, which then requires t time units to reach A. As a consequence, only after $2t$ time units A will detect that some other node started as well.



S = 8 bits * 64= 512 bits

R = 10 M bps / 100 Mbps / 1000 Mbps

$\tau = D / v$

$v = 2*10^8$ m/s

➔$2\tau = S/R = 2D/v$ ➔ $D = S*v/(2*R)$

Problem 3.14 — Solution

Problem 3.15 — Solution

Problem 3.16 — Solution

There are three stations (n = 3). Let q_r be the probability of transmitting a packet in a slot for every station. For the sake of having a concrete numerical value to work with, we assume $q_r = 0.4$, but any other value is acceptable. There are eight situations as follows:

(1) No station transmits: $P_0 = (1-q_r)^3 = (0.6)^3 = 0.216$

(2) Only ***A*** transmits: $P_A = q_r \cdot (1-q_r)^2 = (0.4)\times(0.6)^2 = 0.144$

(3) Only B transmits: $P_B = q_r \cdot (1-q_r)^2 = (0.4)\times(0.6)^2 = 0.144$

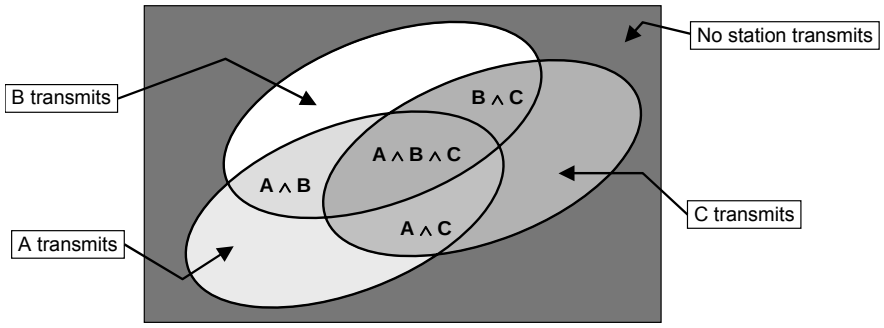
(4) Only C transmits: $P_C = q_r \cdot (1-q_r)^2 = (0.4) \times (0.6)^2 = 0.144$

(5) ***A*** and B transmit, but C does not transmit: $P_{AB} = q_r{}^2 \cdot (1-q_r) = (0.4)^2 \times (0.6) = 0.096$

(6) B and C transmit, but ***A*** does not transmit: $P_{BC} = q_r{}^2 \cdot (1-q_r) = (0.4)^2 \times (0.6) = 0.096$

(7) C and ***A*** transmit, but B does not transmit: $P_{CA} = q_r{}^2 \cdot (1-q_r) = (0.4)^2 \times (0.6) = 0.096$

(8) ***A***, B and C all transmit: $P_{ABC} = q_r{}^3 = (0.4)^3 = 0.064$



(a) Network perspective:

Probability of channel idle P_{idle}, a successful transmission $P_{success}$ and collision $P_{collision}$

$$P_{idle} = \Pr[no\_station\_transmits] = P_0 = 0.216$$

$$P_{success} = \Pr[only\_one\_station\_transmits] = P_A + P_B + P_C = 0.432$$

$$P_{collision} = \Pr[at\_least\_two\_stations\_transmit] = P_{AB} + P_{BC} + P_{CA} + P_{ABC} = 0.352$$

Or use Bernoulli formula

$$P_{idle} = \Pr[no\_station\_transmits] = Q_r(0,3) = \binom{3}{0} \cdot (1-q_r)^3 \cdot q_r^0 = (1-q_r)^3 = (0.6)^3 = 0.216$$

$$P_{success} = \Pr[only\_one\_station\_transmits] = Q_r(1,3) = \binom{3}{1} \cdot (1-q_r)^{3-1} \cdot q_r^1$$

$$= \binom{3}{1} \cdot (1-q_r)^2 \cdot q_r = 3 \times (0.6)^2 \times 0.4 = 0.432$$

$$P_{collision} = \Pr[at\_least\_two\_stations\_transmit] = Q_r(2,3) + Q_r(3,3) = 1 - P_{idle} - P_{success} = 0.352$$

(b) Station Perspective:

b.1 When station A is in its backoff stage, the probabilities that it will observe certain channel state are as follows:

$$P_{idle\_backoff} = \frac{\Pr[no\_station\_transmits]}{\Pr[A\_doesn't\_transmit]} = \frac{P_0}{1-(P_A+P_{AB}+P_{CA}+P_{ABC})} = \frac{P_0}{1-q_r} = \frac{0.216}{0.6} = 0.36$$

$$P_{success\_backoff} = \frac{\Pr[only\_one\_ststion\_transmits]}{\Pr[A\_doesn't\_transmit]} = \frac{P_B+P_C}{1-(P_A+P_{AB}+P_{CA}+P_{ABC})} = \frac{P_B+P_C}{1-q_{rc}} = 0.48$$

$$P_{collision\_backoff} = \frac{\Pr[at\_least\_two\_stations\_transmit]}{\Pr[A\_doesn't\_transmit]} = \frac{P_{AB}+P_{BC}+P_{CA}+P_{ABC}}{1-(P_A+P_{AB}+P_{CA}+P_{ABC})} = 0.16$$

Or use Bernoulli formula

$$P_{idle\_backoff} = \frac{\Pr[no\_node\_transmits]}{\Pr[A\_doesn't\_transmit]} = \frac{\Pr[A\_doesn't\_transmit\_and\_no\_station\_transmits]}{\Pr[A\_doesn't\_transmit]}$$

$$= \frac{\Pr[A\_doesn't\_transmit]\times\Pr[none\_of\_n-1\_stations\_transmits]}{\Pr[A\_doesn't\_transmit]}$$

$$= \Pr[none\_of\_n-1\_nodes\_transmits]$$

$$= \binom{n-1}{0}\cdot(1-q_r)^{n-1} = \binom{2}{0}\cdot(1-q_r)^2 = (0.6)^2 = 0.36$$

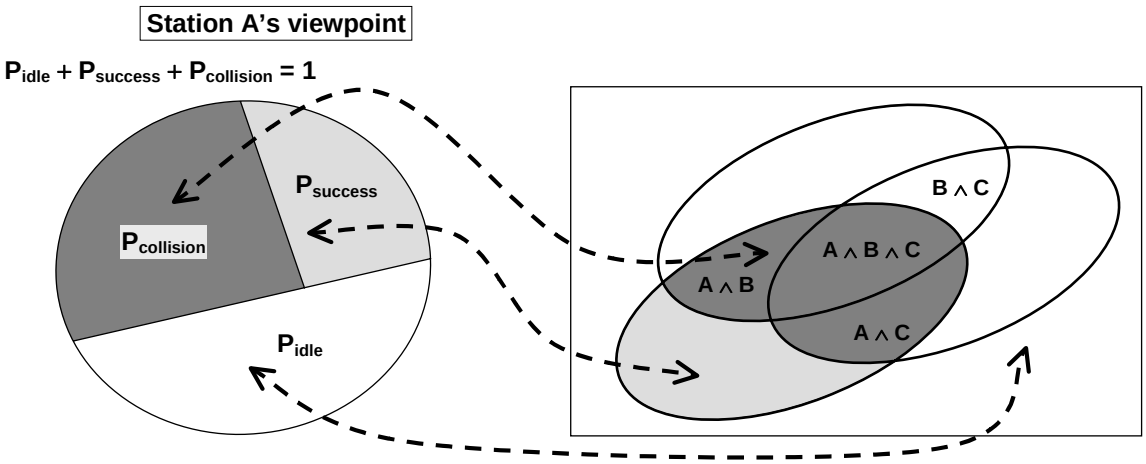
$$P_{success\_backoff} = \frac{\Pr[only\_one\_station\_transmits]}{\Pr[A\_doesn't\_transmit]}$$

$$= \frac{\Pr[A\_doesn't\_transmit]\times\Pr[only\_one\_of\_n-1\_stations\_transmits]}{\Pr[A\_doesn't\_transmit]}$$

$$= \Pr[only\_one\_of\_n-1\_stations\_transmits]$$

$$= \binom{n-1}{1}\cdot(1-q_r)^{n-2}\cdot q_r = \binom{2}{1}\cdot(1-q_r)\cdot q_r = 2\times 0.6\times 0.4 = 0.48$$

$$P_{collision\_backoff} = 1 - P_{idle\_backoff} - P_{success\_backoff} = 0.16$$



b.2 When station A is already transmitting a packet, it can observe only success or collision

$$P_{success\_xmit} = \frac{\Pr[only\_A\_transmits]}{\Pr[at\_least\_A\_transmits]} = \frac{P_A}{P_A + P_{AB} + P_{CA} + P_{ABC}} = \frac{P_A}{q_c} = 0.36$$

$$P_{collision\_xmit} = \frac{\Pr[A\_and\_others\_transmit]}{\Pr[at\_least\_A\_transmits]} = \frac{P_{AB} + P_{CA} + P_{ABC}}{P_A + P_{AB} + P_{CA} + P_{ABC}} = 0.64$$

Or use Bernoulli formula

$$P_{success\_xmit} = \frac{\Pr[only\_A\_transmits]}{\Pr[at\_least\_A\_transmits]}$$

$$= \frac{\Pr[at\_least\_A\_transmits] \times \Pr[none\_of\_n-1\_stations\_transmit]}{\Pr[at\_least\_A\_transmits]}$$

$$= \Pr[none\_of\_n-1\_stations\_transmit]$$

$$= \binom{n-1}{0} \cdot (1-q_r)^{n-1} = \binom{2}{0} \cdot (1-q_r)^2 = (0.6)^2 = 0.36$$

$$P_{collision\_xmit} = \frac{\Pr[A\_and\_others\_transmit]}{\Pr[at\_least\_A\_transmits]} = 1 - P_{success\_xmit}$$

(c) Relation between network and station perspective:

The difference between these two perspectives is that the network (outside observer) has no knowledge of what any of the three stations is doing, whereas the station knows how itself is behaving. The network perspective must be conditioned through the eyes of the station when discerning how station perceives the network.

Conditional probabilities (for an individual station) relate to the channel probabilities like so:

$$\widetilde{P}_{success} = \frac{P_{success}}{P_{success} + P_{collision}} \qquad \widetilde{P}_{collision} = 1 - \widetilde{P}_{success}$$

Problem 3.17 — Solution

Problem 3.18 — Solution

Problem 3.19 — Solution

The probability of success is the sum of the probabilities of success in each slot before slot K:

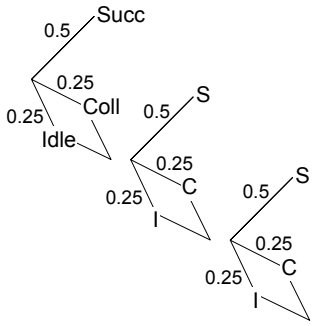
$$P_{succ} = m \cdot p_1 \cdot (1 - p_1)^{m-1} + m \cdot p_2 \cdot (1 - p_1 - p_2)^{m-1} + \ldots + m \cdot p_{K-1} \cdot (1 - p_1 - \ldots - p_{K-1})^{m-1}$$

$$= m \cdot \sum_{s=1}^{K-1} p_s \cdot \left(1 - \sum_{r=1}^{s} p_r\right)^{m-1}$$

Problem 3.20 — Solution

(a)

Probability of transmission p = ½, the success happens if either station transmits alone:



$$P_{\text{success}} = \binom{2}{1} \cdot (1 - p)^{2-1} \cdot p = 2 \times ½ \times ½ = 0.5$$

(b)

The first transmission ends in collision if both stations transmit simultaneously

$$P_{\text{collision}} = \binom{2}{2} \cdot (1 - p)^{0} \cdot p^2 = 1 \times 1 \times ¼ = 0.25$$

Since the waiting times are selected *independent* of the number of pervious collisions (i.e., the successive events are independent of each other), the probability of contention ending on the second round of retransmissions is

$$P_{\text{collision}} \times P_{\text{success}} = 0.5 \times 0.25 = 0.125$$

(c)

Similarly, the probability of contention ending on the third round of retransmissions is

$$P_{\text{collision}} \times P_{\text{success}} \times P_{\text{success}} = 0.5 \times 0.25 \times 0.25 = 0.0312$$

(d)

Regular nonpersistent CSMA with the normal binary exponential backoff algorithm works in such a way that if the channel is busy, select a random period to wait; when the waiting period expires, sense the channel; if idle, transmit; if busy; wait again for a random period, selected from the *same range*; and so on until the channel becomes idle and transmission occurs.

If the transmission is successful, that's it.

If the transmission ends in collision, *double the range* from which the waiting period is drawn and repeat the above procedure.

Unlike the regular nonpersistent CSMA, in the modified nonpersistent CSMA there are only two choices for waiting time. Therefore, half of the backlogged stations will choose one way and the other half will choose the other way.

In conclusion, regular nonpersistent CSMA performs better than the modified one under heavy loads and the modified algorithm performs better under light loads.

Problem 3.21 — Solution

Problem 3.22 — Solution

The analysis is performed for the steady state—we assume that the average number of backlogged nodes remains constant. Although at one time a node may enter an empty state from backlogged state, there will be another node transitioning in the opposite way, so that the system is maintained in a balanced state, with constant $\overline{n} = E[n] > 0$.

As a consequence of non-zero $\overline{n}$, the channel is working in a saturated traffic condition. That is, since at anytime there is at least one backlogged node, the channel is always "busy"—from the channel point of view packets always arrives regardless from which node. (Note that saturated channel does not imply saturated stations.)

The problem statement gives a system of linear equations $\mathbf{b} = \mathbf{P} \cdot \mathbf{b}$, where $\mathbf{P}$ is the transition probability matrix in steady state [Bertsekas & Gallagher, 1992]. Given the constraint that $\widetilde{p}_{succ} + \widetilde{p}_{coll} = 1$ and $\Sigma_k \, b_k = 1$, we can solve for $\mathbf{b}$. Thus, $\mathbf{b}$ is the eigenvector of the transition matrix with eigenvalue 1:

$$(\mathbf{P}-\mathbf{I})^{\mathrm{T}}\cdot\mathbf{b}=\begin{bmatrix}\widetilde{p}_{succ}-1 & \widetilde{p}_{succ} & \widetilde{p}_{succ} & \dots & \widetilde{p}_{succ} & 1\\ \widetilde{p}_{coll} & -1 & 0 & \dots & 0 & 0\\ 0 & \widetilde{p}_{coll} & -1 & \dots & 0 & 0\\ \dots & \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \widetilde{p}_{coll} & -1 & 0\\ 0 & 0 & \dots & 0 & \widetilde{p}_{coll} & -1\end{bmatrix}\cdot\begin{bmatrix}b_0\\ b_1\\ b_2\\ \dots\\ b_{\ell-1}\\ b_\ell\end{bmatrix}=\begin{bmatrix}0\\ 0\\ 0\\ \dots\\ 0\\ 0\end{bmatrix} \qquad \text{(S-3.1)}$$

The probabilities of success and collision, $\widetilde{p}_{succ}$ and $\widetilde{p}_{coll}$ respectively, are given in the problem statement, and these two equations combined will yield b_k. [These probabilities do not depend on the number of collisions k the packet already suffered, since they depend on the other stations, rather than on the internal state of the station under consideration.]

Use Matlab or Mathematica software to solve Eq. (S-3.1) for vector **b** and plot it as a graph.

Problem 3.23 — Solution

Packets are discarded when the retry limit ℓ is reached. The drop probability is

$$p_{drop}=1-\sum_{k=0}^{\ell}q_k=1-\sum_{k=0}^{\ell}\left(\widetilde{p}_{coll}^{\,k}\cdot\widetilde{p}_{succ}\right)=1-\widetilde{p}_{succ}\cdot\frac{\widetilde{p}_{coll}\cdot\left(1-\widetilde{p}_{coll}^{\,\ell}\right)}{1-\widetilde{p}_{coll}}=\left(\widetilde{p}_{coll}\right)^{\ell+1}$$

by the complementary probability property.

These discarded packets waste the channel capacity and in turn reduce the throughput of the system. For a discarded packet, the channel time it wastes is:

$$t_{drop}=\sum_{k=0}^{\ell}t_{coll,k}$$

where $t_{coll,k}$ is the collision time at backoff stage k.

Problem 3.24 — Solution

(a) Hidden terminal:

A and C want to transmit to B.

A sends an RTS to B. [A $\xrightarrow{RTS}$ B C]

B replies with a CTS [A $\xleftarrow{CTS}$ B $\xrightarrow{CTS}$ C]

C overhears a CTS and realizes that B is busy, so refrains from transmitting.

(b) Exposed terminal:

B wants to transmit to A. C wants to transmit to D.

B sends an RTS to A, C hears B's RTS. [A $\xleftarrow{RTS}$ B $\xrightarrow{RTS}$ C]

A replies with a CTS, which B hears but C *does not*. [A $\xrightarrow{CTS}$ B C]

C concludes that D cannot hear the CTS either, and sends to D an RTS after waiting for A's CTS to reach B. [B $\xleftarrow{RTS}$ C $\xrightarrow{RTS}$ D]

If D cannot hear A's CTS, D replies to C immediately with CTS.

If D can hear A's CTS (but C cannot hear A, since A is hidden from it and that is why it decided to send RTS to D), D would reply with a CTS only when A is finished.

In either case, C→D transmission is not unnecessarily affected by B→A transmission.

Problem 3.25 — Solution

Suppose both stations collide on their first RTS. One of them, say *A*, will pick a lower random number timeout value than the other—call it *B*. So, *A* goes next, and sends a steady stream of packets to the receiver. At some point *B* times out and sends an RTS. This will collide with *A*'s packet. *A* will choose a random backoff interval in the range [0, 1], whereas *B*, which is seeing its second consecutive collision, will choose a random backoff interval in the range [0, 2]. It is likely that *B* will choose longer timeout, and that *A* will resume its transmission. In the next collision, *A* chooses a timeout in the range [0, 1], but *B* chooses a timeout in the range [0, 4]. Again, *B* is more likely to lose. Eventually, *B* gets shut out, while *A* gets the entire link bandwidth.

Problem 3.26 — Solution

Problem 3.27 — Solution

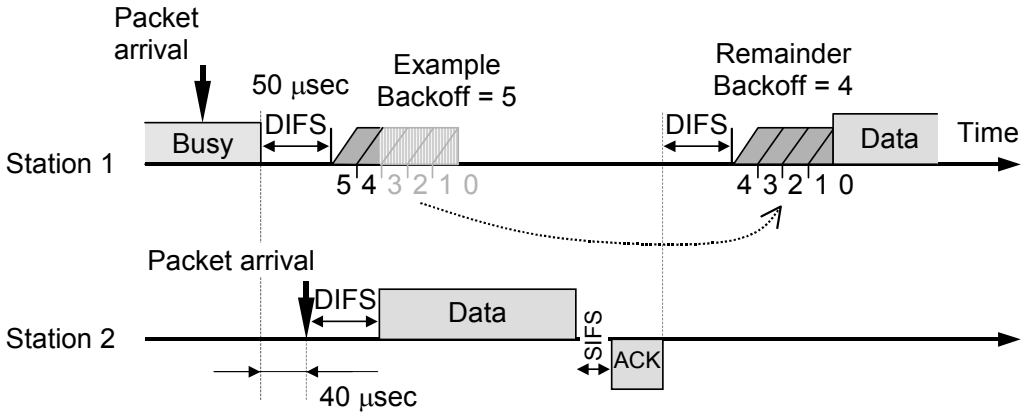
Problem 3.28 — Solution

Problem 3.29 — Solution

Problem 4.1 — Solution

The timing diagrams for transmission of the two newly arrived packets are as shown:

Since the duration of DIFS = 50 μsec (Table 4-2), the packet at Station 2 will arrive during the channel idle period. Hence, it waits only for a DIFS period (no backoff countdown!) and gets transmitted.

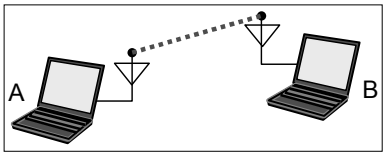


Notice that this is the case regardless of whether the previous channel busy state was due to Station 1's transmission or due to some other station's transmission. Station 1 must enter the backoff countdown state if the channel is busy at the time the new packet arrived. If the distinction was made such that a station does not enter countdown if the previous transmission was from itself, this would be unfair for other stations—this station could seize the channel and transmit for as long as it has something to transmit.

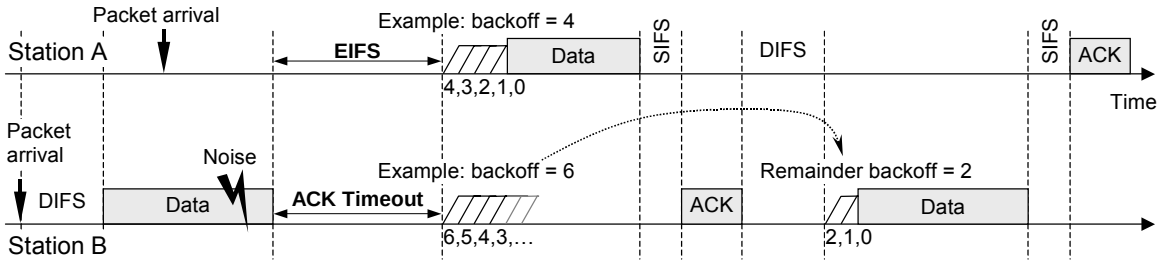
Check also Figures 1 in references [Bianchi, 2000; Carvalho & Garcia-Luna-Aceves, 2003].

Problem 4.2 — Solution

There is no access point since this is an independent BSS (IBSS). Remember that 802.11 stations cannot transmit and receive simultaneously, so once a station starts receiving a packet, it cannot contend for transmission until the next transmission round.



The solution is as follows (see Figure 4-8). We make arbitrary assumption that packet arrives first at *B* and likewise that after the first transmission *A* claims the channel before *B*. Notice that the durations of EIFS and ACT_Timeout are the same.



Problem 4.3 — Solution

Because, in order to distinguish a true transmission from a channel noise, the station must detect a packet being transmitted (at least its preamble).

Problem 4.4 — Solution

The probability of a successful atomic transmission in the basic mode (Data + ACK) is:

$$p = 1 - (1 - p_f) \cdot (1 - p_r)$$

The probability of success after $k - 1$ failures is:

$$s(k) = p^{k-1} \cdot (1 - p)$$

The expected number of retransmissions, assuming that there is no limit for retries is:

$$ETX = \sum_{k=1}^{\infty} k \cdot s(k) = \frac{1}{1 - p}$$

Problem 4.5 — Solution

If you can send shorter frames, the chance that a frame is damaged during transmission is smaller. Fragmentation allows the protocol to fragment frames in order to increase transmission reliability, without letting the network layer be aware of this. With 802.3, frame loss due to noise is rare; with 802.11 it is common.

Problem 4.6 — Solution

Problem 4.7 — Solution

For every successfully received packet we record its arrival time and the instantaneous value of signal (RSSI) and noise levels. Notice that very few packets are actually lost since MAC retries transmission until it is successful or the upper retry bound ℓ is reached.

Note that packet's signal/noise are recorded at different times. According to the standard ([37], p. 150 and 213), RSSI is a one-byte number measured by the physical layer of the energy observed at the antenna used to receive the current PPDU (physical layer frame). It is measured between the beginning of the start frame delimiter (SFD) and the end of the PLCP header error check (HEC), which is 64 bits long for DSSS physical layer. Absolute accuracy of the RSSI reading is not specified and actual implementations are vendor specific. Of course, RSSI comprises both "signal" and noise, since the receiver has no means of separating them.

Noise, on the other hand, is the background RF strength within the receiver bandwidth when no signal/carrier is detected. The IEEE standard [37] does not specify the noise measurement. If we consider the worst-case scenario of the saturated channel, the only silence time for the noise acquisition is during the SIFS (short interframe space) and DIFS (distributed IFS) periods, when there should be no valid signal present on the medium. Thus we speculate that the Wi-Fi vendors implement the noise measurement between the end of the current PPDU and the end of the SIFS period. (It also appears easier to associate the noise measurement with the PPDU *after* its reception rather than the one before.) As a result, RSSI and noise for a received packet are

measured at different times. The SNR during the transmission of packet t is computed as SNR(t) = RSSI(t) – Noise(t).

Problem 4.8 — Solution

Problem 4.9 — Solution

Problem 4.10 — Solution

Problem 4.11 — Solution

Problem 4.12 — Solution

Problem 4.13 — Solution

To simplify the analysis, we make an (incorrect) assumption that the reception range (the distance at which a signal can be decoded well in the receiver) equals to the sensing range.

(a)

Covered STAs	Hidden STAs	Channel state observed by A	Channel state observed by AP
0 transmit	**0** transmit	**IDLE**	**IDLE**
	1 transmit	**IDLE** until (*i*) it hears CTS or (*ii*) counts down to zero and transmits. After that: **SUCCESS** if (*i*), **COLLISION** if (*ii*)	**SUCCESS**, sends CTS if (*i*) A remains silent; **COLLISION**, remains silent if (*ii*) A transmits
	≥2 transmit	**IDLE**	**COLLISION**, remains silent
	0 transmit	**SUCCESS** sees an RTS in 1 backoff slot followed by CTS	**SUCCESS**, sends CTS

1 transmit	**≥1** transmit	**IDLE** until (*i*) it hears covered STA's RTS or (*ii*) counts down to zero and transmits. After that: **COLLISION**	**COLLISION**, remains silent
≥2 transmit	**≥0** transmit	**COLLISION**	**COLLISION**, remains silent

(b)

Assuming the uniform distribution of all nodes as well as the backlogged ones and the same coverage radius for all stations, an "average" station will have $n_h = \rho \cdot n$ hidden nodes and $n_c = (1-\rho) \cdot n$ covered nodes, where ρ=0.41 as derived in [53].

(c)

A node will observe collision state if at least two covered stations transmit or exactly one covered- and at least one hidden station transmit. The collision state probability is:

$$p_{coll}(n_h) = \left[1 - Q_c(0, n_c) - Q_c(1, n_c)\right] + Q_c(1, n_c) \cdot \left[1 - Q_v(0, n_h)\right]$$

$$= \left[1 - (1-q_c)^{n_c - 1} - (n_c - 1) \cdot q_c \cdot (1-q_c)^{n_c - 2}\right] + (n_c - 1) \cdot q_c \cdot (1-q_c)^{n_c - 1} \cdot \left[1 - (1-q_v)^{n_h - 1}\right]$$

Also,

$$p_{busy}(n_h) = p_{succ}(n_h) + p_{coll}(n_h) \qquad \text{and} \qquad p_{idle}(n_h) = 1 - p_{busy}(n_h)$$

The conditional probabilities of success and collision given that a node transmitted are:

$$\tilde{p}_{succ} = Q_c(0, n_c) \cdot Q_v(0, n_h) = (1-q_c)^{n_c} \cdot (1-q_v)^{n_h}$$

$$\tilde{p}_{coll} = 1 - \tilde{p}_{succ}$$

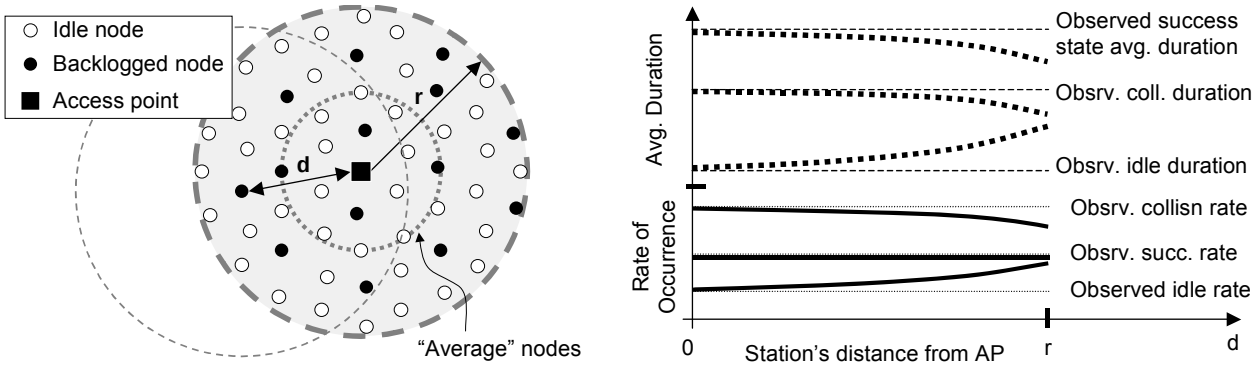


FIGURE for Problem 4.13(d):

(a) All nodes are uniformly distributed in the AP's area of coverage and the nodes further away from the AP have less covered and more hidden nodes. (b) As a result, the times for which they observe different channel states differ, as do the numbers of different states occurrences (shown are qualitative relationships).

(d)

Solution shown in the figure.

Problem 4.14 — Solution

Problem 4.15 — Solution

Problem 4.16 — Solution

Problem 5.1 — Solution

— Solution

References

1. N. Abramson, "The ALOHA System—Another alternative for computer communications," in *1970 Fall Joint Comput. Conf., AFIPS Conf. Proc*. vol. 37, Montvale. NJ: AFIPS Press, pp. 281-285, 1970.
2. A. Acharya, A. Misra, and S. Bansal, "MACA-P: A MAC for concurrent transmissions in multi-hop wireless networks," *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications (PerCom 2003)*, Dallas-Fort Worth, TX, pp. 505-508, March 2003. Also appears as IBM Research Report RC 22528, Online at: http://www.research.ibm.com/
3. A. Acharya, A. Misra, S. Bansal, "A label-switching packet forwarding architecture for multi-hop wireless LANs," *Proceedings of the 5th ACM International Workshop on Wireless Mobile Multimedia (WoWMoM 2002)*, Atlanta, GA, pp. 33-40, September 2002. Also appears as IBM Research Report RC22512, Online at: http://www.research.ibm.com/
4. I. Aldridge, Analysis of Existing Wireless Communication Protocols, COMS E6998-5 course project taught by Prof. M. Lerner, Columbia University, NY, Summer 2000. Online at: http://www.columbia.edu/~ir94/wireless.html
5. M.-S. Alouini and A. J. Goldsmith, "A unified approach for calculating error rates of linearly modulated signals over generalized fading channels," *IEEE Transactions on Communication*, vol. 47, pp. 1324-1334, September 1999.
6. J. S. Belrose, The Sounds of a Spark Transmitter: Telegraphy and Telephony. Online at: http://www.physics.otago.ac.nz/ursi/belrose/spark.html
7. Y. Benkler, "From consumers to users: Shifting the deeper structures of regulation toward sustainable commons and user access," *Federal Communications Law Journal*, vol. 52, no. 3, pp. 561-579, May 2000. Online at: http://www.law.indiana.edu/fclj/pubs/v52/no3/benkler1.pdf
8. J. M. Berger and B. B. Mandelbrot, "A new model for error clustering in telephone circuits," *IBM Journal of Research and Development*, vol. 7, no. 3, pp. 224-235, July 1963.
9. H. L. Bertoni, *Radio Propagation for Modern Wireless Systems*. Prentice Hall PTR, Upper Saddle River, NJ, 2000.
10. D. Bertsekas and R. Gallagher, *Data Networks*. 2nd edition, Prentice Hall, Upper Saddle River, NJ, 1992.
11. D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, Belmont, MA: Athena Scientific, 2002.
12. V. Bharghavan, A. Demers, S. Shenker, and L. Zhang, "MACAW: A media access protocol for wireless LANs," *Proceedings of ACM SIGCOMM'94: Applications, Technologies, Architectures, and Protocols for Computer Communication*, September 1994.
13. G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535-547, March 2000.
14. E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: Information theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619-2692, 1998.
15. S. Chesire and M. Baker, "A wireless network in MosquitoNet," *IEEE Micro*, pp. 44-52, February 1996.

16. M.Chiani, D. Dardari, and M. K. Simon, "New exponential bounds and approximations for the computation of error probability in fading channels," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 840-845, July 2003.

17. S. Corson and J. Macker, "Mobile ad hoc networking (MANET): Routing protocol performance issues and evaluations considerations," *IETF RFC*, No. 2501, January 1999.

18. V. Erceg, L. J. Greenstein, S. Y. Tjandra, S. R. Parkoff, A. Gupta, B. Kulic, A. A. Julius, and R. Bianchi, "An empirically based path loss model for wireless channels in suburban environments," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 7, pp. 1205-1211, July 1999.

19. Federal Communications Commission, Office of Engineering and Technology, Spectrum Inventory Table 137 MHz to 100 GHz. Online at: http://www.fcc.gov/oet/info/database/spectrum/

20. R. Flickenger, *Building Wireless Community Networks*. O'Reilly & Associates, Inc., Sebastopol, CA, 2001.

21. C. L. Fullmer and J. J. Garcia-Luna-Aceves, "Solutions to hidden terminal problems in wireless networks," *Proceedings of ACM SIGCOMM'97: Applications, Technologies, Architectures, and Protocols for Computer Communication*, Cannes, France, 1997.

22. M. S. Gast. *802.11 Wireless Networks: The Definitive Guide*. O'Reilly & Associates, Inc., Sebastopol, CA, 2002.

23. A. J. Goldsmith and S.-G., Chua, "Variable rate variable power M-QAM for fading channels, *IEEE Transactions on Communications*, vol. 46, no. 5, pp. 1218-1230, May 1997.

24. M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *Proceedings of the IEEE INFOCOM*, 2001.

25. P. Gupta, R. Gray, and P. R. Kumar, An Experimental Scaling Law for Ad Hoc Networks," ECE Department, University of Illinois at Urbana-Champaign, May 2001.

26. P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 3, pp. 388-404, March 2000.

27. J. F. Hayes, *Modeling and Analysis of Computer Communications Networks*. Plenum Press, New York, NY, 1984.

28. G. Holland, N. Vaidya and P. Bahl, "A rate-adaptive MAC protocol for multi-hop wireless networks," *Proc. ACM MobiCom'01*, Rome, Italy, pp. 236-250, July 2001.

29. Intersil, "HFA3861B: Direct sequence spread spectrum baseband processor," January 2000.

30. T. Jiang, N. D. Sidiropoulos, and G. B. Giannakis, "Kalman filtering for power estimation in mobile communications," *IEEE Transactions on Wireless Communications*, January 2003. http://www.comsoc.org/livepubs/twc/Public/2003/Jan/151_02twc01-jiang.html

31. A. Kamerman and L. Monteban, "WaveLAN-II: A high-performance wireless LAN for the unlicensed band," *Bell Labs Technical Journal*, pp. 118-133, Summer 1997.

32. G. Keiser, *Local Area Networks*. 2nd edition. McGraw-Hill, New York, NY, 2002.

33. L. Kleinrock and J. Silvester, "Spatial reuse in multihop packet radio networks," *Proceedings of the IEEE*, vol. 75, no. 1, pp. 156-167, January 1987.

34. L. Kleinrock and F. A. Tobagi, "Packet switching in radio channels: Part I—Carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE Transactions on Communications*, vol. COM-23, pp. 1400-1416, December 1975.

35. Y. Y. Kim and S. Li, "Modeling multipath fading channel dynamics for packet data performance analysis," *Wireless Networks*, vol. 6, no. 6, pp. 481-492, 2001.

36. A. Kopke, A. Willig, and H. Karl, "Chaotic maps as parsimonious bit error models of wireless channels," *Proceedings of the Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, vol. 1, pp. 513-523, March/April 2003.

37. J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach Featuring the Internet*. 3rd edition. Pearson Education, Inc. (Addison-Wesley), Boston, MA, 2005.

38. LAN/MAN Standards Committee of the IEEE Computer Society, ANSI/IEEE Std. 802.11—1999 edition, "Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications," 1999. Online at: http://standards.ieee.org

39. LAN/MAN Standards Committee of the IEEE Computer Society, IEEE Std. 802.11a—1999, "Supplement to IEEE standard for information technology telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: High-speed physical layer in the 5 GHz band," September 1999. Online at: http://standards.ieee.org

40. LAN/MAN Standards Committee of the IEEE Computer Society, IEEE Std. 802.11b—1999, "Supplement to IEEE standard for information technology telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-speed physical layer extension in the 2.4 GHz band," January 2000. Online at: http://standards.ieee.org

41. B. P. Lathi, *Modern Digital and Analog Communication Systems*. 2nd edition. Holt, Rinehart and Winston, Inc., Orlando, FL, 1989.

42. G. Lauer, J. Jubin, and J. Tornow, "Survivable protocols for large scale packet radio networks," *Proceedings of the IEEE Globecom '84*, pp. 468-471, November 1984.

43. W. C. Y. Lee, "Estimate of local average power of a mobile radio signal," *IEEE Transactions on Vehicular Technology*, vol. VT-34, pp. 22-27, February 1985.

44. W. Lehr and L. W. McKnight, "Wireless Internet access: 3G vs. WiFi?" *Telecommunications Policy (Elsevier)*, vol. 27, no. 5, pp. 351-370, June 2003.

45. J. Li, C. Blake, D. S. J. De Couto, H. I. Lee, and R. Morris, "Capacity of ad hoc wireless networks," *Proceedings of the Annual International Conference on Mobile Computing and Networking (MobiCom'01)*, July 2001.

46. B. B. Mandelbrot and J. W. van Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422-437, 1968.

47. A. Marshall, Amateur Radio—A Brief History. Online at: http://www.vistech.net/users/w1fji/spark.html
See also: Wireless Museum at: http://www.vistech.net/users/w1fji/museum.html

48. MeshNetworks, Inc., Maitland, FL, http://www.meshnetworks.com/. Last visited: October 2003.

49. MeshNetworks, Inc., "Corporate and technology overview," MeshNetworks White Paper, 2002.

50. L. E. Miller, Wireless Propagation Bibliography, Wireless Communications Technologies Group, National Institute of Standards and Technology (NIST). Online at: http://w3.antd.nist.gov/wctg/manet/wirelesspropagation_bibliog.html, Last visited: November 2003.

51. C. S. R. Murthy and B. S. Manoj, *Ad Hoc Wireless Networks: Architectures and Protocols*. Prentice Hall PTR, Upper Saddle River, NJ, 2004.

52. National Telecommunications and Information Administration, Home Page. Online at: http://www.ntia.doc.gov/

53. S. Y. Ni, Y. C. Tseng, Y. S. Chen, and J. P. Sheu, "The broadcast storm problem in a mobile ad hoc network," *Proceedings of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99)*, pp. 152-162, 1999.

54. B. O'Hara and A. Petrick, *The IEEE 802.11 Handbook: A Designer's Companion*. IEEE Press, New York, NY, 1999.

55. A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. 4th edition, McGraw-Hill, New York, NY, 2001.

56. C. E. Perkins (Editor), *Ad Hoc Networks*. Addison-Wesley, Upper Saddle River, NJ, 2001.

57. C. E. Perkins, *Mobile IP Design Principles and Practices*. Prentice Hall PTR, Upper Saddle River, NJ, 1998.

58. L. L. Peterson and B. S. Davie, *Computer Networks: A Systems Approach*. 3rd edition. Morgan Kaufmann Publishers, San Francisco, CA, 2003.

59. J. del Prado and S. Choi, "Link adaptation strategy for IEEE 802.11 WLAN via received signal strength measurement," *Proc. IEEE Int'l Conf. Communications (ICC'03)*, Anchorage, Alaska, pp. 1108-1113, May 2003.

60. J. G. Proakis, *Digital Communications*. 4th edition. McGraw-Hill, Inc., New York, NY, 2001.

61. T. S. Rappaport, *Wireless Communications: Principle and Practice*. 2nd edition. Prentice Hall PTR, Upper Saddle River, NJ, 2002.

62. D. Schuler, *New Community Networks: Wired for Change*. ACM Press, New York, NY, 1996.

63. C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, 1949.

64. T. J. Shepard, "A channel access scheme for large dense packet radio networks," *Proc. ACM Conf. Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'96)*, Palo Alto, CA, pp. 219-230, August 1996.

65. B. Sklar, "Rayleigh fading channels in mobile digital communication systems—Part I: Characterization," *IEEE Communications Magazine*, vol. 35, no.7, pp. 90-100, July 1997.(a)
Reprinted in [Tantaratana & Ahmed 1998].

66. B. Sklar, "Rayleigh fading channels in mobile digital communication systems—Part II: Mitigation," *IEEE Communications Magazine*, vol. 35, no.9, pp. 148-155, September 1997.(b)
Reprinted in [Tantaratana & Ahmed 1998].

67. K. Sohraby, M. L. Molle, and A. N. Venetsanopoulos, "Comments on 'Throughput analysis for persistent CSMA systems,' " *IEEE Transactions on Communications*, vol. COM-35, no. 2, pp. 240-243, February 1987.

68. G. L. Stüber, *Principles of Mobile Communication*. 2nd edition. Kluwer Academic Publishers, Boston, MA, 2001.

69. H. Takagi and L. Kleinrock, "Throughput analysis for persistent CSMA systems,'' *IEEE Transactions on Communications*, vol. COM-33, no. 7, pp. 627-638, July 1985.

70. A. S. Tanenbaum, *Computer Networks*. 4th edition, Prentice Hall PTR, Upper Saddle River, NJ, 2003.

71. D. Tang and M. Baker, "Analysis of a metropolitan-area wireless network," *Proc. ACM/IEEE Mobicom '99*, Seattle, WA, pp. 13-23, August 1999.

72. S. Tantaratana and K. M. Ahmed (eds.), *Wireless Applications of Spread Spectrum Systems: Selected Readings*. IEEE Press, Piscataway, NJ, 1998.

73. Y. C. Tay, K. Jamieson, H. Balakrishnan, "Collision-minimizing CSMA and its applications to wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1048-1057, August 2004.

74. S. Toumpis and A.J. Goldsmith, "Capacity regions for wireless ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 736-748, July 2003.

75. C. Tschudin, H. Lundgren, and E. Nordström, "Embedding MANETs in the real world," *Proceedings of the Personal Wireless Communication Conference (PWC 2003)*, Venice, Italy, September 2003.

76. Y. Wang and J. J. Garcia-Luna-Aceves, "A new hybrid channel access scheme for ad hoc networks," *ACM WINET Journal*, Special Issue on Ad-Hoc Networks, vol. 10, no. 4, July 2004.

77. W. Webb, *The Future of Wireless Communications*, Artech House, Norwood, MA, 2001.

78. S. Weinstein, *IEEE Communications Magazine*, pp. 26-28, February 2002.

79. S. Woolley, "Dead air (Cover story)," *Forbes*, vol.170, no.11, pp.138-150, 25 November 2002. Online at: http://www.forbes.com/forbes/2002/1125/138.html

80. J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, John Wiley & Sons, Inc., New York, 1965. Reissued by Waveland Press, Inc., Prospect Heights, IL, 1990.

81. H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: Analysis and enhancement," *Proceedings of the Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom 2002)*, vol. 2, pp. 599-607, June 2002.

82. R. D. Yates and D. J. Goodman, *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*. 2nd edition, John Wiley & Sons, Inc., New York, NY, 2004.

83. W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," *Proceedings of the 21st International Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002)*, New York, NY, June 2002. Online at: http://www.isi.edu/scadds/publications.html

84. E. Ziouva and T. Antonakopoulos, "CSMA/CA performance under high traffic conditions: Throughput and delay analysis," *Computer Communications (Elsevier)*, vol. 25, no. 3, pp.313-321, 15 February 2002.

Acronyms and Abbreviations

3G — Third Generation
ACK — Acknowledgement
AIMD — Additive Increase/Multiplicative Decrease
AQM — Active Queue Management
AP — Access Point
AF — Assumed Forwarding
ASCII — American Standard Code for Information Interchange
AWGN — Additive White Gaussian Noise
BDPDR — Bounded Delay Packet Delivery Ratio
BER — Bit Error Rate
API — Application Programming Interface
ARQ — Automatic Repeat Request
bps — bits per second
CBR — Constant Bit Rate
CIDR — Classless Interdomain Routing
CORBA — Common Object Request Broker Architecture
CoS — Class of Service
CPU — Central Processing Unit
CTS — Clear To Send
DBS — Direct Broadcast Satellite
DCF — Distributed Coordination Function
DiffServ — Differentiated Services
dupACK — Duplicate Acknowledgement
DV — Distance Vector
EF — Expedited Forwarding
FCFS — First Come First Served
FDM — Frequency Division Multiplexing
FEC — Forward Error Correction
FIFO — First In First Out
FIRO — First In Random Out
FQ — Fair Queuing
FTP — File Transfer Protocol
GBN — Go-Back-N
GPS — Generalized Processor Sharing
GUI — Graphical User Interface
HTML — HyperText Markup Language
HTTP — HyperText Transport Protocol
IEEE — Institute of Electrical and Electronics Engineers
IETF — Internet Engineering Task Force
IntServ — Integrated Services
IP — Internet Protocol
IPv4 — Internet Protocol version 4
j.n.d. — just noticeable difference
Kbps — Kilo bits per second
LAN — Local Area Network
LCFS — Last Come First Served
LS — Link State
MAC — Medium Access Control
MANET — Mobile Ad-hoc Network
Mbps — Mega bits per second
MPEG — Moving Picture Experts Group
MSS — Maximum Segment Size
MTU — Maximum Transmission Unit
NAK — Negative Acknowledgement
NAT — Network Address Translation
NIC — Network Interface Card
PAN — Personal Area Network
PC — Personal Computer
PDA — Personal Digital Assistant
pdf — probability distribution function
pmf — probability mass function
PHB — Per-Hop Behavior
PtMP — Point-to-Multipoint
PtP — Point-to-Point
QoS — Quality of Service
P2P — Peer-to-Peer
RED — Random Early Detection
RFC — Request For Comments
RFID — Radio Frequency IDentification
RPC — Remote Procedure Call
RSSI — Receive(r) Signal Strength Index/Indication
RSVP — Resource ReSerVation Protocol
RTS — Request To Send
RTT — Round-Trip Time
SIP — Session Initiation Protocol
SN — Sequence Number
SNR — Signal-to-Noise Ratio
SR — Selective Repeat

SSTresh — Slow-Start Threshold
TCP — Transport Control Protocol
TDM — Time Division Multiplexing
UDP — User Datagram Protocol
VBR — Variable Bit Rate
VLSI — Very Large Scale Integration
VoIP — Voice over IP
W3C — World Wide Web Consortium
WAN — Wide Area Network
WAP — Wireless Access Protocol
WEP — Wired Equivalent Privacy
WFQ — Weighted Fair Queuing
Wi-Fi — Wireless Fidelity (synonym for IEEE 802.11)

Index

X

xDSL …

Y

Z

…