2 Wireless transmission

This book focuses on higher layer aspects of mobile communications, the computer science element rather than on the radio and transmission aspects, the electrical engineering part. This chapter introduces only those fundamental aspects of wireless transmission which are necessary to understand the problems of higher layers and the complexity needed to handle transmission impairments. Wherever appropriate, the reader is referred to literature giving a deeper insight into the topic. To avoid too many details blurring the overall picture, this chapter sometimes simplifies the real-world characteristics of wireless transmission. Readers who are more interested in the details of wireless transmission, calculation of propagation characteristics etc. are referred to Pahlavan (2002) or Stallings (2002).

While transmission over different wires typically does not cause interference,[1] this is an important topic in wireless transmission. The frequencies used for transmission are all regulated. The first section gives a general overview of these frequencies. The following sections recall some basic facts about signals, antennas, and signal propagation. The varying propagation characteristics create particular complications for radio transmission, frequently causing transmission errors. Multiplexing is a major design topic in this context, because the medium is always shared. Multiplexing schemes have to ensure low interference between different senders.

Modulation is needed to transmit digital data via certain frequencies. A separate section of this chapter presents standard modulation schemes that will reoccur together with the wireless communication systems presented in chapters 4 to 7. The next section discusses spread spectrum, a special transmission technique that is more robust against errors. A short introduction to cellular systems concludes this chapter.

1 However, if the transmitted frequencies are too high for a certain wire crosstalk takes place. This is a common problem, e.g., for DSL or Powerline installations, especially if many wires are bundled.

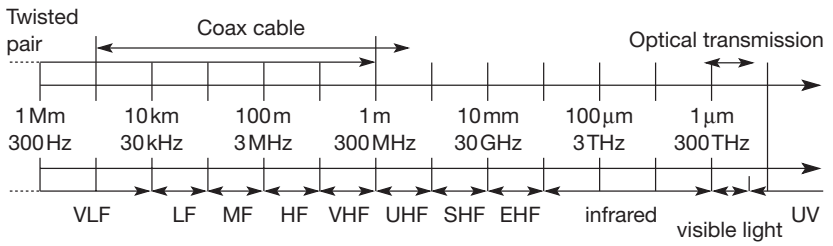

Figure 2.1 Frequency spectrum

2.1 Frequencies for radio transmission

Radio transmission can take place using many different frequency bands. Each frequency band exhibits certain advantages and disadvantages. Figure 2.1 gives a rough overview of the frequency spectrum that can be used for data transmission. The figure shows frequencies starting at 300 Hz and going up to over 300 THz.

Directly coupled to the frequency is the wavelength λ via the equation:

$$\lambda = c/f,$$

where $c \cong 3 \cdot 10^8$ m/s (the speed of light in vacuum) and f the frequency. For traditional wired networks, frequencies of up to several hundred kHz are used for distances up to some km with twisted pair copper wires, while frequencies of several hundred MHz are used with coaxial cable (new coding schemes work with several hundred MHz even with twisted pair copper wires over distances of some 100 m). Fiber optics are used for frequency ranges of several hundred THz, but here one typically refers to the wavelength which is, e.g., 1500 nm, 1350 nm etc. (infra red).

Radio transmission starts at several kHz, the **very low frequency (VLF)** range. These are very long waves. Waves in the **low frequency (LF)** range are used by submarines, because they can penetrate water and can follow the earth's surface. Some radio stations still use these frequencies, e.g., between 148.5 kHz and 283.5 kHz in Germany. The **medium frequency (MF)** and **high frequency (HF)** ranges are typical for transmission of hundreds of radio stations either as amplitude modulation **(AM)** between 520 kHz and 1605.5 kHz, as short wave **(SW)** between 5.9 MHz and 26.1 MHz, or as frequency modulation **(FM)** between 87.5 MHz and 108 MHz. The frequencies limiting these ranges are typically fixed by national regulation and, vary from country to country. Short waves are typically used for (amateur) radio transmission around the world, enabled by reflection at the ionosphere. Transmit power is up to 500 kW – which is quite high compared to the 1 W of a mobile phone.

As we move to higher frequencies, the TV stations follow. Conventional analog TV is transmitted in ranges of 174–230 MHz and 470–790 MHz using the very high frequency **(VHF)** and ultra high frequency **(UHF)** bands. In this range,

digital audio broadcasting (DAB) takes place as well (223–230 MHz and 1452–1472 MHz) and digital TV is planned or currently being installed (470–862 MHz), reusing some of the old frequencies for analog TV. UHF is also used for mobile phones with analog technology (450–465 MHz), the digital GSM (890–960 MHz, 1710–1880 MHz), digital cordless telephones following the DECT standard (1880–1900 MHz), 3G cellular systems following the UMTS standard (1900–1980 MHz, 2020–2025 MHz, 2110–2190 MHz) and many more. VHF and especially UHF allow for small antennas and relatively reliable connections for mobile telephony.

Super high frequencies (SHF) are typically used for directed microwave links (approx. 2–40 GHz) and fixed satellite services in the C-band (4 and 6 GHz), Ku-band (11 and 14 GHz), or Ka-band (19 and 29 GHz). Some systems are planned in the **extremely high frequency (EHF)** range which comes close to infra red. All radio frequencies are regulated to avoid interference, e.g., the German regulation covers 9 kHz–275 GHz.

The next step into higher frequencies involves optical transmission, which is not only used for fiber optical links but also for wireless communications. **Infra red (IR)** transmission is used for directed links, e.g., to connect different buildings via laser links. The most widespread IR technology, infra red data association (IrDA), uses wavelengths of approximately 850–900 nm to connect laptops, PDAs etc. Finally, visible light has been used for wireless transmission for thousands of years. While light is not very reliable due to interference, but it is nevertheless useful due to built-in human receivers.

2.1.1 Regulations

As the examples in the previous section have shown, radio frequencies are scarce resources. Many national (economic) interests make it hard to find common, worldwide regulations. The International Telecommunications Union (ITU) located in Geneva is responsible for worldwide coordination of telecommunication activities (wired and wireless). ITU is a sub-organization of the UN. The ITU Radiocommunication sector (ITU-R) handles standardization in the wireless sector, so it also handles frequency planning (formerly known as Consultative Committee for International Radiocommunication, CCIR).

To have at least some success in worldwide coordination and to reflect national interests, the ITU-R has split the world into three regions: **Region 1** covers Europe, the Middle East, countries of the former Soviet Union, and Africa. **Region 2** includes Greenland, North and South America, and **region 3** comprises the Far East, Australia, and New Zealand. Within these regions, national agencies are responsible for further regulations, e.g., the Federal Communications Commission (FCC) in the US. Several nations have a common agency such as European Conference for Posts and Telecommunications (CEPT) in Europe. While CEPT is still responsible for the general planning, many tasks have been transferred to other agencies (confusing anybody following the regulation

process). For example, the European Telecommunications Standards Institute (ETSI) is responsible for standardization and consists of national standardization bodies, public providers, manufacturers, user groups, and research institutes.

To achieve at least some harmonization, the ITU-R holds, the World Radio Conference (WRC), to periodically discuss and decide frequency allocations for all three regions. This is obviously a difficult task as many regions or countries may have already installed a huge base of a certain technology and will be reluctant to change frequencies just for the sake of harmonization. Harmonization is, however, needed as soon as satellite communication is used. Satellites, especially the new generation of low earth-orbiting satellites (see chapter 5) do not 'respect' national regulations, but should operate worldwide. While it is difficult to prevent other nations from setting up a satellite system it is much simpler to ban the necessary devices or the infrastructure needed for operation. Satellite systems should operate on frequencies available worldwide to support global usage with a single device.

Table 2.1 gives some examples for frequencies used for (analog and digital) mobile phones, cordless telephones, wireless LANs, and other radio frequency (RF) systems for countries in the three regions representing the major economic power. Older systems like Nordic Mobile Telephone (NMT) are not available all over Europe, and sometimes they have been standardized with different national frequencies. The newer (digital) systems are compatible throughout Europe (standardized by ETSI).

Table 2.1 Example systems and their frequency allocations (all values in MHz)

	Europe	**US**	**Japan**
Mobile phones	**NMT**	**AMPS, TDMA, CDMA**	**PDC**
	453–457 463–467	824–849 869–894	810–826 940–956 1429–1465 1477–1513
	GSM	**GSM, TDMA, CDMA**	
	890–915 935–960 1710–1785 1805–1880	1850–1910 1930–1990	
	UMTS (FDD)/ W–CDMA 1920–1980 2110–2190		**FOMA/ W–CDMA** 1920–1980 2110–2170

	UMTS (TDD) 1900–1920 2020–2025		
Cordless telephones	**CT1+** 885–887 930–932	**PACS** 1850–1910 1930–1990	**PHS** 1895–1918
	CT2 864–868	**PACS–UB** 1910–1930	**JCT** 254–380
	DECT 1880–1900		
Wireless LANs	**IEEE 802.11** 2400–2483	**IEEE 802.11** 902–928 2400–2483	**IEEE 802.11** 2400–2497
	HiperLAN2, IEEE 802.11a 5150–5350 5470–5725	**HiperLAN2, IEEE 802.11a** 5150–5350 5725–5825	**HiperLAN2, IEEE 802.11a** 5150–5250
Others	**RF-Control** 27, 128, 418, 433, 868	**RF-Control** 315, 915	**RF-Control** 426, 868
	Satellite (e.g., Iridium, Globalstar) 1610–1626, 2483–2500		

While older analog **mobile phone** systems like NMT or its derivatives at 450 MHz are still available, Europe is heavily dominated by the fully digital GSM (see chapter 4.1) at 900 MHz and 1800 MHz (also known as DCS1800, Digital Cellular System). In contrast to Europe, the US FCC allowed several cellular technologies in the same frequency bands around 850 MHz. Starting from the analog advanced mobile phone system (AMPS), this led to the co-existence of several solutions, such as dual mode mobile phones supporting digital time division multiple access (TDMA) service and analog AMPS according to the standard IS-54. All digital TDMA phones according to IS-136 (also known as NA-TDMA, North American TDMA) and digital code division multiple access (CDMA) phones according to IS-95 have been developed. The US did not adopt a common mobile phone system, but waited for market forces to decide. This led to many islands of different systems and, consequently, as in Europe, full coverage, is not available in the US. The long discussions about the pros and cons of TDMA and CDMA

also promoted the worldwide success of GSM. GSM is available in over 190 countries and used by more than 800 million people (GSM World, 2002). A user can roam with the same mobile phone from Zimbabwe, via Uzbekistan, Sweden, Singapore, USA, Tunisia, Russia, Canada, Italy, Greece, Germany, China, and Belgium to Austria.

Another system, the personal digital cellular (**PDC**), formerly known as Japanese digital cellular (JDC) was established in Japan. Quite often mobile phones covering many standards have been announced, however, industry is still waiting for a cheap solution. Chapter 11 will discuss this topic again in the context of software defined radios (SDR). New frequency bands, e.g., for the universal mobile telecommunications system (**UMTS**) or the freedom of mobile multi-media access (**FOMA**) are located at 1920–1980 MHz and 2110–2170/2190 MHz (see chapter 4).

Many different **cordless telephone** standards exist around the world. However, this is not as problematic as the diversity of mobile phone standards. Some older analog systems such as cordless telephone (**CT1+**) are still in use, but digital technology has been introduced for cordless telephones as well. Examples include **CT2**, the first digital cordless telephone introduced in the UK, digital enhanced cordless telecommunications (**DECT**) as a European standard (see section 4.2), personal access communications system (PACS) and PACS-Unlicensed Band (**PACS-UB**) in the US, as well as personal handyphone system (PHS) as replacement for the analog Japanese cordless telephone (**JCT**) in Japan. Mobile phones covering, e.g., DECT and GSM are available but they have not been a commercial success.

Finally, the area of **WLAN** standards is of special interest for wireless, mobile computer communication on a campus or in buildings. Here the computer industry developed products within the license-free **ISM** band, of which the most attractive is located at 2.4 GHz and is available for license-free operation almost everywhere around the world (with national differences limiting frequencies, transmit power etc.). The most widespread standard in this area is **IEEE 802.11b**, which is discussed in chapter 7 (together with other members of the 802.11 family). The wireless LAN standards **HiperLAN2** and **IEEE 802.11a** operate in the 5 GHz range, but depending on the region on different frequencies with different restrictions.

Many more frequencies have been assigned for trunk radio (e.g., trans-European trunked radio (TETRA), 380–400 MHz, 410–430 MHz, 450–470 MHz – depending on national regulations), paging services, terrestrial flight telephone system (TFTS), 1670–1675 MHz and 1800–1805 MHz, satellite services (Iridium: 1610–1626 MHz, Globalstar: 1610–1626 MHz and 2483–2500 MHz, see chapter 5) etc. Higher frequencies are of special interest for high bit-rate transmission, although these frequencies face severe shadowing by many obstacles. License-free bands at 17.2, 24 and even 61 GHz are under consideration for commercial use. Additionally, a lot of license-free wireless communication takes place at lower frequencies. Garage openers, car locks, wireless headsets, radio frequency identifications (RFID) etc. operate on, e.g., 433 or 868 MHz.

2.2 Signals

Signals are the physical representation of data. Users of a communication system can only exchange data through the transmission of signals. Layer 1 of the ISO/OSI basic reference model is responsible for the conversion of data, i.e., bits, into signals and vice versa (Halsall, 1996), (Stallings, 1997 and 2002).

Signals are functions of time and location. Signal parameters represent the data values. The most interesting types of signals for radio transmission are **periodic signals**, especially **sine waves** as carriers. (The process of mapping of data onto a carrier is explained in section 2.6.) The general function of a sine wave is:

$$g(t) = A_t \sin(2\pi f_t t + \varphi_t)$$

Signal parameters are the **amplitude** A, the **frequency** f, and the **phase shift** φ. The amplitude as a factor of the function g may also change over time, thus A_t, (see section 2.6.1). The frequency f expresses the periodicity of the signal with the period $T = 1/f$. (In equations, ω is frequently used instead of $2\pi f$.) The frequency f may also change over time, thus f_t, (see section 2.6.2). Finally, the phase shift determines the shift of the signal relative to the same signal without a shift. An example for shifting a function is shown in Figure 2.2. This shows a sine function without a phase shift and the same function, i.e., same amplitude and frequency, with a phase shift φ. Section 2.6.3 shows how shifting the phase can be used to represent data.

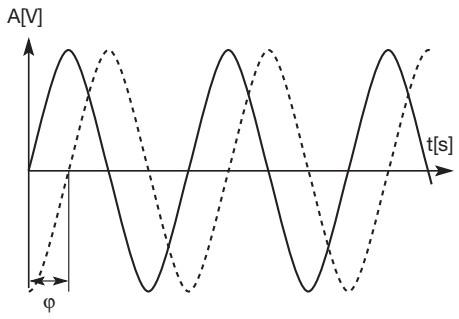

Figure 2.2 Time domain representation of a signal

Sine waves are of special interest, as it is possible to construct every periodic signal g by using only sine and cosine functions according to a fundamental equation of **Fourier**:

$$g(t) = \frac{1}{2}c + \sum_{n=1}^{\infty} a_n \sin(2\pi nft) + \sum_{n=1}^{\infty} b_n \cos(2\pi nft)$$

In this equation the parameter c determines the **Direct Current (DC)** component of the signal, the coefficients a_n and b_n are the amplitudes of the nth sine and cosine function. The equation shows that an infinite number of sine and cosine functions is needed to construct arbitrary periodic functions. However, the frequencies of these functions (the so-called **harmonics**) increase with a growing parameter n and are a multiple of the **fundamental frequency** f. The bandwidth of any medium, air, cable, transmitter etc. is limited and, there is an upper limit for the frequencies. In reality therefore, it is enough to consider a limited number of sine and cosine functions to construct periodic

functions – all real transmitting systems exhibit these bandwidth limits and can never transmit arbitrary periodic functions. It is sufficient for us to know that we can think of transmitted signals as composed of one or many sine functions. The following illustrations always represent the example of one sine function, i.e., the case of a single frequency.

A typical way to represent signals is the time domain (see Figure 2.2). Here the amplitude A of a signal is shown versus time (time is mostly measured in seconds s, amplitudes can be measured in, e.g., volt V). This is also the typical representation known from an oscilloscope. A phase shift can also be shown in this representation.

Representations in the time domain are problematic if a signal consists of many different frequencies (as the Fourier equation indicates). In this case, a better representation of a signal is the **frequency domain** (see Figure 2.3). Here the amplitude of a certain frequency part of the signal is shown versus the frequency. Figure 2.3 only shows one peak and the signal consists only of a single frequency part (i.e., it is a single sine function). Arbitrary periodic functions would have many peaks, known as the frequency spectrum of a signal. A tool to display frequencies is a spectrum analyzer. Fourier transformations are a mathematical tool for translating from the time domain into the frequency domain and vice versa (using the inverse Fourier transformation).

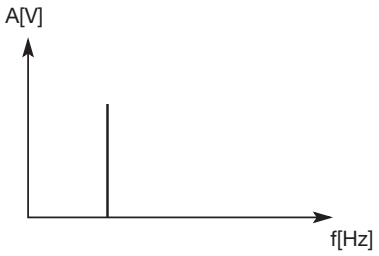

Figure 2.3 Frequency domain representation of a signal

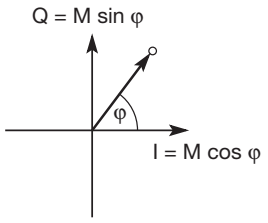

Figure 2.4 Phase domain representation of a signal

A third way to represent signals is the **phase domain** shown in Figure 2.4. This representation, also called phase state or signal constellation diagram, shows the amplitude M of a signal and its phase φ in polar coordinates. (The length of the vector represents the amplitude, the angle the phase shift.) The x-axis represents a phase of 0 and is also called **In-Phase (I)**. A phase shift of 90° or $\pi/2$ would be a point on the y-axis, called **Quadrature (Q)**.

2.3 Antennas

As the name wireless already indicates, this communication mode involves 'getting rid' of wires and transmitting signals through space without guidance. We do not need any 'medium' (such as an ether) for the transport of electromagnetic waves. Somehow, we have to couple the energy from the transmitter to the out-

side world and, in reverse, from the outside world to the receiver. This is exactly what **antennas** do. Antennas couple electromagnetic energy to and from space to and from a wire or coaxial cable (or any other appropriate conductor).

A theoretical reference antenna is the **isotropic radiator**, a point in space radiating equal power in all directions, i.e., all points with equal power are located on a sphere with the antenna as its center. The **radiation pattern** is symmetric in all directions (see Figure 2.5, a two dimensional cross-section of the real three-dimensional pattern).

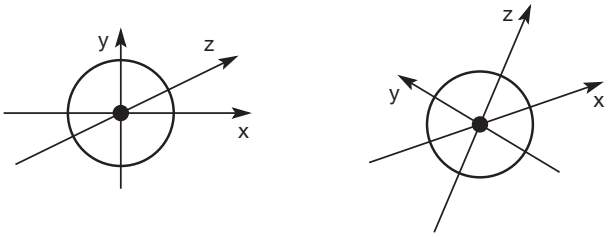

Figure 2.5
Radiation pattern of an isotropic radiator

However, such an antenna does not exist in reality. Real antennas all exhibit **directive effects**, i.e., the intensity of radiation is not the same in all directions from the antenna. The simplest real antenna is a thin, center-fed **dipole**, also called Hertzian dipole, as shown in Figure 2.6 (right-hand side). The dipole consists of two collinear conductors of equal length, separated by a small feeding gap. The length of the dipole is not arbitrary, but, for example, half the wavelength λ of the signal to transmit results in a very efficient radiation of the energy. If mounted on the roof of a car, the length of $\lambda/4$ is efficient. This is also known as Marconi antenna.

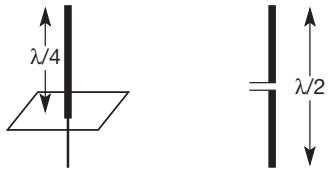

Figure 2.6
Simple antennas

A $\lambda/2$ dipole has a uniform or **omni-directional** radiation pattern in one plane and a figure eight pattern in the other two planes as shown in Figure 2.7. This type of antenna can only overcome environmental challenges by boosting the power level of the signal. Challenges could be mountains, valleys, buildings etc.

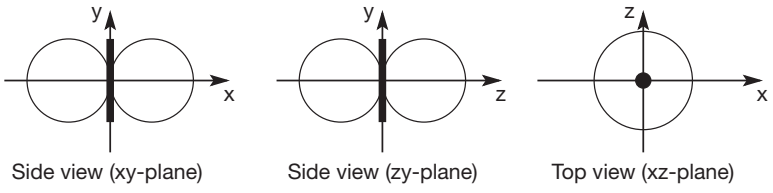

Figure 2.7
Radiation pattern of a simple dipole

If an antenna is positioned, e.g., in a valley or between buildings, an omnidirectional radiation pattern is not very useful. In this case, **directional antennas** with certain fixed preferential transmission and reception directions can be used. Figure 2.8 shows the radiation pattern of a directional antenna with the main lobe in the direction of the x-axis. A special example of directional antennas is constituted by satellite dishes.

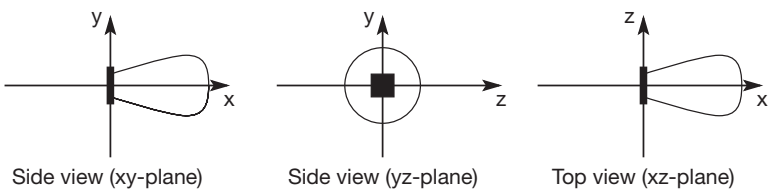

Figure 2.8 Radiation pattern of a directed antenna

Directed antennas are typically applied in cellular systems as presented in section 2.8. Several directed antennas can be combined on a single pole to construct a **sectorized antenna**. A cell can be sectorized into, for example, three or six sectors, thus enabling frequency reuse as explained in section 2.8. Figure 2.9 shows the radiation patterns of these sectorized antennas.

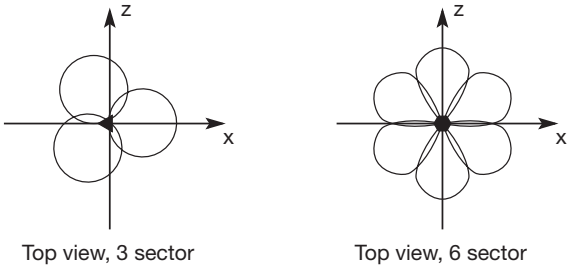

Figure 2.9 Radiation patterns of sectorized antennas

Two or more antennas can also be combined to improve reception by counteracting the negative effects of multi-path propagation (see section 2.4.3). These antennas, also called **multi-element antenna arrays**, allow different diversity schemes. One such scheme is **switched diversity** or **selection diversity**, where the receiver always uses the antenna element with the largest output. **Diversity combining** constitutes a combination of the power of all signals to produce gain. The phase is first corrected (cophasing) to avoid cancellation. As shown in Figure 2.10, different schemes are possible. On the left, two $\lambda/4$ antennas are combined with a distance of $\lambda/2$ between them on top of a ground plane. On the right, three standard $\lambda/2$ dipoles are combined with a distance of $\lambda/2$ between them. Spacing could also be in multiples of $\lambda/2$.

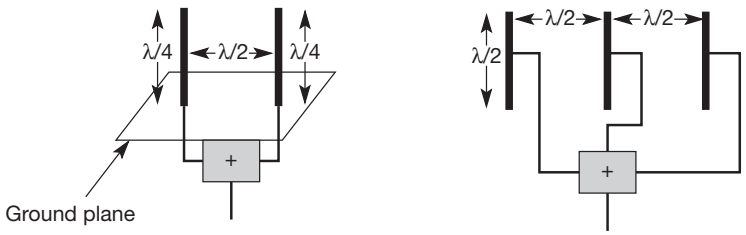

Figure 2.10
Diversity antenna systems

A more advanced solution is provided by **smart antennas** which combine multiple antenna elements (also called antenna array) with signal processing to optimize the radiation/reception pattern in response to the signal environment. These antennas can adapt to changes in reception power, transmission conditions and many signal propagation effects as discussed in the following section. Antenna arrays can also be used for beam forming. This would be an extreme case of a directed antenna which can follow a single user thus using space division multiplexing (see section 2.5.1). It would not just be base stations that could follow users with an individual beam. Wireless devices, too, could direct their electromagnetic radiation, e.g., away from the human body towards a base station. This would help in reducing the absorbed radiation. Today's handset antennas are omni-directional as the integration of smart antennas into mobiles is difficult and has not yet been realized.

2.4 Signal propagation

Like wired networks, wireless communication networks also have senders and receivers of signals. However, in connection with signal propagation, these two networks exhibit considerable differences. In wireless networks, the signal has no wire to determine the direction of propagation, whereas signals in wired networks only travel along the wire (which can be twisted pair copper wires, a coax cable, but also a fiber etc.). As long as the wire is not interrupted or damaged, it typically exhibits the same characteristics at each point. One can precisely determine the behavior of a signal travelling along this wire, e.g., received power depending on the length. For wireless transmission, this predictable behavior is only valid in a vacuum, i.e., without matter between the sender and the receiver. The situation would be as follows (Figure 2.11):

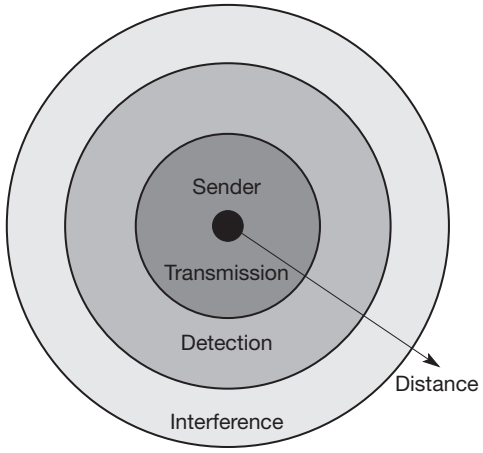

Figure 2.11
Ranges for transmission, detection, and interference of signals

- **Transmission range**: Within a certain radius of the sender transmission is possible, i.e., a receiver receives the signals with an error rate low enough to be able to communicate and can also act as sender.
- **Detection range**: Within a second radius, detection of the transmission is possible, i.e., the transmitted power is large enough to differ from background noise. However, the error rate is too high to establish communication.
- **Interference range**: Within a third even larger radius, the sender may interfere with other transmission by adding to the background noise. A receiver will not be able to detect the signals, but the signals may disturb other signals.

This simple and ideal scheme led to the notion of **cells** around a transmitter (as briefly discussed in section 2.8). However, real life does not happen in a vacuum, radio transmission has to contend with our atmosphere, mountains, buildings, moving senders and receivers etc. In reality, the three circles referred to above will be bizarrely-shaped polygons with their shape being time and frequency dependent. The following paragraphs discuss some problems arising in this context, thereby showing the differences between wireless and wired transmission.

2.4.1 Path loss of radio signals

In free space radio signals propagate as light does (independently of their frequency), i.e., they follow a straight line (besides gravitational effects). If such a straight line exists between a sender and a receiver it is called **line-of-sight (LOS)**. Even if no matter exists between the sender and the receiver (i.e., if there is a vacuum), the signal still experiences the **free space loss**. The received power P_r is proportional to $1/d^2$ with d being the distance between sender and receiver (**inverse square law**). The reason for this phenomenon is quite simple. Think of the sender being a point in space. The sender now emits a signal with certain energy. This signal travels away from the sender at the speed of light as a wave with a spherical shape. If there is no obstacle, the sphere continuously grows with the sending energy equally distributed over the sphere's surface. This surface area s grows with the increasing distance d from the center according to the equation $s = 4\pi d^2$.

Even without any matter between sender and receiver, additional parameters are important. The received power also depends on the wavelength and the gain of receiver and transmitter antennas. As soon as there is any matter between sender and receiver, the situation becomes more complex. Most radio transmission takes place through the atmosphere – signals travel through air, rain, snow, fog, dust particles, smog etc. While the **path loss** or **attenuation** does not cause too much trouble for short distances, e.g., for LANs (see chapter 7), the atmosphere heavily influences transmission over long distances, e.g., satellite transmission (see chapter 5). Even mobile phone systems are influenced by weather conditions such as heavy rain. Rain can absorb much of the radiated energy of the antenna (this effect is used in a microwave oven to cook), so communication links may break down as soon as the rain sets in.

Depending on the frequency, radio waves can also penetrate objects. Generally the lower the frequency, the better the penetration. Long waves can be transmitted through the oceans to a submarine while high frequencies can be blocked by a tree. The higher the frequency, the more the behavior of the radio waves resemble that of light – a phenomenon which is clear if one considers the spectrum shown in Figure 2.1.

Radio waves can exhibit three fundamental propagation behaviors depending on their frequency:

- **Ground wave** (<2 MHz): Waves with low frequencies follow the earth's surface and can propagate long distances. These waves are used for, e.g., submarine communication or AM radio.
- **Sky wave** (2–30 MHz): Many international broadcasts and amateur radio use these short waves that are reflected[2] at the ionosphere. This way the waves can bounce back and forth between the ionosphere and the earth's surface, travelling around the world.
- **Line-of-sight** (>30 MHz): Mobile phone systems, satellite systems, cordless telephones etc. use even higher frequencies. The emitted waves follow a (more or less) straight line of sight. This enables direct communication with satellites (no reflection at the ionosphere) or microwave links on the ground. However, an additional consideration for ground-based communication is that the waves are bent by the atmosphere due to refraction (see next section).

Almost all communication systems presented in this book work with frequencies above 100 MHz so, we are almost exclusively concerned with LOS communication. But why do mobile phones work even without an LOS?

2.4.2 Additional signal propagation effects

As discussed in the previous section, signal propagation in free space almost follows a straight line, like light. But in real life, we rarely have a line-of-sight between the sender and receiver of radio signals. Mobile phones are typically used in big cities with skyscrapers, on mountains, inside buildings, while driving through an alley etc. Hare several effects occur in addition to the attenuation caused by the distance between sender and receiver, which are again very much frequency dependent.

An extreme form of attenuation is **blocking** or **shadowing** of radio signals due to large obstacles (see Figure 2.12, left side). The higher the frequency of a signal, the more it behaves like light. Even small obstacles like a simple wall, a truck on the street, or trees in an alley may block the signal. Another effect is the **reflection** of signals as shown in the middle of Figure 2.12. If an object is large compared to the wavelength of the signal, e.g., huge buildings, mountains,

2 Compared to, e.g., the surface of a building, the ionosphere is not really a hard reflecting surface. In the case of sky waves the 'reflection' is caused by refraction.

or the surface of the earth, the signal is reflected. The reflected signal is not as strong as the original, as objects can absorb some of the signal's power. Reflection helps transmitting signals as soon as no LOS exists. This is the standard case for radio transmission in cities or mountain areas. Signals transmitted from a sender may bounce off the walls of buildings several times before they reach the receiver. The more often the signal is reflected, the weaker it becomes. Finally, the right side of Figure 2.12 shows the effect of **refraction**. This effect occurs because the velocity of the electromagnetic waves depends on the density of the medium through which it travels. Only in vacuum does it equal *c*. As the figure shows, waves that travel into a denser medium are bent towards the medium. This is the reason for LOS radio waves being bent towards the earth: the density of the atmosphere is higher closer to the ground.

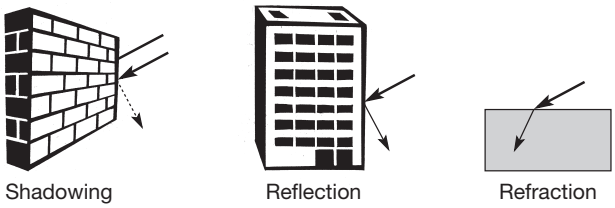

Figure 2.12 Blocking (shadowing), reflection and refraction of waves

While shadowing and reflection are caused by objects much larger than the wavelength of the signals (and demonstrate the typical 'particle' behavior of radio signals), the following two effects exhibit the 'wave' character of radio signals. If the size of an obstacle is in the order of the wavelength or less, then waves can be **scattered** (see Figure 2.13, left side). An incoming signal is scattered into several weaker outgoing signals. In school experiments, this is typically demonstrated with laser light and a very small opening or obstacle, but here we have to take into consideration that the typical wavelength of radio transmission for, e.g., GSM or AMPS is in the order of some 10 cm. Thus, many objects in the environment can cause these scattering effects. Another effect is **diffraction** of waves. As shown on the right side of Figure 2.13, this effect is very similar to scattering. Radio waves will be deflected at an edge and propagate in different directions. The result of scattering and diffraction are patterns with varying signal strengths depending on the location of the receiver.

Effects like attenuation, scattering, diffraction, and refraction all happen simultaneously and are frequency and time dependent. It is very difficult to predict the precise strength of signals at a certain point in space. How do mobile phone operators plan the coverage of their antennas, the location of the antennas, the direction of the beams etc.? Two or three dimensional maps are used with a resolution down to several meters. With the help of, e.g., ray tracing or radiosity techniques similar to rendering 3D graphics, the signal quality can roughly be calculated in advance. Additionally, operators perform a lot of measurements during and after installation of antennas to fill gaps in the coverage.

Figure 2.13 Scattering and diffraction of waves

2.4.3 Multi-path propagation

Together with the direct transmission from a sender to a receiver, the propagation effects mentioned in the previous section lead to one of the most severe radio channel impairments, called **multi-path propagation**. Figure 2.14 shows a sender on the left and one possible receiver on the right. Radio waves emitted by the sender can either travel along a straight line, or they may be reflected at a large building, or scattered at smaller obstacles. This simplified figure only shows three possible paths for the signal. In reality, many more paths are possible. Due to the finite speed of light, signals travelling along different paths with different lengths arrive at the receiver at different times. This effect (caused by multi-path propagation) is called **delay spread**: the original signal is spread due to different delays of parts of the signal. This delay spread is a typical effect of radio transmission, because no wire guides the waves along a single path as in the case of wired networks (however, a similar effect, dispersion, is known for high bit-rate optical transmission over multi-mode fiber, see Halsall, 1996, or Stallings, 1997). Notice that this effect has nothing to do with possible movements of the sender or receiver. Typical values for delay spread are approximately 3 μs in cities, up to 12 μs can be observed. GSM, for example, can tolerate up to 16 μs of delay spread, i.e., almost a 5 km path difference.

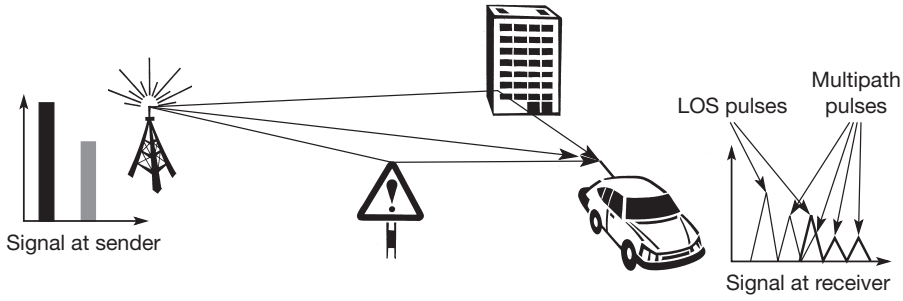

Figure 2.14 Multi-path propagation and intersymbol interference

What are the **effects** of this delay spread on the signals representing the data? The first effect is that a short impulse will be smeared out into a broader impulse, or rather into several weaker impulses. In Figure 2.14 only three possible paths are shown and, thus, the impulse at the sender will result in three smaller impulses at the receiver. For a real situation with hundreds of different paths, this implies that a single impulse will result in many weaker impulses at the receiver. Each path has a different attenuation and, the received pulses have different power. Some of the received pulses will be too weak even to be detected (i.e., they will appear as noise).

Now consider the second impulse shown in Figure 2.14. On the sender side, both impulses are separated. At the receiver, both impulses interfere, i.e., they overlap in time. Now consider that each impulse should represent a symbol, and that one or several symbols could represent a bit. The energy intended for one symbol now spills over to the adjacent symbol, an effect which is called **intersymbol interference (ISI)**. The higher the symbol rate to be transmitted, the worse the effects of ISI will be, as the original symbols are moved closer and closer to each other. ISI limits the bandwidth of a radio channel with multi-path propagation (which is the standard case). Due to this interference, the signals of different symbols can cancel each other out leading to misinterpretations at the receiver and causing transmission errors.

In this case, knowing the channel characteristics can be a great help. If the receiver knows the delays of the different paths (or at least the main paths the signal takes), it can compensate for the distortion caused by the channel. The sender may first transmit a **training sequence** known by the receiver. The receiver then compares the received signal to the original training sequence and programs an **equalizer** that compensates for the distortion (Wesel, 1998), (Pahlavan, 2002), (Stallings, 2002).

While ISI and delay spread already occur in the case of fixed radio transmitters and receivers, the situation is even worse if receivers, or senders, or both, move. Then the channel characteristics change over time, and the paths a signal can travel along vary. This effect is well known (and audible) with analog radios while driving. The power of the received signal changes considerably over time. These quick changes in the received power are also called **short-term fading**. Depending on the different paths the signals take, these signals may have a different phase and cancel each other as shown in Figure 2.15. The receiver now has to try to constantly adapt to the varying channel characteristics, e.g., by changing the parameters of the equalizer. However, if these changes are too fast, such as driving on a highway through a city, the receiver cannot adapt fast enough and the error rate of transmission increases dramatically.

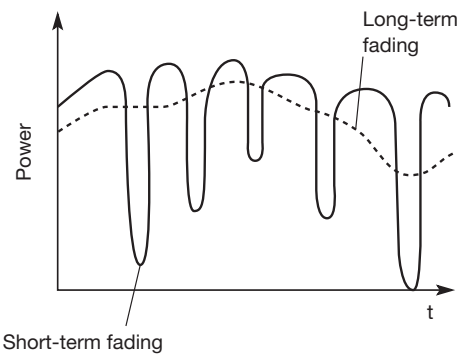

Figure 2.15 Short-term and long-term fading

An additional effect shown in Figure 2.15 is the **long-term fading** of the received signal. This long-term fading, shown here as the average power over time, is caused by, for example, varying distance to the sender or more remote obstacles. Typically, senders can compensate for long-term fading by increasing/decreasing sending power so that the received signal always stays within certain limits.

There are many more effects influencing radio transmission which will not be discussed in detail – for example, the **Doppler shift** caused by a moving sender or receiver. While this effect is audible for acoustic waves already at low speed, it is also a topic for radio transmission from or to fast moving transceivers. One example of such a transceiver could be a satellite (see chapter 5) – there Doppler shift causes random frequency shifts. The interested reader is referred to Anderson (1995), (Pahlavan, 2002), and (Stallings, 2002) for more information about the characteristics of wireless communication channels. For the present it will suffice to know that multi-path propagation limits the maximum bandwidth due to ISI and that moving transceivers cause additional problems due to varying channel characteristics.

2.5 Multiplexing

Multiplexing is not only a fundamental mechanism in communication systems but also in everyday life. Multiplexing describes how several users can share a medium with minimum or no interference. One example, is highways with several lanes. Many users (car drivers) use the same medium (the highways) with hopefully no interference (i.e., accidents). This is possible due to the provision of several lanes (space division multiplexing) separating the traffic. In addition, different cars use the same medium (i.e., the same lane) at different points in time (time division multiplexing).

While this simple example illustrates our everyday use of multiplexing, the following examples will deal with the use of multiplexing in wireless communications. Mechanisms controlling the use of multiplexing and the assignment of a medium to users (the traffic regulations), are discussed in chapter 3 under the aspect of medium access control.

2.5.1 Space division multiplexing

For wireless communication, multiplexing can be carried out in four dimensions: **space**, **time**, **frequency**, and **code**. In this field, the task of multiplexing is to assign space, time, frequency, and code to each communication channel with a minimum of interference and a maximum of medium utilization. The term communication channel here only refers to an association of sender(s) and receiver(s) who want to exchange data. Characteristics of communication channels (e.g., bandwidth, error rate) will be discussed together with certain technologies in chapters 4 to 7.

Figure 2.16 shows six channels k_i and introduces a three dimensional coordinate system. This system shows the dimensions of code c, time t and frequency f. For this first type of multiplexing, **space division multiplexing (SDM)**, the (three dimensional) space s_i is also shown. Here space is represented via circles indicating the interference range as introduced in Figure 2.11. How is the separation of the different channels achieved? The channels k_1 to k_3 can be mapped onto the three 'spaces' s_1 to s_3 which clearly separate the channels and prevent the interference ranges from overlapping. The space between the interference ranges is sometimes called **guard space**. Such a guard space is needed in all four multiplexing schemes presented.

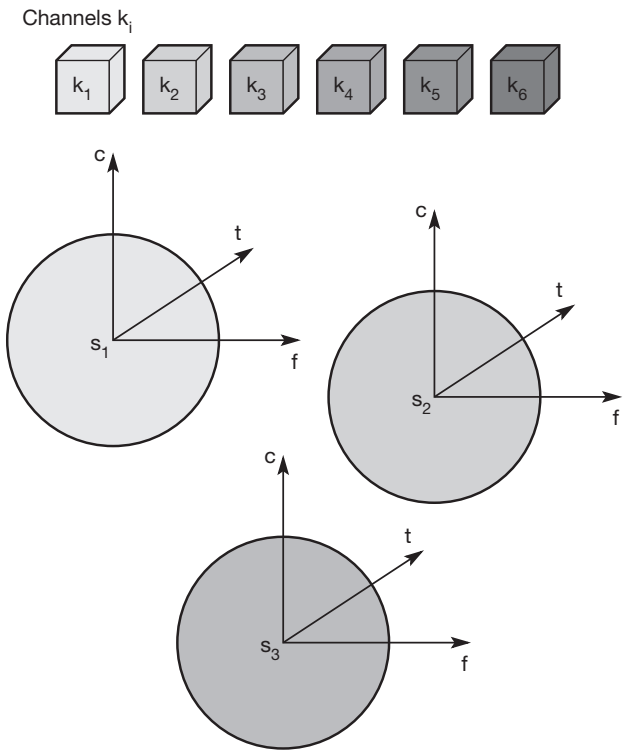

Figure 2.16 Space division multiplexing (SDM)

For the remaining channels (k_4 to k_6) three additional spaces would be needed. In our highway example this would imply that each driver had his or her own lane. Although this procedure clearly represents a waste of space, this is exactly the principle used by the old analog telephone system: each subscriber is given a separate pair of copper wires to the local exchange. In wireless transmission, SDM implies a separate sender for each communication channel with a wide enough distance between senders. This multiplexing scheme is used, for example, at FM radio stations where the transmission range is limited to a certain region –

many radio stations around the world can use the same frequency without interference. Using SDM, obvious problems arise if two or more channels were established within the same space, for example, if several radio stations want to broadcast in the same city. Then, one of the following multiplexing schemes must be used (frequency, time, or code division multiplexing).

2.5.2 Frequency division multiplexing

Frequency division multiplexing (FDM) describes schemes to subdivide the frequency dimension into several non-overlapping frequency bands as shown in Figure 2.17. Each channel k_i is now allotted its own frequency band as indicated. Senders using a certain frequency band can use this band continuously. Again, **guard spaces** are needed to avoid frequency band overlapping (also called **adjacent channel interference**). This scheme is used for radio stations within the same region, where each radio station has its own frequency. This very simple multiplexing scheme does not need complex coordination between sender and receiver: the receiver only has to tune in to the specific sender.

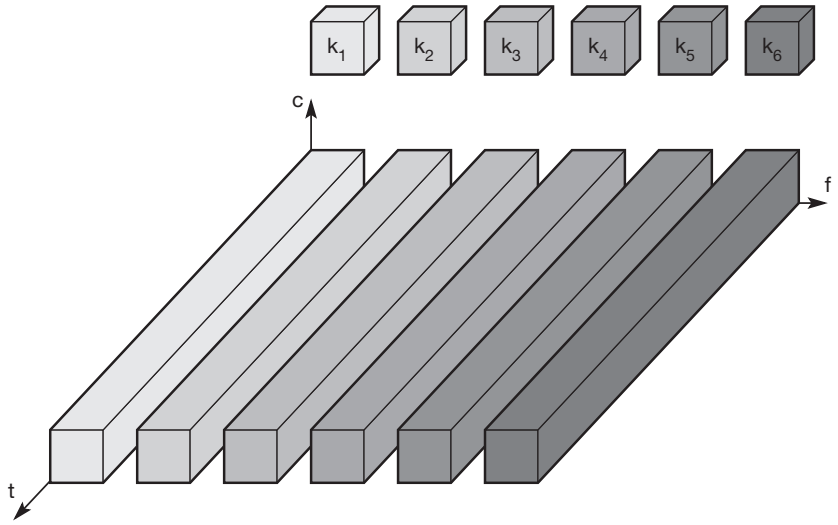

Figure 2.17 Frequency division multiplexing (FDM)

However, this scheme also has disadvantages. While radio stations broadcast 24 hours a day, mobile communication typically takes place for only a few minutes at a time. Assigning a separate frequency for each possible communication scenario would be a tremendous waste of (scarce) frequency resources. Additionally, the fixed assignment of a frequency to a sender makes the scheme very inflexible and limits the number of senders.

2.5.3 Time division multiplexing

A more flexible multiplexing scheme for typical mobile communications is **time division multiplexing (TDM)**. Here a channel k_i is given the whole bandwidth for a certain amount of time, i.e., all senders use the same frequency but at different points in time (see Figure 2.18). Again, **guard spaces**, which now represent time gaps, have to separate the different periods when the senders use the medium. In our highway example, this would refer to the gap between two cars. If two transmissions overlap in time, this is called co-channel interference. (In the highway example, interference between two cars results in an accident.) To avoid this type of interference, precise synchronization between different senders is necessary. This is clearly a disadvantage, as all senders need precise clocks or, alternatively, a way has to be found to distribute a synchronization signal to all senders. For a receiver tuning in to a sender this does not just involve adjusting the frequency, but involves listening at exactly the right point in time. However, this scheme is quite flexible as one can assign more sending time to senders with a heavy load and less to those with a light load.

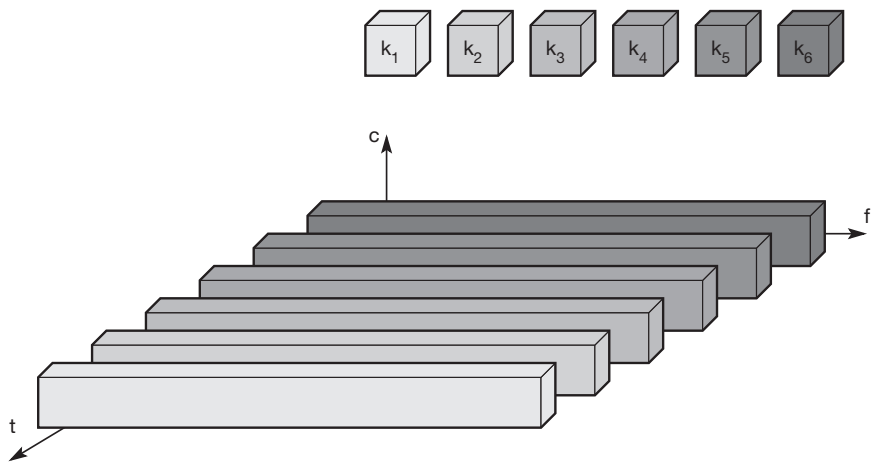

Figure 2.18 Time division multiplexing (TDM)

Frequency and time division multiplexing can be combined, i.e., a channel k_i can use a certain frequency band for a certain amount of time as shown in Figure 2.19. Now guard spaces are needed both in the time and in the frequency dimension. This scheme is more robust against frequency selective interference, i.e., interference in a certain small frequency band. A channel may use this band only for a short period of time. Additionally, this scheme provides some (weak) protection against tapping, as in this case the sequence of frequencies a sender uses has to be known to listen in to a channel. The mobile phone standard GSM uses this combination of frequency and time division multiplexing for transmission between a mobile phone and a so-called base station (see section 4.1).

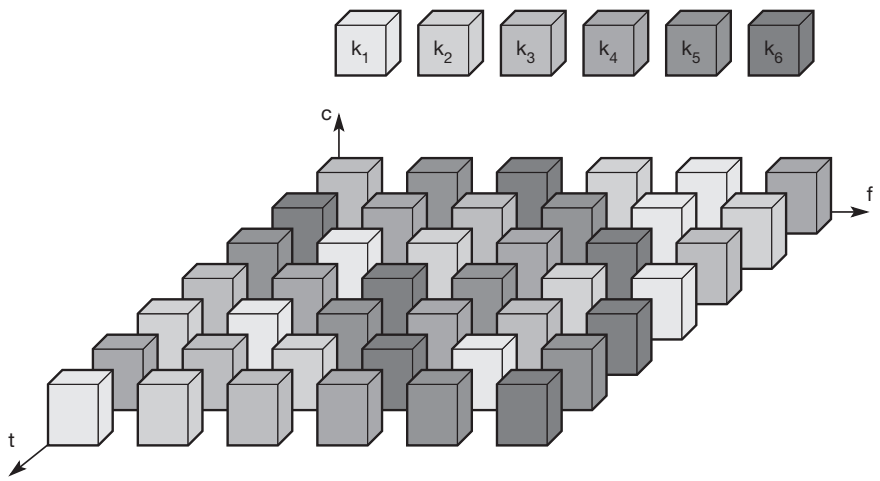

Figure 2.19
Frequency and time division multiplexing combined

A disadvantage of this scheme is again the necessary coordination between different senders. One has to control the sequence of frequencies and the time of changing to another frequency. Two senders will interfere as soon as they select the same frequency at the same time. However, if the frequency change (also called frequency hopping) is fast enough, the periods of interference may be so small that, depending on the coding of data into signals, a receiver can still recover the original data. (This technique is discussed in section 2.7.2.)

2.5.4 Code division multiplexing

While SDM and FDM are well known from the early days of radio transmission and TDM is used in connection with many applications, **code division multiplexing (CDM)** is a relatively new scheme in commercial communication systems. First used in military applications due to its inherent security features (together with spread spectrum techniques, see section 2.7), it now features in many civil wireless transmission scenarios thanks to the availability of cheap processing power (explained in more detail in section 3.5). Figure 2.20 shows how all channels k_i use the same frequency at the same time for transmission. Separation is now achieved by assigning each channel its own 'code', **guard spaces** are realized by using codes with the necessary 'distance' in code space, e.g., **orthogonal codes**. The technical realization of CDM is discussed in section 2.7 and chapter 3 together with the medium access mechanisms. An excellent book dealing with all aspects of CDM is Viterbi (1995).

The typical everyday example of CDM is a party with many participants from different countries around the world who establish communication channels, i.e., they talk to each other, using the same frequency range (approx. 300–6000 Hz depending on a person's voice) at the same time. If everybody speaks the same language, SDM is needed to be able to communicate (i.e., standing in groups,

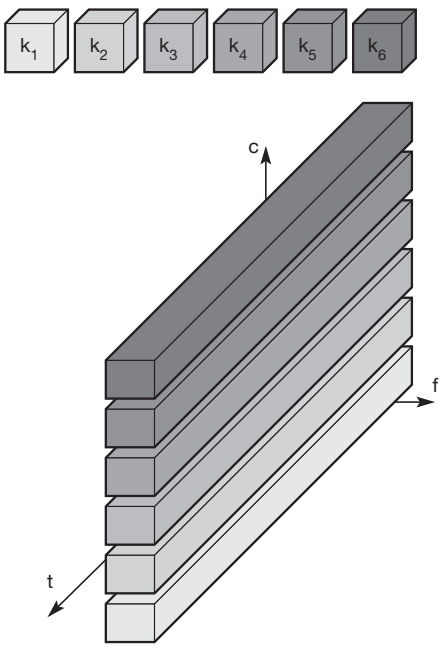

Figure 2.20 Code division multiplexing (CDM)

talking with limited transmit power). But as soon as another code, i.e., another language, is used, one can tune in to this language and clearly separate communication in this language from all the other languages. (The other languages appear as background noise.) This explains why CDM has built-in security: if the language is unknown, the signals can still be received, but they are useless. By using a secret code (or language), a secure channel can be established in a 'hostile' environment. (At parties this may cause some confusion.) Guard spaces are also of importance in this illustrative example. Using, e.g., Swedish and Norwegian does not really work; the languages are too close. But Swedish and Finnish are 'orthogonal' enough to separate the communication channels.

The main advantage of CDM for wireless transmission is that it gives good protection against interference and tapping. Different codes have to be assigned, but code space is huge compared to the frequency space. Assigning individual codes to each sender does not usually cause problems. The main disadvantage of this scheme is the relatively high complexity of the receiver (see section 3.5). A receiver has to know the code and must separate the channel with user data from the background noise composed of other signals and environmental noise. Additionally, a receiver must be precisely synchronized with the transmitter to apply the decoding correctly. The voice example also gives a hint to another problem of CDM receivers. All signals should reach a receiver with almost equal strength, otherwise some signals could drain others. If some people close to a receiver talk very loudly the language does not matter. The receiver cannot listen to any other person. To apply CDM, precise power control is required.

2.6 Modulation

Section 2.2 introduced the basic function of a sine wave which already indicates the three basic modulation schemes (typically, the cosine function is used for explanation):

$$g(t) = A_t \cos(2\pi f_t t + \varphi_t)$$

This function has three parameters: amplitude A_t, frequency f_t, and phase φ_t which may be varied in accordance with data or another modulating signal. For **digital modulation**, which is the main topic in this section, digital data (0 and 1) is translated into an analog signal (baseband signal). Digital modulation is required if digital data has to be transmitted over a medium that only allows for analog transmission. One example for wired networks is the old analog telephone system – to connect a computer to this system a modem is needed. The modem then performs the translation of digital data into analog signals and vice versa. Digital transmission is used, for example, in wired local area networks or within a computer (Halsall, 1996), (Stallings, 1997). In wireless networks, however, digital transmission cannot be used. Here, the binary bit-stream has to be translated into an analog signal first. The three basic methods for this translation are **amplitude shift keying (ASK)**, **frequency shift keying (FSK)**, and **phase shift keying (PSK)**. These are discussed in more detail in the following sections.

Apart from the translation of digital data into analog signals, wireless transmission requires an additional modulation, an **analog modulation** that shifts the center frequency of the baseband signal generated by the digital modulation up to the radio carrier. For example, digital modulation translates a 1 Mbit/s bit-stream into a baseband signal with a bandwidth of 1 MHz. There are several reasons why this baseband signal cannot be directly transmitted in a wireless system:

- **Antennas**: As shown in section 2.3, an antenna must be the order of magnitude of the signal's wavelength in size to be effective. For the 1 MHz signal in the example this would result in an antenna some hundred meters high, which is obviously not very practical for handheld devices. With 1 GHz, antennas a few centimeters in length can be used.
- **Frequency division multiplexing**: Using only baseband transmission, FDM could not be applied. Analog modulation shifts the baseband signals to different carrier frequencies as required in section 2.5.2. The higher the carrier frequency, the more bandwidth that is available for many baseband signals.
- **Medium characteristics**: Path-loss, penetration of obstacles, reflection, scattering, and diffraction – all the effects discussed in section 2.4 depend heavily on the wavelength of the signal. Depending on the application, the right carrier frequency with the desired characteristics has to be chosen: long waves for submarines, short waves for handheld devices, very short waves for directed microwave transmission etc.

As for digital modulation, three different basic schemes are known for analog modulation: **amplitude modulation (AM)**, **frequency modulation (FM)**, and **phase modulation (PM)**. The reader is referred to Halsall (1996) and Stallings (2002) for more details about these analog modulation schemes.

Figure 2.21 shows a (simplified) block diagram of a radio transmitter for digital data. The first step is the digital modulation of data into the analog baseband signal according to one of the schemes presented in the following

sections. The analog modulation then shifts the center frequency of the analog signal up to the radio carrier. This signal is then transmitted via the antenna.

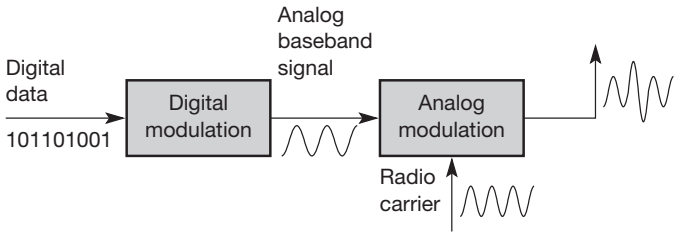

Figure 2.21 Modulation in a transmitter

The receiver (see Figure 2.22) receives the analog radio signal via its antenna and demodulates the signal into the analog baseband signal with the help of the known carrier. This would be all that is needed for an analog radio tuned in to a radio station. (The analog baseband signal would constitute the music.) For digital data, another step is needed. Bits or frames have to be detected, i.e., the receiver must synchronize with the sender. How synchronization is achieved, depends on the digital modulation scheme. After synchronization, the receiver has to decide if the signal represents a digital 1 or a 0, reconstructing the original data.

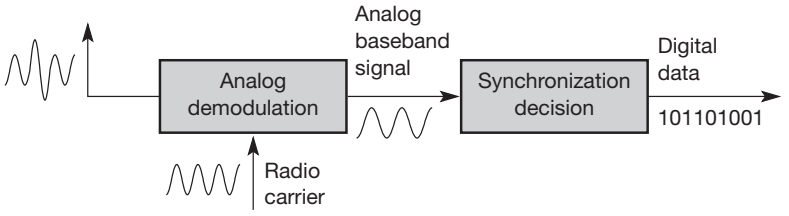

Figure 2.22 Demodulation and data reconstruction in a receiver

The digital modulation schemes presented in the following sections differ in many issues, such as **spectral efficiency** (i.e., how efficiently the modulation scheme utilizes the available frequency spectrum), **power efficiency** (i.e., how much power is needed to transfer bits – which is very important for portable devices that are battery dependent), and **robustness** to multi-path propagation, noise, and interference (Wesel, 1998).

2.6.1 Amplitude shift keying

Figure 2.23 illustrates **amplitude shift keying (ASK)**, the most simple digital modulation scheme. The two binary values, 1 and 0, are represented by two different amplitudes. In the example, one of the amplitudes is 0 (representing the binary 0). This simple scheme only requires low bandwidth, but is very susceptible to interference. Effects like multi-path propagation, noise, or path loss heavily influence the amplitude. In a wireless environment, a constant amplitude

cannot be guaranteed, so ASK is typically not used for wireless radio transmission. However, the wired transmission scheme with the highest performance, namely optical transmission, uses ASK. Here, a light pulse may represent a 1, while the absence of light represents a 0. The carrier frequency in optical systems is some hundred THz. ASK can also be applied to wireless infra red transmission, using a directed beam or diffuse light (see chapter 7, Wireless LANs).

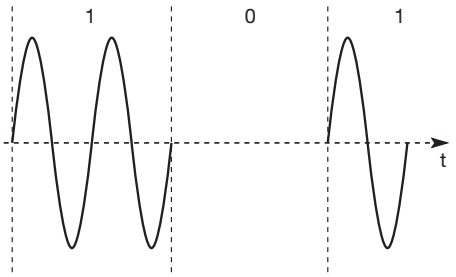

Figure 2.23 Amplitude shift keying (ASK)

2.6.2 Frequency shift keying

A modulation scheme often used for wireless transmission is **frequency shift keying (FSK)** (see Figure 2.24). The simplest form of FSK, also called **binary FSK (BFSK)**, assigns one frequency f_1 to the binary 1 and another frequency f_2 to the binary 0. A very simple way to implement FSK is to switch between two oscillators, one with the frequency f_1 and the other with f_2, depending on the input. To avoid sudden changes in phase, special frequency modulators with **continuous phase modulation**, **(CPM)** can be used. Sudden changes in phase cause high frequencies, which is an undesired side-effect.

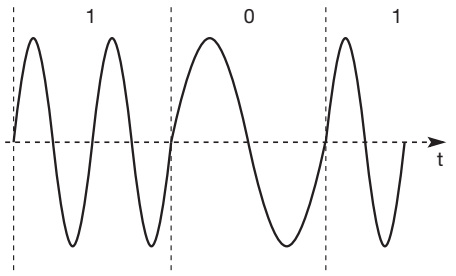

Figure 2.24 Frequency shift keying (FSK)

A simple way to implement demodulation is by using two bandpass filters, one for f_1 the other for f_2. A comparator can then compare the signal levels of the filter outputs to decide which of them is stronger. FSK needs a larger bandwidth compared to ASK but is much less susceptible to errors.

2.6.3 Phase shift keying

Finally, **phase shift keying (PSK)** uses shifts in the phase of a signal to represent data. Figure 2.25 shows a phase shift of 180° or π as the 0 follows the 1 (the same happens as the 1 follows the 0). This simple scheme, shifting the phase by 180° each time the value of data changes, is also called **binary PSK (BPSK)**. A simple

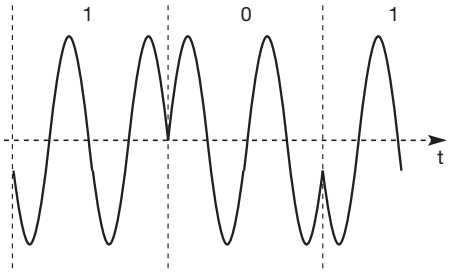

Figure 2.25 Phase shift keying (PSK)

implementation of a BPSK modulator could multiply a frequency f with +1 if the binary data is 1 and with –1 if the binary data is 0.

To receive the signal correctly, the receiver must synchronize in frequency and phase with the transmitter. This can be done using a **phase lock loop (PLL)**. Compared to FSK, PSK is more resistant to interference, but receiver and transmitter are also more complex.

2.6.4 Advanced frequency shift keying

A famous FSK scheme used in many wireless systems is **minimum shift keying (MSK)**. MSK is basically BFSK without abrupt phase changes, i.e., it belongs to CPM schemes. Figure 2.26 shows an example for the implementation of MSK. In a first step, data bits are separated into even and odd bits, the duration of each bit being doubled. The scheme also uses two frequencies: f_1, the lower frequency, and f_2, the higher frequency, with $f_2 = 2f_1$.

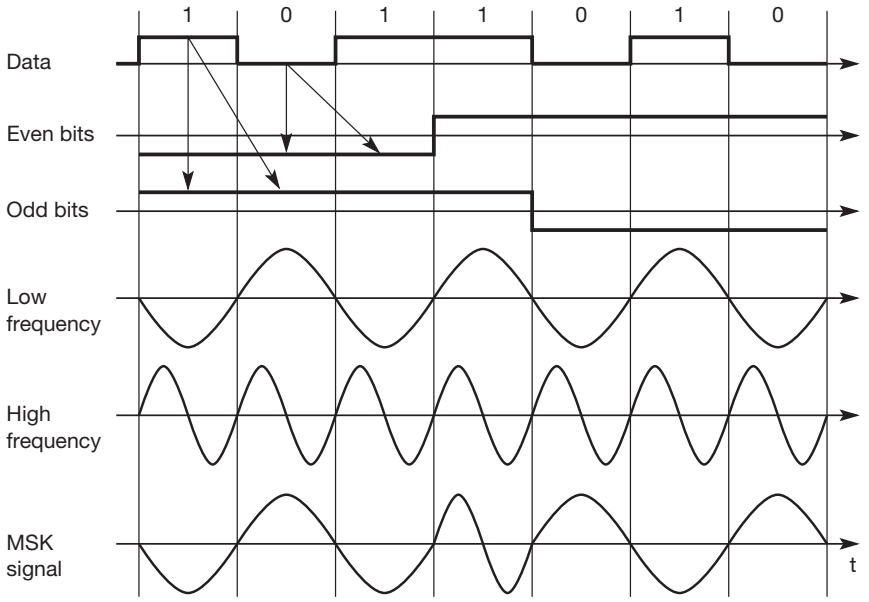

Figure 2.26 Minimum shift keying (MSK)

According to the following scheme, the lower or higher frequency is chosen (either inverted or non-inverted) to generate the MSK signal:

- if the even and the odd bit are both 0, then the higher frequency f_2 is inverted (i.e., f_2 is used with a phase shift of 180°);
- if the even bit is 1, the odd bit 0, then the lower frequency f_1 is inverted. This is the case, e.g., in the fifth to seventh columns of Figure 2.26,

- if the even bit is 0 and the odd bit is 1, as in columns 1 to 3, f_1 is taken without changing the phase,
- if both bits are 1 then the original f_2 is taken.

A high frequency is always chosen if even and odd bits are equal. The signal is inverted if the odd bit equals 0. This scheme avoids all phase shifts in the resulting MSK signal.

Adding a so-called Gaussian lowpass filter to the MSK scheme results in **Gaussian MSK (GMSK)**, which is the digital modulation scheme for many European wireless standards (see chapter 4 for GSM, DECT). The filter reduces the large spectrum needed by MSK.

2.6.5 Advanced phase shift keying

The simple PSK scheme can be improved in many ways. The basic BPSK scheme only uses one possible phase shift of 180°. The left side of Figure 2.27 shows BPSK in the phase domain (which is typically the better representation compared to the time domain in Figure 2.25). The right side of Figure 2.27 shows **quadrature PSK (QPSK)**, one of the most common PSK schemes (sometimes also called quaternary PSK). Here, higher bit rates can be achieved for the same bandwidth by coding two bits into one phase shift. Alternatively, one can reduce the bandwidth and still achieve the same bit rates as for BPSK.

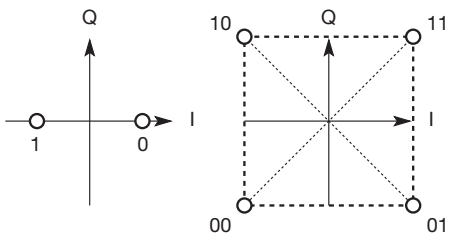

Figure 2.27 BPSK and QPSK in the phase domain

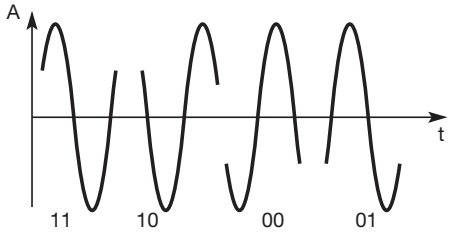

Figure 2.28 QPSK in the time domain

QPSK (and other PSK schemes) can be realized in two variants. The phase shift can always be relative to a **reference signal** (with the same frequency). If this scheme is used, a phase shift of 0 means that the signal is in phase with the reference signal. A QPSK signal will then exhibit a phase shift of 45° for the data 11, 135° for 10, 225° for 00, and 315° for 01 – with all phase shifts being relative to the reference signal. The transmitter 'selects' parts of the signal as shown in Figure 2.28 and concatenates them. To reconstruct data, the receiver has to compare the incoming signal with the reference signal. One problem of this scheme involves producing a reference signal at the receiver. Transmitter and receiver have to be synchronized very often, e.g., by using special synchronization patterns before user data arrives or via a pilot frequency as reference.

One way to avoid this problem is to use **differential QPSK (DQPSK)**. Here the phase shift is not relative to a reference signal but to the phase of the previous two bits. In this case, the receiver does not need the reference signal but only compares two signals to reconstruct data. DQPSK is used in US wireless technologies IS-136 and PACS and in Japanese PHS.

One could now think of extending the scheme to more and more angles for shifting the phase. For instance, one can think of coding 3 bits per phase shift using 8 angles. Additionally, the PSK scheme could be combined with ASK as is done for example in **quadrature amplitude modulation (QAM)** for standard 9,600 bit/s modems (left side of Figure 2.29). Here, three different amplitudes and 12 angles are combined coding 4 bits per phase/amplitude change. Problems occur for wireless communication in case of noise or ISI. The more 'points' used in the phase domain, the harder it is to separate them. DQPSK has been proven as one of the most efficient schemes under these considerations (Wesel, 1998).

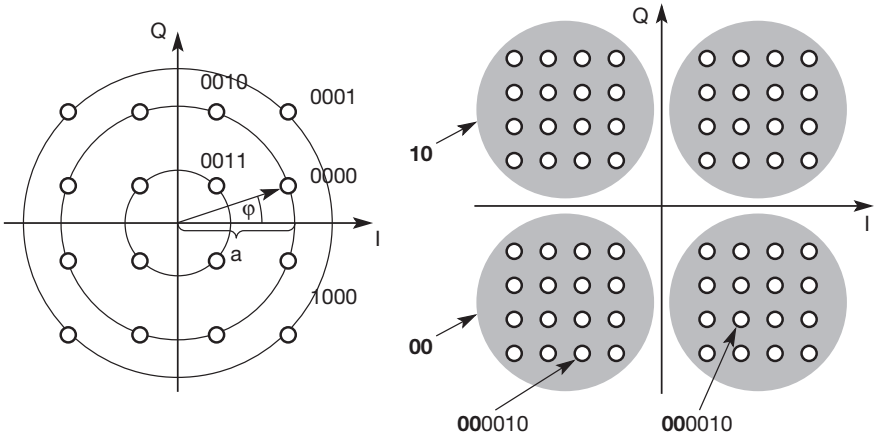

Figure 2.29 16 quadrature amplitude modulation and hierarchical 64 QAM

A more advanced scheme is a hierarchical modulation as used in the digital TV standard DVB-T. The right side of Figure 2.29 shows a 64 QAM that contains a QPSK modulation. A 64 QAM can code 6 bit per symbol. Here the two most significant bits are used for the QPSK signal embedded in the QAM signal. If the reception of the signal is good the entire QAM constellation can be resolved. Under poor reception conditions, e.g., with moving receivers, only the QPSK portion can be resolved. A high priority data stream in DVB-T is coded with QPSK using the two most significant bits. The remaining 4 bits represent low priority data. For TV this could mean that the standard resolution data stream is coded with high priority, the high resolution information with low priority. If the signal is distorted, at least the standard TV resolution can be received.

2.6.6 Multi-carrier modulation

Special modulation schemes that stand somewhat apart from the others are **multi-carrier modulation (MCM)**, **orthogonal frequency division multiplexing (OFDM)** or **coded OFDM (COFDM)** that are used in the context of the European digital radio system DAB (see section 6.3) and the WLAN standards IEEE 802.11a and HiperLAN2 (see chapter 7). The main attraction of MCM is its good ISI mitigation property. As explained in section 2.4.3, higher bit rates are more vulnerable to ISI. MCM splits the high bit rate stream into many lower bit rate streams (see Figure 2.30), each stream being sent using an independent carrier frequency. If, for example, n symbols/s have to be transmitted, each subcarrier transmits n/c symbols/s with c being the number of subcarriers. One symbol could, for example represent 2 bit as in QPSK. DAB, for example, uses between 192 and 1,536 of these subcarriers. The physical layer of HiperLAN2 and IEEE 802.11a uses 48 subcarriers for data.

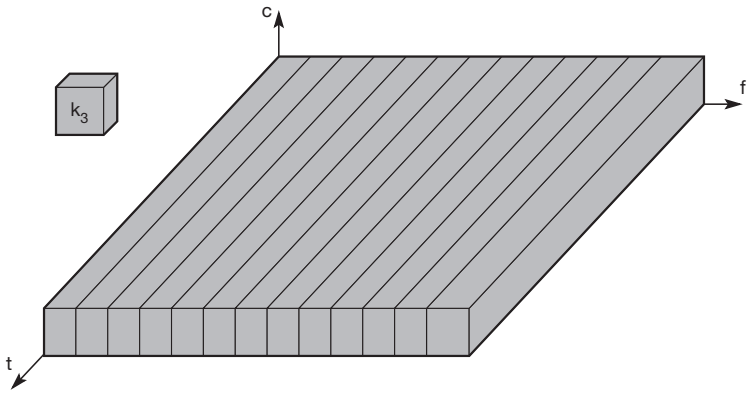

Figure 2.30 Parallel data transmission on several subcarriers with lower rate

Figure 2.31 shows the superposition of orthogonal frequencies. The maximum of one subcarrier frequency appears exactly at a frequency where all other subcarriers equal zero.

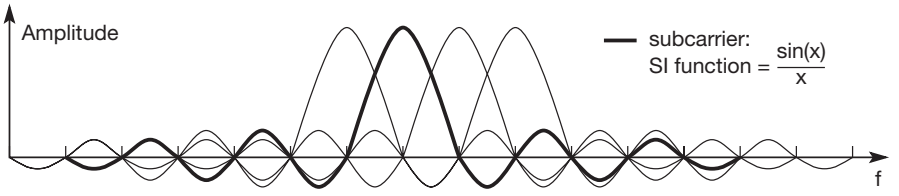

Figure 2.31 Superposition of orthogonal frequencies

Using this scheme, frequency selective fading only influences some subcarriers, and not the whole signal – an additional benefit of MCM. Typically, MCM transmits symbols with guard spaces between single symbols or groups of symbols. This helps the receiver to handle multi-path propagation. OFDM is a special

method of implementing MCM using orthogonal carriers. Computationally, this is a very efficient algorithm based on fast Fourier transform (FFT) for modulation/demodulation. If additional error-control coding across the symbols in different subcarriers is applied, the system is referred to as COFDM. More details about the implementation of MCM, OFDM, and COFDM can be found in Wesel (1998), Pahlavan (2002), ETSI (1997) and in section 6.3 or chapter 7.

2.7 Spread spectrum

As the name implies, **spread spectrum** techniques involve spreading the bandwidth needed to transmit data – which does not make sense at first sight. Spreading the bandwidth has several advantages. The main advantage is the resistance to **narrowband interference**. In Figure 2.32, diagram i) shows an idealized narrowband signal from a sender of user data (here power density dP/df versus frequency f). The sender now spreads the signal in step ii), i.e., converts the narrowband signal into a broadband signal. The energy needed to transmit the signal (the area shown in the diagram) is the same, but it is now spread over a larger frequency range. The power level of the spread signal can be much lower than that of the original narrowband signal without losing data. Depending on the generation and reception of the spread signal, the power level of the user signal can even be as low as the background noise. This makes it difficult to distinguish the user signal from the background noise and thus hard to detect.

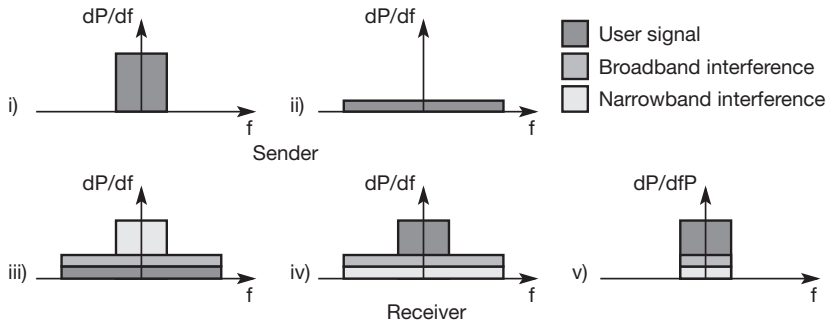

Figure 2.32 Spread spectrum: spreading and despreading

During transmission, narrowband and broadband interference add to the signal in step iii). The sum of interference and user signal is received. The receiver now knows how to despread the signal, converting the spread user signal into a narrowband signal again, while spreading the narrowband interference and leaving the broadband interference. In step v) the receiver applies a bandpass filter to cut off frequencies left and right of the narrowband signal. Finally, the receiver can reconstruct the original data because the power level of the user signal is high enough, i.e., the signal is much stronger than the remaining interference. The following sections show how spreading can be performed.

Just as spread spectrum helps to deal with narrowband interference for a single channel, it can be used for several channels. Consider the situation shown in Figure 2.33. Six different channels use FDM for multiplexing, which means that each channel has its own narrow frequency band for transmission. Between each frequency band a guard space is needed to avoid adjacent channel interference. As mentioned in section 2.5.2, this method requires careful frequency planning. Additionally, Figure 2.33 depicts a certain channel quality. This is frequency dependent and is a measure for interference at this frequency. Channel quality also changes over time – the diagram only shows a snapshot at one moment. Depending on receiver characteristics, channels 1, 2, 5, and 6 could be received while the quality of channels 3 and 4 is too bad to reconstruct transmitted data. Narrowband interference destroys the transmission of channels 3 and 4. This illustration only represents a snapshot and the situation could be completely different at the next moment. All in all, communication may be very difficult using such narrowband signals.

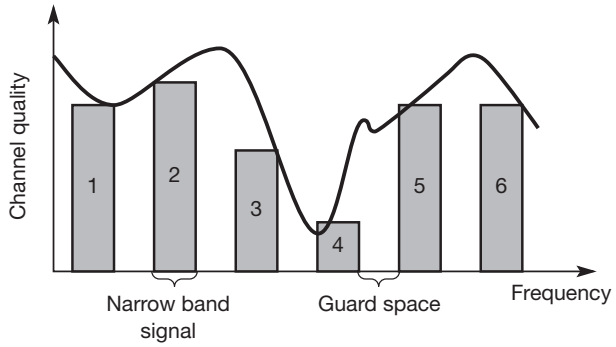

Figure 2.33 Narrowband interference without spread spectrum

How can spread spectrum help in such a situation? As already shown, spread spectrum can increase resistance to narrowband interference. The same technique is now applied to all narrowband signals. As shown in Figure 2.34, all narrowband signals are now spread into broadband signals using the same frequency range. No more frequency planning is needed (under these simplified assumptions), and all senders use the same frequency band. But how can receivers recover their signal?

To separate different channels, CDM is now used instead of FDM. This application shows the tight coupling of CDM and spread spectrum (explained in more detail in chapter 3). Spreading of a narrowband signal is achieved using a special code as shown in sections 2.7.1 and 2.7.2. Each channel is allotted its own code, which the receivers have to apply to recover the signal. Without knowing the code, the signal cannot be recovered and behaves like background noise. This is the security effect of spread spectrum if a secret code is used for

spreading. Features that make spread spectrum and CDM very attractive for military applications are the coexistence of several signals without coordination (apart from the fact that the codes must have certain properties), robustness against narrowband interference, relative high security, and a characteristic like background noise. Only the appropriate (secret) codes have to be exchanged.

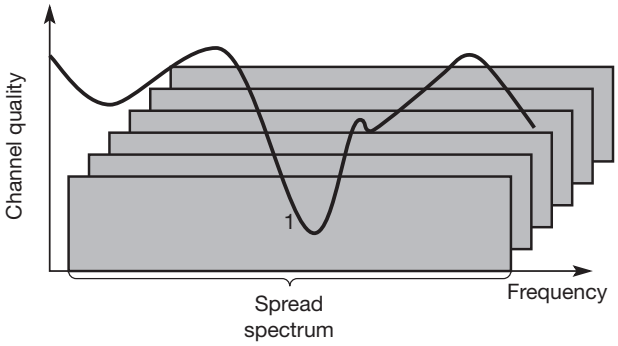

Figure 2.34 Spread spectrum to avoid narrowband interference

Apart from military uses, the combination of spread spectrum and CDM is becoming more and more attractive for everyday applications. As mentioned before, frequencies are a scarce resource around the world, particularly license-free bands. Spread spectrum now allows an overlay of new transmission technology at exactly the same frequency at which current narrowband systems are already operating. This is used by US mobile phone systems. While the frequency band around 850 MHz had already been in use for TDM and FDM systems (AMPS and IS-54), the introduction of a system using CDM (IS-95) was still possible.

Spread spectrum technologies also exhibit drawbacks. One disadvantage is the increased complexity of receivers that have to despread a signal. Today despreading can be performed up to high data rates thanks to digital signal processing. Another problem is the large frequency band that is needed due to the spreading of the signal. Although spread signals appear more like noise, they still raise the background noise level and may interfere with other transmissions if no special precautions are taken.

Spreading the spectrum can be achieved in two different ways as shown in the following two sections.

2.7.1 Direct sequence spread spectrum

Direct sequence spread spectrum (DSSS) systems take a user bit stream and perform an (XOR) with a so-called **chipping sequence** as shown in Figure 2.35. The example shows that the result is either the sequence 0110101 (if the user bit equals 0) or its complement 1001010 (if the user bit equals 1). While each user bit has a duration t_b, the chipping sequence consists of smaller pulses, called **chips**, with a duration t_c. If the chipping sequence is generated properly it

appears as random noise: this sequence is also sometimes called **pseudo-noise** sequence. The **spreading factor** $s = t_b/t_c$ determines the bandwidth of the resulting signal. If the original signal needs a bandwidth w, the resulting signal needs s·w after spreading. While the spreading factor of the very simple example is only 7 (and the chipping sequence 0110101 is not very random), civil applications use spreading factors between 10 and 100, military applications use factors of up to 10,000. Wireless LANs complying with the standard IEEE 802.11 (see section 7.3) use, for example, the sequence 10110111000, a so-called Barker code, if implemented using DSSS. Barker codes exhibit a good robustness against interference and insensitivity to multi-path propagation. Other known Barker codes are 11, 110, 1110, 11101, 1110010, and 1111100110101 (Stallings, 2002).

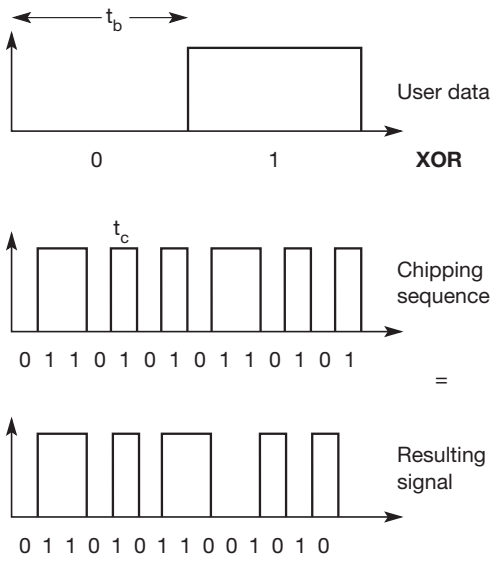

Figure 2.35
Spreading with DSSS

Up to now only the spreading has been explained. However, transmitters and receivers using DSSS need additional components as shown in the simplified block diagrams in Figure 2.36 and Figure 2.37. The first step in a DSSS transmitter, Figure 2.36 is the spreading of the user data with the chipping sequence (**digital modulation**). The spread signal is then modulated with a radio carrier as explained in section 2.6 (**radio modulation**). Assuming for example a user signal with a bandwidth of 1 MHz. Spreading with the above 11-chip Barker code would result in a signal with 11 MHz bandwidth. The radio carrier then shifts this signal to the carrier frequency (e.g., 2.4 GHz in the ISM band). This signal is then transmitted.

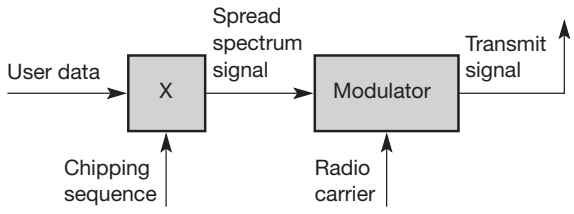

Figure 2.36
DSSS transmitter

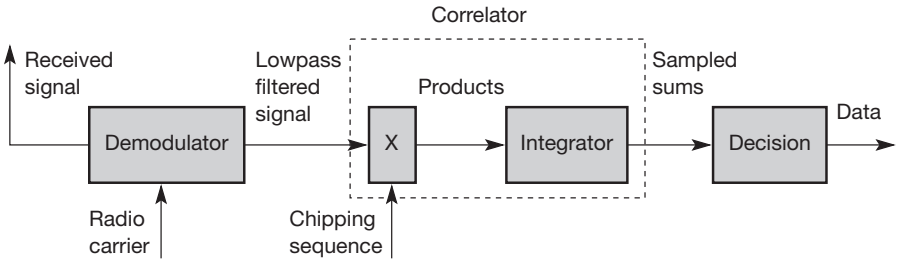

Figure 2.37
DSSS receiver

The DSSS receiver is more complex than the transmitter. The receiver only has to perform the inverse functions of the two transmitter modulation steps. However, noise and multi-path propagation require additional mechanisms to reconstruct the original data. The first step in the receiver involves demodulating the received signal. This is achieved using the same carrier as the transmitter reversing the modulation and results in a signal with approximately the same bandwidth as the original spread spectrum signal. Additional filtering can be applied to generate this signal.

While demodulation is well known from ordinary radio receivers, the next steps constitute a real challenge for DSSS receivers, contributing to the complexity of the system. The receiver has to know the original chipping sequence, i.e., the receiver basically generates the same pseudo random sequence as the transmitter. Sequences at the sender and receiver have to be precisely synchronized because the receiver calculates the product of a chip with the incoming signal. This comprises another XOR operation as explained in section 3.5, together with a medium access mechanism that relies on this scheme. During a bit period, which also has to be derived via synchronization, an **integrator** adds all these products. Calculating the products of chips and signal, and adding the products in an integrator is also called correlation, the device a **correlator**. Finally, in each bit period a **decision unit** samples the sums generated by the integrator and decides if this sum represents a binary 1 or a 0.

If transmitter and receiver are perfectly synchronized and the signal is not too distorted by noise or multi-path propagation, DSSS works perfectly well according to the simple scheme shown. Sending the user data 01 and applying the 11-chip Barker code 10110111000 results in the spread 'signal' 10110111000001001000111. On the receiver side, this 'signal' is XORed bit-wise after demodulation with the same Barker code as chipping sequence. This results in the sum of products equal to 0 for the first bit and to 11 for the second bit. The decision unit can now map the first sum (=0) to a binary 0, the second sum (=11) to a binary 1 – this constitutes the original user data.

In real life, however, the situation is somewhat more complex. Assume that the demodulated signal shows some distortion, e.g., 101**0**01**0**100001**1**01000111. The sum of products for the first bit would be 2, 10 for the second bit. Still, the decision unit can map, e.g., sums less than 4 to a binary 0 and sums larger than

7 to a binary 1. However, it is important to stay synchronized with the transmitter of a signal. But what happens in case of multi-path propagation? Then several paths with different delays exist between a transmitter and a receiver. Additionally, the different paths may have different path losses. In this case, using so-called rake receivers provides a possible solution. A **rake receiver** uses n correlators for the n strongest paths. Each correlator is synchronized to the transmitter plus the delay on that specific path. As soon as the receiver detects a new path which is stronger than the currently weakest path, it assigns this new path to the correlator with the weakest path. The output of the correlators are then combined and fed into the decision unit. Rake receivers can even take advantage of the multi-path propagation by combining the different paths in a constructive way (Viterbi, 1995).

2.7.2 Frequency hopping spread spectrum

For **frequency hopping spread spectrum (FHSS)** systems, the total available bandwidth is split into many channels of smaller bandwidth plus guard spaces between the channels. Transmitter and receiver stay on one of these channels for a certain time and then hop to another channel. This system implements FDM and TDM. The pattern of channel usage is called the **hopping sequence**, the time spend on a channel with a certain frequency is called the **dwell time**. FHSS comes in two variants, slow and fast hopping (see Figure 2.38).

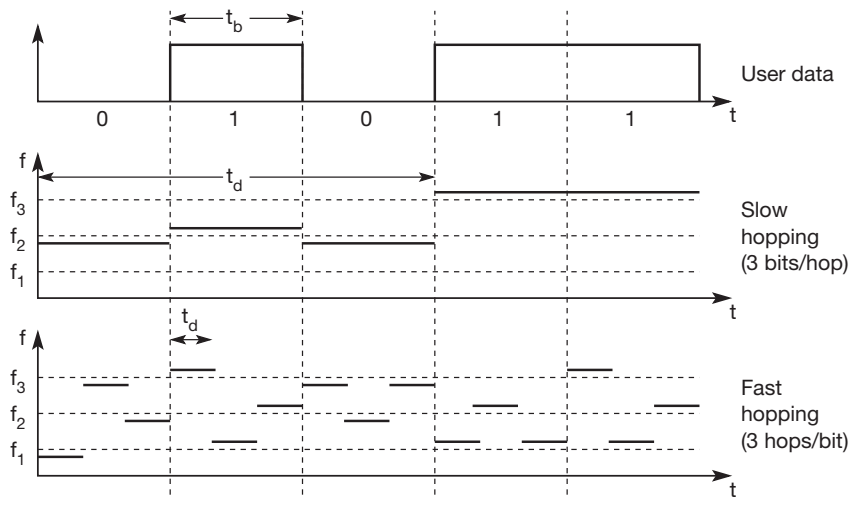

Figure 2.38 Slow and fast frequency hopping

In **slow hopping**, the transmitter uses one frequency for several bit periods.[3] Figure 2.38 shows five user bits with a bit period t_b. Performing slow hopping, the transmitter uses the frequency f_2 for transmitting the first three bits during the dwell time t_d. Then, the transmitter hops to the next frequency f_3. Slow hopping systems are typically cheaper and have relaxed tolerances, but they are not as immune to narrowband interference as fast hopping systems. Slow frequency hopping is an option for GSM (see section 4.1).

For **fast hopping** systems, the transmitter changes the frequency several times during the transmission of a single bit. In the example of Figure 2.38, the transmitter hops three times during a bit period. Fast hopping systems are more complex to implement because the transmitter and receiver have to stay synchronized within smaller tolerances to perform hopping at more or less the same points in time. However, these systems are much better at overcoming the effects of narrowband interference and frequency selective fading as they only stick to one frequency for a very short time.

Another example of an FHSS system is Bluetooth, which is presented in section 7.5. Bluetooth performs 1,600 hops per second and uses 79 hop carriers equally spaced with 1 MHz in the 2.4 GHz ISM band.

Figures 2.39 and 2.40 show simplified block diagrams of FHSS transmitters and receivers respectively. The first step in an FHSS transmitter is the modulation of user data according to one of the digital-to-analog modulation schemes, e.g., FSK or BPSK, as discussed in section 2.6. This results in a narrowband signal, if FSK is used with a frequency f_0 for a binary 0 and f_1 for a binary 1. In the next step, frequency hopping is performed, based on a hopping sequence. The hopping sequence is fed into a frequency synthesizer generating the carrier frequencies f_i. A second modulation uses the modulated narrowband signal and the carrier frequency to generate a new spread signal with frequency of f_i+f_0 for a 0 and f_i+f_1 for a 1 respectively. If different FHSS transmitters use hopping sequences that never overlap, i.e., if two transmitters never use the same frequency f_i at the same time, then these two transmissions do not interfere. This requires the coordination of all transmitters and their hopping sequences. As for DSSS systems, pseudo-random hopping sequences can also be used without coordination. These sequences only have to fulfill certain properties to keep interference minimal.[4] Two or more transmitters may choose the same frequency for a hop, but dwell time is short for fast hopping systems, so interference is minimal.

The receiver of an FHSS system has to know the hopping sequence and must stay synchronized. It then performs the inverse operations of the modulation to reconstruct user data. Several filters are also needed (these are not shown in the simplified diagram in Figure 2.40).

---

3 Another definition refers to the number of hops per signal element instead of bits.

4 These sequences should have a low cross-correlation. More details are given in section 3.5.

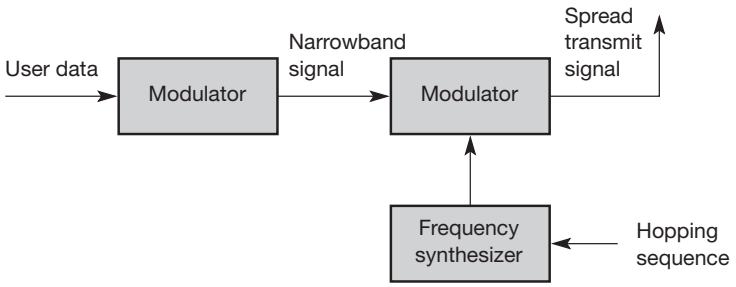

Figure 2.39
FHSS transmitter

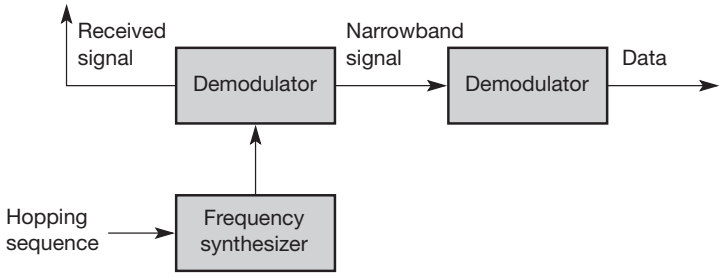

Figure 2.40
FHSS receiver

Compared to DSSS, spreading is simpler using FHSS systems. FHSS systems only use a portion of the total band at any time, while DSSS systems always use the total bandwidth available. DSSS systems on the other hand are more resistant to fading and multi-path effects. DSSS signals are much harder to detect – without knowing the spreading code, detection is virtually impossible. If each sender has its own pseudo-random number sequence for spreading the signal (DSSS or FHSS), the system implements CDM. More details about spread spectrum applications and their theoretical background can be found in Viterbi (1995), Peterson (1995), Ojanperä (1998), and Dixon (1994).

2.8 Cellular systems

Cellular systems for mobile communications implement SDM. Each transmitter, typically called a **base station**, covers a certain area, a **cell**. Cell radii can vary from tens of meters in buildings, and hundreds of meters in cities, up to tens of kilometers in the countryside. The shape of cells are never perfect circles or hexagons (as shown in Figure 2.41), but depend on the environment (buildings, mountains, valleys etc.), on weather conditions, and sometimes even on system load. Typical systems using this approach are mobile telecommunication systems (see chapter 4), where a mobile station within the cell around a base station communicates with this base station and vice versa.

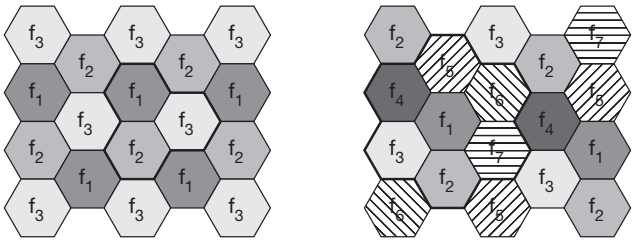

Figure 2.41
Cellular system with three and seven cell clusters

In this context, the question arises as to why mobile network providers install several thousands of base stations throughout a country (which is quite expensive) and do not use powerful transmitters with huge cells like, e.g., radio stations, use.

Advantages of cellular systems with small cells are the following:

- **Higher capacity**: Implementing SDM allows frequency reuse. If one transmitter is far away from another, i.e., outside the interference range, it can reuse the same frequencies. As most mobile phone systems assign frequencies to certain users (or certain hopping patterns), this frequency is blocked for other users. But frequencies are a scarce resource and, the number of concurrent users per cell is very limited. Huge cells do not allow for more users. On the contrary, they are limited to less possible users per km^2. This is also the reason for using very small cells in cities where many more people use mobile phones.
- **Less transmission power**: While power aspects are not a big problem for base stations, they are indeed problematic for mobile stations. A receiver far away from a base station would need much more transmit power than the current few Watts. But energy is a serious problem for mobile handheld devices.
- **Local interference only**: Having long distances between sender and receiver results in even more interference problems. With small cells, mobile stations and base stations only have to deal with 'local' interference.
- **Robustness**: Cellular systems are decentralized and so, more robust against the failure of single components. If one antenna fails, this only influences communication within a small area.

Small cells also have some **disadvantages**:

- **Infrastructure needed**: Cellular systems need a complex infrastructure to connect all base stations. This includes many antennas, switches for call forwarding, location registers to find a mobile station etc, which makes the whole system quite expensive.

- **Handover needed**: The mobile station has to perform a handover when changing from one cell to another. Depending on the cell size and the speed of movement, this can happen quite often.
- **Frequency planning**: To avoid interference between transmitters using the same frequencies, frequencies have to be distributed carefully. On the one hand, interference should be avoided, on the other, only a limited number of frequencies is available.

To avoid interference, different transmitters within each other's interference range use FDM. If FDM is combined with TDM (see Figure 2.19), the hopping pattern has to be coordinated. The general goal is never to use the same frequency at the same time within the interference range (if CDM is not applied). Two possible models to create cell patterns with minimal interference are shown in Figure 2.41. Cells are combined in **clusters** – on the left side three cells form a cluster, on the right side seven cells form a cluster. All cells within a cluster use disjointed sets of frequencies. On the left side, one cell in the cluster uses set f_1, another cell f_2, and the third cell f_3. In real-life transmission, the pattern will look somewhat different. The hexagonal pattern is chosen as a simple way of illustrating the model. This pattern also shows the repetition of the same frequency sets. The transmission power of a sender has to be limited to avoid interference with the next cell using the same frequencies.

To reduce interference even further (and under certain traffic conditions, i.e., number of users per km^2) **sectorized antennas** can be used. Figure 2.42 shows the use of three sectors per cell in a cluster with three cells. Typically, it makes sense to use sectorized antennas instead of omni-directional antennas for larger cell radii.

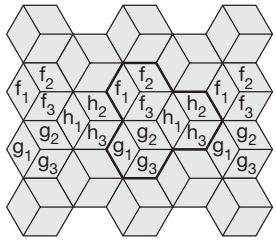

Figure 2.42
Cellular system with three cell clusters and three sectors per cell

The fixed assignment of frequencies to cell clusters and cells respectively, is not very efficient if traffic load varies. For instance, in the case of a heavy load in one cell and a light load in a neighboring cell, it could make sense to 'borrow' frequencies. Cells with more traffic are dynamically allotted more frequencies. This scheme is known as **borrowing channel allocation (BCA)**, while the first fixed scheme is called **fixed channel allocation (FCA)**. FCA is used in the GSM system as it is much simpler to use, but it requires careful traffic analysis before installation.

A **dynamic channel allocation (DCA)** scheme has been implemented in DECT (see section 4.2). In this scheme, frequencies can only be borrowed, but it is also possible to freely assign frequencies to cells. With dynamic assignment of frequencies to cells, the danger of interference with cells using the same frequency exists. The 'borrowed' frequency can be blocked in the surrounding cells.

Cellular systems using CDM instead of FDM do not need such elaborate channel allocation schemes and complex frequency planning. Here, users are separated through the code they use, not through the frequency. Cell planning faces another problem – the cell size depends on the current load. Accordingly, **CDM cells** are commonly said to '**breathe**'. While a cell can cover a larger area under a light load, it shrinks if the load increases. The reason for this is the growing noise level if more users are in a cell. (Remember, if you do not know the code, other signals appear as noise, i.e., more and more people join the party.) The higher the noise, the higher the path loss and the higher the transmission errors. Finally, mobile stations further away from the base station drop out of the cell. (This is similar to trying to talk to someone far away at a crowded party.) Figure 2.43 illustrates this phenomenon with a user transmitting a high bit rate stream within a CDM cell. This additional user lets the cell shrink with the result that two users drop out of the cell. In a real-life scenario this additional user could request a video stream (high bit rate) while the others use standard voice communication (low bit rate).

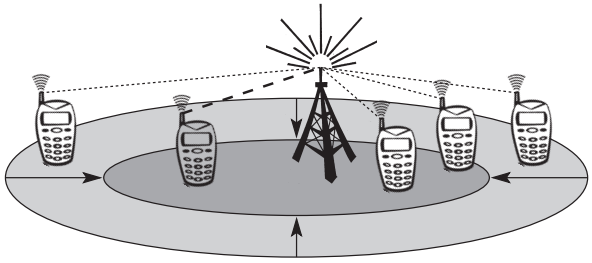

Figure 2.43 Cell breathing depending on the current load

2.9 Summary

This chapter introduced the basics of wireless communications, leaving out most formulae found in books dedicated to wireless transmission and the effects of radio propagation. However, the examples, mechanisms, and problems discussed will hopefully give the reader a good idea as to why wireless communication is fundamentally different from wired communication and why protocols and applications on higher layers have to follow different principles to take the missing wire into account.

A topic of worldwide importance is the regulation and harmonization of frequencies used for radio transmission. The chapter showed many different systems using either different or the same frequencies. Hopefully, the future will bring more frequencies which are available worldwide to avoid more expensive multi-mode devices. At least some harmonization has taken and continues to take place in the area of WLANs (see chapter 7) and 3G mobile phone systems (see chapter 4).

As electromagnetic waves are the basis for wireless communication, antennas are needed for the transmission and reception of waves. While base stations of mobile phone systems often use directed antennas, omni-directional antennas are the choice for mobile devices. On the way from sender to receiver, many things can happen to electromagnetic waves. The standard effects, such as shadowing, fading, reflection, diffraction, and scattering have been presented. All these effects lead to one of the biggest problems in wireless communication: multi-path propagation. Multi-path propagation limits the bandwidth of a channel due to intersymbol interference, i.e., one symbol is 'smeared' into another symbol due to delay spread.

As we only have one 'medium' for wireless transmission, several multiplexing schemes can be applied to raise overall capacity. The standard schemes are SDM, FDM, TDM, and CDM. To achieve FDM, data has to be 'translated' into a signal with a certain carrier frequency. Therefore, two modulation steps can be applied. Digital modulation encodes data into a baseband signal, whereas analog modulation then shifts the centre frequency of the signal up to the radio carrier. Some advanced schemes have been presented that can code many bits into a single phase shift, raising the efficiency.

With the help of spread spectrum technology, several features can be implemented. One is (at least some) security – without knowing the spreading code, the signal appears as noise. As we will see in more detail in chapter 3, spread spectrum lays the basis for special medium access schemes using the code space. Spread spectrum also makes a transmission more robust against narrowband interference, as the signal is spread over a larger bandwidth so, the narrowband interference only influences a small fraction of the signal.

Finally, this chapter has presented the concept of cellular systems. Cellular systems implement SDM to raise the overall capacity of mobile phone systems. While these systems require detailed planning (i.e., matching the cell size with the traffic expected), it presents one of the basic solutions for using the scarce frequency resources efficiently.

2.10 Review exercises

1. Frequency regulations may differ between countries. Check out the regulations valid for your country (within Europe the European Radio Office may be able to help you, www.ero.dk, for the US try the FCC, www.fcc.gov, for Japan ARIB, www.arib.or.jp).
2. Why can waves with a very low frequency follow the earth's surface? Why are they not used for data transmission in computer networks?
3. Why does the ITU-R only regulate 'lower' frequencies (up to some hundred GHz) and not higher frequencies (in the THz range)?

4. What are the two different approaches in regulation regarding mobile phone systems in Europe and the US? What are the consequences?
5. Why is the international availability of the same ISM bands important?
6. Is it possible to transmit a digital signal, e.g., coded as square wave as used inside a computer, using radio transmission without any loss? Why?
7. Is a directional antenna useful for mobile phones? Why? How can the gain of an antenna be improved?
8. What are the main problems of signal propagation? Why do radio waves not always follow a straight line? Why is reflection both useful and harmful?
9. Name several methods for ISI mitigation. How does ISI depend on the carrier frequency, symbol rate, and movement of sender/receiver? What are the influences of ISI on TDM schemes?
10. What are the means to mitigate narrowband interference? What is the complexity of the different solutions?
11. Why, typically, is digital modulation not enough for radio transmission? What are general goals for digital modulation? What are typical schemes?
12. Think of a phase diagram and the points representing bit patterns for a PSK scheme (see Figure 2.29). How can a receiver decide which bit pattern was originally sent when a received 'point' lies somewhere in between other points in the diagram? Why is it difficult to code more and more bits per phase shift?
13. What are the main benefits of a spread spectrum system? How can spreading be achieved? What replaces the guard space in Figure 2.33 when compared to Figure 2.34? How can DSSS systems benefit from multi-path propagation?
14. What are the main reasons for using cellular systems? How is SDM typically realized and combined with FDM? How does DCA influence the frequencies available in other cells?
15. What limits the number of simultaneous users in a TDM/FDM system compared to a CDM system? What happens to the transmission quality of connections if the load gets higher in a cell, i.e., how does an additional user influence the other users in the cell?

2.11 References

Anderson, J.B., Rappaport, T.S., Yoshida, S. (1995) 'Propagation measurements and models for wireless communications channels,' *IEEE Communications Magazine*, 33, (1).

Dixon, R. (1994) *Spread spectrum systems with commercial applications*. John Wiley.

ETSI (1997) *Digital Audio Broadcasting (DAB) to mobile, portable, and fixed receivers*, European Telecommunications Standards Institute, ETS 300 401.

GSM World (2002), GSM Association, http://www.gsmworld.com/.

Halsall, F. (1996) *Data communications, computer networks and open systems*. Addison-Wesley Longman.
Ojanperä, T., Prasad, R. (1998) *Wideband CDMA for Third Generation Mobile Communications*. Artech House.
Pahlavan, K., Krishnamurthy, P. (2002) *Principles of Wireless Network*. Prentice Hall.
Peterson, R., Ziemer, R., Borth, D. (1995) *Introduction to spread spectrum communications*. Prentice Hall.
Stallings, W. (1997) *Data and computer communications*. Prentice Hall.
Stallings, W. (2002) *Wireless Communications and Networking*. Prentice Hall.
Viterbi, A. (1995) *CDMA: Principles of spread spectrum communication*. Addison-Wesley Longman.
Wesel, E. (1998) *Wireless multimedia communications: networking video, voice, and data*. Addison-Wesley Longman.

3 Medium access control

This chapter introduces several **medium access control (MAC)** algorithms which are specifically adapted to the wireless domain. Medium access control comprises all mechanisms that regulate user access to a medium using SDM, TDM, FDM, or CDM. MAC is thus similar to traffic regulations in the highway/multiplexing example introduced in chapter 2. The fact that several vehicles use the same street crossing in TDM, for example, requires rules to avoid collisions; one mechanism to enforce these rules is traffic lights. While the previous chapter mainly introduced mechanisms of the physical layer, layer 1, of the ISO/OSI reference model, MAC belongs to layer 2, the **data link control layer (DLC)**. Layer 2 is subdivided into the **logical link control (LLC)**, layer 2b, and the MAC, layer 2a (Halsall, 1996). The task of DLC is to establish a reliable point to point or point to multi-point connection between different devices over a wired or wireless medium. The basic MAC mechanisms are introduced in the following sections, whereas LLC and higher layers, as well as specific relevant technologies will be presented in later chapters together with mobile and wireless systems.

This chapter aims to explain why special MACs are needed in the wireless domain and why standard MAC schemes known from wired networks often fail. (In contrast to wired networks, hidden and exposed terminals or near and far terminals present serious problems here.) Then, several MAC mechanisms will be presented for the multiplexing schemes introduced in chapter 2. While SDM and FDM are typically used in a rather fixed manner, i.e., a certain space or frequency (or frequency hopping pattern) is assigned for a longer period of time; the main focus of this chapter is on TDM mechanisms. TDM can be used in a very flexible way, as tuning in to a certain frequency does not present a problem, but time can be allocated on demand and in a distributed fashion. Well-known algorithms are Aloha (in several versions), different reservation schemes, or simple polling.

Finally, the use of CDM is discussed again to show how a MAC scheme using CDM has to assign certain codes to allow the separation of different users in code space. This chapter also shows that one typically does not use a single scheme in its pure form but mixes schemes to benefit from the specific advantages. A comparison of the four basic schemes concludes the chapter.

3.1 Motivation for a specialized MAC

The main question in connection with MAC in the wireless is whether it is possible to use elaborated MAC schemes from wired networks, for example, CSMA/CD as used in the original specification of IEEE 802.3 networks (aka Ethernet).

So let us consider **carrier sense multiple access with collision detection, (CSMA/CD)** which works as follows. A sender senses the medium (a wire or coaxial cable) to see if it is free. If the medium is busy, the sender waits until it is free. If the medium is free, the sender starts transmitting data and continues to listen into the medium. If the sender detects a collision while sending, it stops at once and sends a jamming signal.

Why does this scheme fail in wireless networks? CSMA/CD is not really interested in collisions at the sender, but rather in those at the receiver. The signal should reach the receiver without collisions. But the sender is the one detecting collisions. This is not a problem using a wire, as more or less the same signal strength can be assumed all over the wire if the length of the wire stays within certain often standardized limits. If a collision occurs somewhere in the wire, everybody will notice it. It does not matter if a sender listens into the medium to detect a collision at its own location while in reality is waiting to detect a possible collision at the receiver.

The situation is different in wireless networks. As shown in chapter 2, the strength of a signal decreases proportionally to the square of the distance to the sender. Obstacles attenuate the signal even further. The sender may now apply carrier sense and detect an idle medium. The sender starts sending – but a collision happens at the receiver due to a second sender. Section 3.1.1 explains this hidden terminal problem. The same can happen to the collision detection. The sender detects no collision and assumes that the data has been transmitted without errors, but a collision might actually have destroyed the data at the receiver. Collision detection is very difficult in wireless scenarios as the transmission power in the area of the transmitting antenna is several magnitudes higher than the receiving power. So, this very common MAC scheme from wired network fails in a wireless scenario. The following sections show some more scenarios where schemes known from fixed networks fail.

3.1.1 Hidden and exposed terminals

Consider the scenario with three mobile phones as shown in Figure 3.1. The transmission range of A reaches B, but not C (the detection range does not reach C either). The transmission range of C reaches B, but not A. Finally, the transmission range of B reaches A and C, i.e., A cannot detect C and vice versa.

A starts sending to B, C does not receive this transmission. C also wants to send something to B and senses the medium. The medium appears to be free, the carrier sense fails. C also starts sending causing a collision at B. But A cannot detect this collision at B and continues with its transmission. A is **hidden** for C and vice versa.

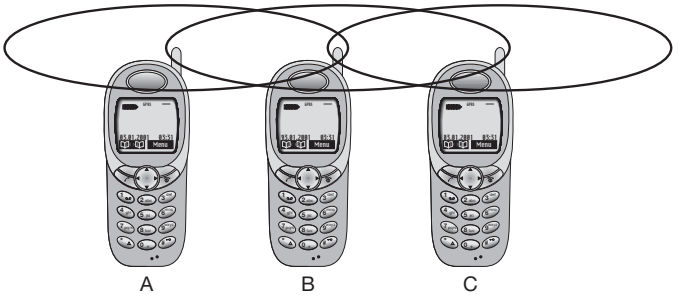

Figure 3.1
Hidden and exposed terminals

While hidden terminals may cause collisions, the next effect only causes unnecessary delay. Now consider the situation that B sends something to A and C wants to transmit data to some other mobile phone outside the interference ranges of A and B. C senses the carrier and detects that the carrier is busy (B's signal). C postpones its transmission until it detects the medium as being idle again. But as A is outside the interference range of C, waiting is not necessary. Causing a 'collision' at B does not matter because the collision is too weak to propagate to A. In this situation, C is **exposed** to B.

3.1.2 Near and far terminals

Consider the situation as shown in Figure 3.2. A and B are both sending with the same transmission power. As the signal strength decreases proportionally to the square of the distance, B's signal drowns out A's signal. As a result, C cannot receive A's transmission.

Now think of C as being an arbiter for sending rights (e.g., C acts as a base station coordinating media access). In this case, terminal B would already drown out terminal A on the physical layer. C in return would have no chance of applying a fair scheme as it would only hear B.

The **near/far effect** is a severe problem of wireless networks using CDM. All signals should arrive at the receiver with more or less the same strength. Otherwise (referring again to the party example of chapter 2) a person standing closer to somebody could always speak louder than a person further away. Even

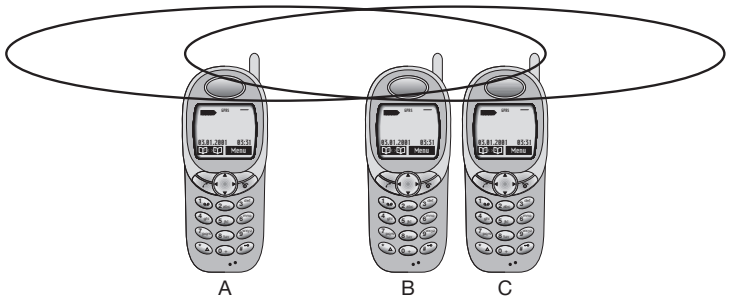

Figure 3.2
Near and far terminals

if the senders were separated by code, the closest one would simply drown out the others. Precise power control is needed to receive all senders with the same strength at a receiver. For example, the UMTS system presented in chapter 4 adapts power 1,500 times per second.

3.2 SDMA

Space Division Multiple Access (SDMA) is used for allocating a separated space to users in wireless networks. A typical application involves assigning an optimal base station to a mobile phone user. The mobile phone may receive several base stations with different quality. A MAC algorithm could now decide which base station is best, taking into account which frequencies (FDM), time slots (TDM) or code (CDM) are still available (depending on the technology). Typically, SDMA is never used in isolation but always in combination with one or more other schemes. The basis for the SDMA algorithm is formed by cells and sectorized antennas which constitute the infrastructure implementing **space division multiplexing (SDM)** (see section 2.5.1). A new application of SDMA comes up together with beam-forming antenna arrays as explained in chapter 2. Single users are separated in space by individual beams. This can improve the overall capacity of a cell (e.g., measured in bit/s/m^2 or voice calls/m^2) tremendously.

3.3 FDMA

Frequency division multiple access (FDMA) comprises all algorithms allocating frequencies to transmission channels according to the **frequency division multiplexing (FDM)** scheme as presented in section 2.5.2. Allocation can either be fixed (as for radio stations or the general planning and regulation of frequencies) or dynamic (i.e., demand driven).

Channels can be assigned to the same frequency at all times, i.e., pure FDMA, or change frequencies according to a certain pattern, i.e., FDMA combined with TDMA. The latter example is the common practice for many wireless systems to circumvent narrowband interference at certain frequencies, known as frequency hopping. Sender and receiver have to agree on a hopping pattern, otherwise the receiver could not tune to the right frequency. Hopping patterns are typically fixed, at least for a longer period. The fact that it is not possible to arbitrarily jump in the frequency space (i.e., the receiver must be able to tune to the right frequency) is one of the main differences between FDM schemes and TDM schemes.

Furthermore, FDM is often used for simultaneous access to the medium by base station and mobile station in cellular networks. Here the two partners typically establish a **duplex channel**, i.e., a channel that allows for simultaneous transmission in both directions. The two directions, mobile station to base station and vice versa are now separated using different frequencies. This scheme is then called **frequency division duplex (FDD)**. Again, both partners have to

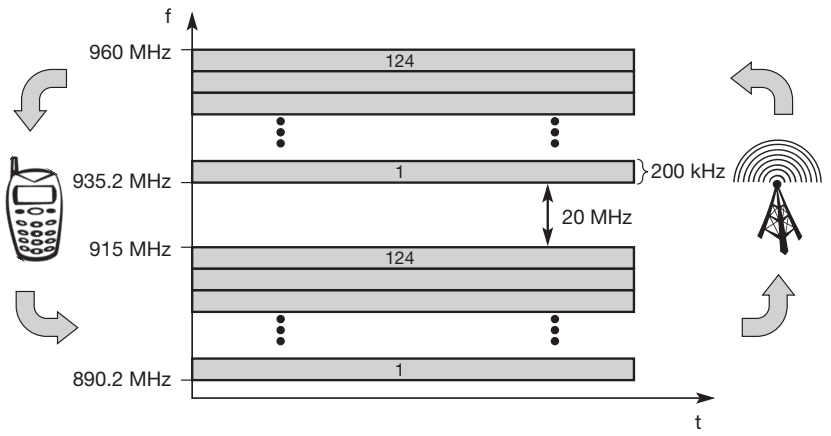

Figure 3.3
Frequency division multiplexing for multiple access and duplex

know the frequencies in advance; they cannot just listen into the medium. The two frequencies are also known as **uplink**, i.e., from mobile station to base station or from ground control to satellite, and as **downlink**, i.e., from base station to mobile station or from satellite to ground control.

As for example FDM and FDD, Figure 3.3 shows the situation in a mobile phone network based on the GSM standard for 900 MHz (see chapter 4). The basic frequency allocation scheme for GSM is fixed and regulated by national authorities. (Certain variations exist regarding the frequencies mentioned in the examples.) All uplinks use the band between 890.2 and 915 MHz, all downlinks use 935.2 to 960 MHz. According to FDMA, the base station, shown on the right side, allocates a certain frequency for up- and downlink to establish a duplex channel with a mobile phone. Up- and downlink have a fixed relation. If the uplink frequency is $f_u = 890\ \text{MHz} + n \cdot 0.2\ \text{MHz}$, the downlink frequency is $f_d = f_u + 45\ \text{MHz}$, i.e., $f_d = 935\ \text{MHz} + n \cdot 0.2\ \text{MHz}$ for a certain channel n. The base station selects the channel. Each channel (uplink and downlink) has a bandwidth of 200 kHz. This illustrates the use of FDM for multiple access (124 channels per direction are available at 900 MHz) and duplex according to a predetermined scheme. Similar FDM schemes for FDD are implemented in AMPS, IS-54, IS-95, IS-136, PACS, and UMTS (FDD mode). Chapter 4 presents some more details regarding the combination of this scheme with TDM as implemented in GSM.

3.4 TDMA

Compared to FDMA, **time division multiple access (TDMA)** offers a much more flexible scheme, which comprises all technologies that allocate certain time slots for communication, i.e., controlling **TDM**. Now tuning in to a certain frequency is not necessary, i.e., the receiver can stay at the same frequency the whole time. Using only one frequency, and thus very simple receivers and transmitters, many different algorithms exist to control medium access. As already mentioned, listening to different frequencies at the same time is quite difficult,

but listening to many channels separated in time at the same frequency is simple. Almost all MAC schemes for wired networks work according to this principle, e.g., Ethernet, Token Ring, ATM etc. (Halsall, 1996), (Stallings, 1997).

Now synchronization between sender and receiver has to be achieved in the time domain. Again this can be done by using a fixed pattern similar to FDMA techniques, i.e., allocating a certain time slot for a channel, or by using a dynamic allocation scheme. Dynamic allocation schemes require an identification for each transmission as this is the case for typical wired MAC schemes (e.g., sender address) or the transmission has to be announced beforehand. MAC addresses are quite often used as identification. This enables a receiver in a broadcast medium to recognize if it really is the intended receiver of a message. Fixed schemes do not need an identification, but are not as flexible considering varying bandwidth requirements. The following sections present several examples for fixed and dynamic schemes as used for wireless transmission. Typically, those schemes can be combined with FDMA to achieve even greater flexibility and transmission capacity.

3.4.1 Fixed TDM

The simplest algorithm for using TDM is allocating time slots for channels in a fixed pattern. This results in a fixed bandwidth and is the typical solution for wireless phone systems. MAC is quite simple, as the only crucial factor is accessing the reserved time slot at the right moment. If this synchronization is assured, each mobile station knows its turn and no interference will happen. The fixed pattern can be assigned by the base station, where competition between different mobile stations that want to access the medium is solved.

Fixed access patterns (at least fixed for some period in time) fit perfectly well for connections with a fixed bandwidth. Furthermore, these patterns guarantee a fixed delay – one can transmit, e.g., every 10 ms as this is the case for standard DECT systems. TDMA schemes with fixed access patterns are used for many digital mobile phone systems like IS-54, IS-136, GSM, DECT, PHS, and PACS.

Figure 3.4 shows how these fixed TDM patterns are used to implement multiple access and a duplex channel between a base station and mobile station. Assigning different slots for uplink and downlink using the same frequency is called **time division duplex (TDD)**. As shown in the figure, the base station uses one out of 12 slots for the downlink, whereas the mobile station uses one out of 12 different slots for the uplink. Uplink and downlink are separated in time. Up to 12 different mobile stations can use the same frequency without interference using this scheme. Each connection is allotted its own up- and downlink pair. In the example below, which is the standard case for the DECT cordless phone system, the pattern is repeated every 10 ms, i.e., each slot has a duration of 417 μs. This repetition guarantees access to the medium every 10 ms, independent of any other connections.

While the fixed access patterns, as shown for DECT, are perfectly apt for connections with a constant data rate (e.g., classical voice transmission with 32 or 64 kbit/s duplex), they are very inefficient for bursty data or asymmetric connections. If temporary bursts in data are sent from the base station to the

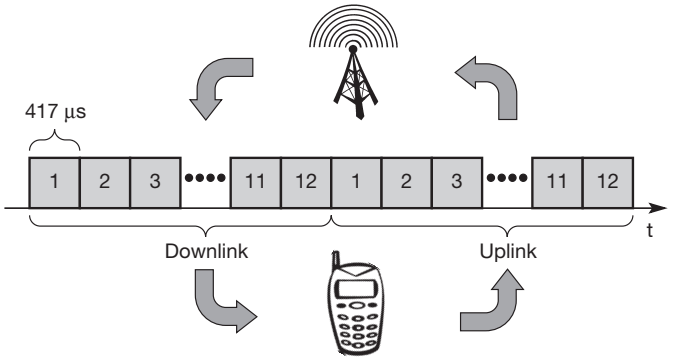

Figure 3.4 Time division multiplexing for multiple access and duplex

mobile station often or vice versa (as in the case of web browsing, where no data transmission occurs while reading a page, whereas clicking on a hyperlink triggers a data transfer from the mobile station, often to the base station, often followed by huge amounts of data returned from the web server). While DECT can at least allocate asymmetric bandwidth (see section 4.2), this general scheme still wastes a lot of bandwidth. It is too static, too inflexible for data communication. In this case, connectionless, demand-oriented TDMA schemes can be used, as the following sections show.

3.4.2 Classical Aloha

As mentioned above, TDMA comprises all mechanisms controlling medium access according to TDM. But what happens if TDM is applied without controlling access? This is exactly what the classical **Aloha** scheme does, a scheme which was invented at the University of Hawaii and was used in the ALOHANET for wireless connection of several stations. Aloha neither coordinates medium access nor does it resolve contention on the MAC layer. Instead, each station can access the medium at any time as shown in Figure 3.5. This is a random access scheme, without a central arbiter controlling access and without coordination among the stations. If two or more stations access the medium at the same time, a **collision** occurs and the transmitted data is destroyed. Resolving this problem is left to higher layers (e.g., retransmission of data).

The simple Aloha works fine for a light load and does not require any complicated access mechanisms. On the classical assumption[1] that data packet arrival follows a Poisson distribution, maximum throughput is achieved for an 18 per cent load (Abramson, 1977), (Halsall, 1996).

1 This assumption is often used for traffic in classical telephone networks but does not hold for today's Internet traffic. Internet traffic is considered as self-similar following – a so-called heavy-tail distribution. An important feature of this distribution is the existence of many values far away from the average. Self-similarity describes the independence of the observed event pattern from the duration of the observation. For example, the interarrival times of www sessions, TCP connection set-ups, IP packets or ATM cells all look similar within their respective timescale (Willinger, 1998a, b).

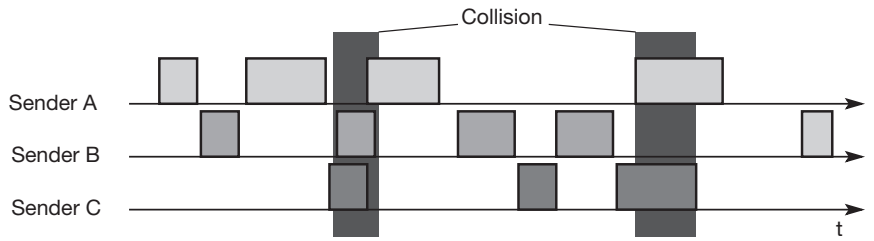

Figure 3.5 Classical Aloha multiple access

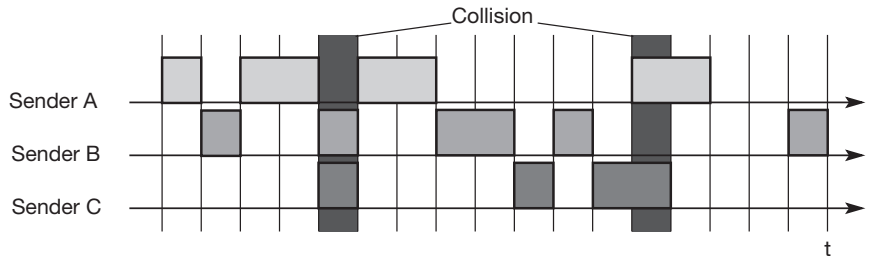

Figure 3.6 Slotted Aloha multiple access

3.4.3 Slotted Aloha

The first refinement of the classical Aloha scheme is provided by the introduction of time slots (**slotted Aloha**). In this case, all senders have to be **synchronized**, transmission can only start at the beginning of a **time slot** as shown in Figure 3.6. Still, access is not coordinated. Under the assumption stated above, the introduction of slots raises the throughput from 18 per cent to 36 per cent, i.e., slotting doubles the throughput.

As we will see in the following sections, both basic Aloha principles occur in many systems that implement distributed access to a medium. Aloha systems work perfectly well under a light load (as most schemes do), but they cannot give any hard transmission guarantees, such as maximum delay before accessing the medium, or minimum throughput. Here one needs additional mechanisms, e.g., combining fixed schemes and Aloha schemes. However, even new mobile communication systems like UMTS have to rely on slotted Aloha for medium access in certain situations (random access for initial connection set-up).

3.4.4 Carrier sense multiple access

One improvement to the basic Aloha is sensing the carrier before accessing the medium. This is what **carrier sense multiple access (CSMA)** schemes generally do (Kleinrock, 1975, Halsall, 1996). Sensing the carrier and accessing the medium only if the carrier is idle decreases the probability of a collision. But, as already mentioned in the introduction, hidden terminals cannot be detected, so, if a hidden terminal transmits at the same time as another sender, a collision might occur at the receiver. This basic scheme is still used in most wireless LANs (this will be explained in more detail in chapter 7).

Several versions of CSMA exist. In **non-persistent CSMA**, stations sense the carrier and start sending immediately if the medium is idle. If the medium is busy, the station pauses a random amount of time before sensing the medium again and repeating this pattern. In **p-persistent CSMA** systems nodes also sense the medium, but only transmit with a probability of p, with the station deferring to the next slot with the probability 1-p, i.e., access is slotted in addition. In **1-persistent CSMA systems**, all stations wishing to transmit access the medium at the same time, as soon as it becomes idle. This will cause many collisions if many stations wish to send and block each other. To create some fairness for stations waiting for a longer time, back-off algorithms can be introduced, which are sensitive to waiting time as this is done for standard Ethernet (Halsall, 1996).

CSMA with collision avoidance (**CSMA/CA**) is one of the access schemes used in wireless LANs following the standard IEEE 802.11. Here sensing the carrier is combined with a back-off scheme in case of a busy medium to achieve some fairness among competing stations. Another, very elaborate scheme is elimination yield – non-preemptive multiple access (**EY-NMPA**) used in the HIPERLAN 1 specification. Here several phases of sensing the medium and accessing the medium for contention resolution are interleaved before one "winner" can finally access the medium for data transmission. Here, priority schemes can be included to assure preference of certain stations with more important data.

3.4.5 Demand assigned multiple access

A general improvement of Aloha access systems can also be achieved by **reservation** mechanisms and combinations with some (fixed) TDM patterns. These schemes typically have a reservation period followed by a transmission period. During the reservation period, stations can reserve future slots in the transmission period. While, depending on the scheme, collisions may occur during the reservation period, the transmission period can then be accessed without collision. Alternatively, the transmission period can be split into periods with and without collision. In general, these schemes cause a higher delay under a light load (first the reservation has to take place), but allow higher throughput due to less collisions.

One basic scheme is **demand assigned multiple access (DAMA)** also called **reservation Aloha**, a scheme typical for satellite systems. DAMA, as shown in Figure 3.7 has two modes. During a contention phase following the slotted Aloha scheme, all stations can try to reserve future slots. For example, different stations on earth try to reserve access time for satellite transmission. Collisions during the reservation phase do not destroy data transmission, but only the short requests for data transmission. If successful, a time slot in the future is reserved, and no other station is allowed to transmit during this slot. Therefore, the satellite collects all successful requests (the others are destroyed) and sends back a reservation list indicating access rights for future slots. All ground stations have to obey this list. To maintain the fixed TDM pattern of reservation and transmission, the stations have to be synchronized from time to time. DAMA is an **explicit reservation** scheme. Each transmission slot has to be reserved explicitly.

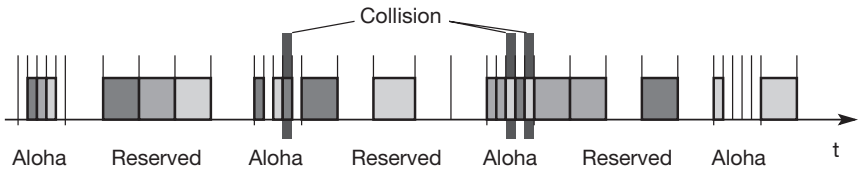

Figure 3.7
Demand assignment multiple access with explicit reservation

3.4.6 PRMA packet reservation multiple access

An example for an **implicit reservation** scheme is **packet reservation multiple access (PRMA)**. Here, slots can be reserved implicitly according to the following scheme. A certain number of slots forms a frame (Figure 3.8 shows eight slots in a frame). The frame is repeated in time (forming frames one to five in the example), i.e., a fixed TDM pattern is applied.

A base station, which could be a satellite, now broadcasts the status of each slot (as shown on the left side of the figure) to all mobile stations. All stations receiving this vector will then know which slot is occupied and which slot is currently free. In the illustration, a successful transmission of data is indicated by the station's name (A to F). In the example, the base station broadcasts the reservation status 'ACDABA-F' to all stations, here A to F. This means that slots one to six and eight are occupied, but slot seven is free in the following transmission. All stations wishing to transmit can now compete for this free slot in Aloha fashion. The already occupied slots are not touched. In the example shown, more than one station wants to access this slot, so a collision occurs. The base station returns the reservation status 'ACDABA-F', indicating that the reservation of slot seven failed (still indicated as free) and that nothing has changed for the other slots. Again, stations can compete for this slot. Additionally, station D has stopped sending in slot three and station F in slot eight. This is noticed by the base station after the second frame.

Before the third frame starts, the base station indicates that slots three and eight are now idle. Station F has succeeded in reserving slot seven as also indicated by the base station. PRMA constitutes yet another combination of fixed

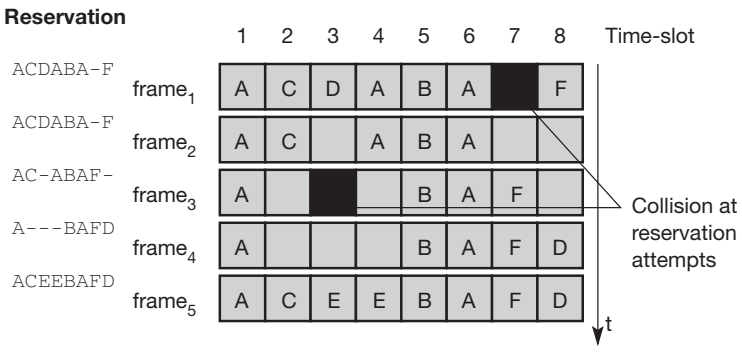

Figure 3.8
Demand assignment multiple access with implicit reservation

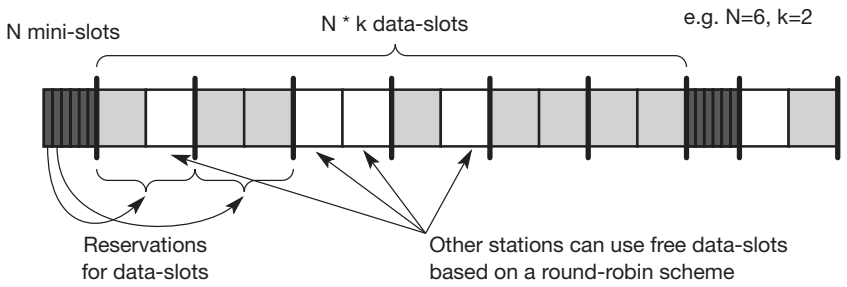

Figure 3.9
Reservation TDMA access scheme

and random TDM schemes with reservation compared to the previous schemes. As soon as a station has succeeded with a reservation, all future slots are implicitly reserved for this station. This ensures transmission with a guaranteed data rate. The slotted aloha scheme is used for idle slots only, data transmission is not destroyed by collision.

3.4.7 Reservation TDMA

An even more fixed pattern that still allows some random access is exhibited by **reservation TDMA** (see Figure 3.9). In a fixed TDM scheme N mini-slots followed by N·k data-slots form a frame that is repeated. Each station is allotted its own mini-slot and can use it to reserve up to k data-slots. This guarantees each station a certain bandwidth and a fixed delay. Other stations can now send data in unused data-slots as shown. Using these free slots can be based on a simple round-robin scheme or can be uncoordinated using an Aloha scheme. This scheme allows for the combination of, e.g., isochronous traffic with fixed bitrates and best-effort traffic without any guarantees.

3.4.8 Multiple access with collision avoidance

Let us go back to one of the initial problems: hidden terminals. How do the previous access schemes solve this? To all schemes with central base stations assigning TDM patterns, the problem of hidden terminals is unknown. If the terminal is hidden for the base station it cannot communicate anyway. But as mentioned above, more or less fixed access patterns are not as flexible as Aloha schemes. What happens when no base station exists at all? This is the case in so-called ad-hoc networks (presented in more detail in chapter 7).

Multiple access with collision avoidance (MACA) presents a simple scheme that solves the hidden terminal problem, does not need a base station, and is still a random access Aloha scheme – but with dynamic reservation. Figure 3.10 shows the same scenario as Figure 3.1 with the hidden terminals. Remember, A and C both want to send to B. A has already started the transmission, but is hidden for C, C also starts with its transmission, thereby causing a collision at B.

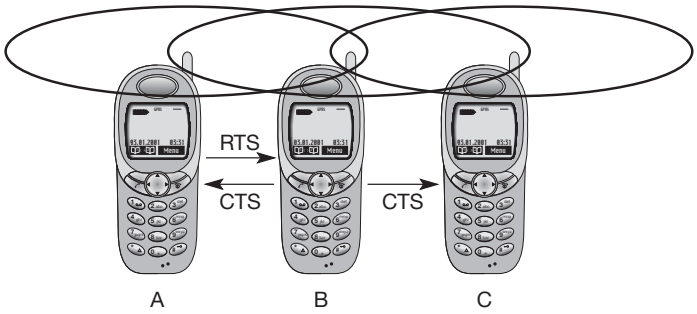

Figure 3.10
MACA can avoid hidden terminals

With MACA, A does not start its transmission at once, but sends a **request to send (RTS)** first. B receives the RTS that contains the name of sender and receiver, as well as the length of the future transmission. This RTS is not heard by C, but triggers an acknowledgement from B, called **clear to send (CTS)**. The CTS again contains the names of sender (A) and receiver (B) of the user data, and the length of the future transmission. This CTS is now heard by C and the medium for future use by A is now reserved for the duration of the transmission. After receiving a CTS, C is not allowed to send anything for the duration indicated in the CTS toward B. A collision cannot occur at B during data transmission, and the hidden terminal problem is solved – provided that the transmission conditions remain the same. (Another station could move into the transmission range of B after the transmission of CTS.)

Still, collisions can occur during the sending of an RTS. Both A and C could send an RTS that collides at B. RTS is very small compared to the data transmission, so the probability of a collision is much lower. B resolves this contention and acknowledges only one station in the CTS (if it was able to recover the RTS at all). No transmission is allowed without an appropriate CTS. This is one of the medium access schemes that is optionally used in the standard IEEE 802.11 (more details can be found in section 7.3).

Can MACA also help to solve the 'exposed terminal' problem? Remember, B wants to send data to A, C to someone else. But C is polite enough to sense the medium before transmitting, sensing a busy medium caused by the transmission from B. C defers, although C could never cause a collision at A.

With MACA, B has to transmit an RTS first (as shown in Figure 3.11) containing the name of the receiver (A) and the sender (B). C does not react to this message as it is not the receiver, but A acknowledges using a CTS which identifies B as the sender and A as the receiver of the following data transmission. C does not receive this CTS and concludes that A is outside the detection range. C can start its transmission assuming it will not cause a collision at A. The problem with exposed terminals is solved without fixed access patterns or a base station. One problem of MACA is clearly the overheads associated with the RTS and CTS transmissions – for short and time-critical data packets, this is

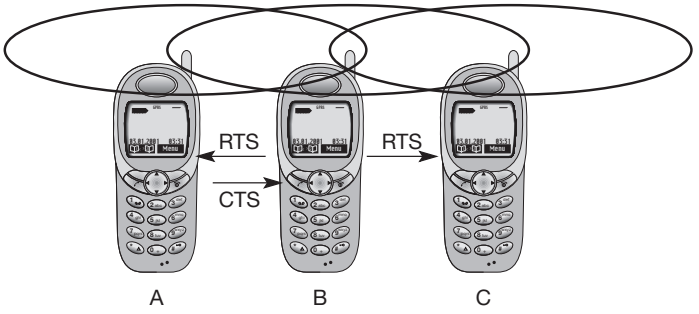

Figure 3.11
MACA can avoid exposed terminals

not negligible. MACA also assumes symmetrical transmission and reception conditions. Otherwise, a strong sender, directed antennas etc. could counteract the above scheme.

Figure 3.12 shows simplified state machines for a sender and receiver. The sender is idle until a user requests the transmission of a data packet. The sender then issues an RTS and waits for the right to send. If the receiver gets an RTS and is in an idle state, it sends back a CTS and waits for data. The sender receives the CTS and sends the data. Otherwise, the sender would send an RTS again after a time-out (e.g., the RTS could be lost or collided). After transmission of the data, the sender waits for a positive acknowledgement to return into an idle state. The receiver sends back a positive acknowledgement if the received data was correct. If not, or if the waiting time for data is too long, the receiver returns into idle state. If the sender does not receive any acknowledgement or a negative acknowledgement, it sends an RTS and again waits for the right to send. Alternatively, a receiver could indicate that it is currently busy via a separate RxBusy. Real implementations have to add more states and transitions, e.g., to limit the number of retries.

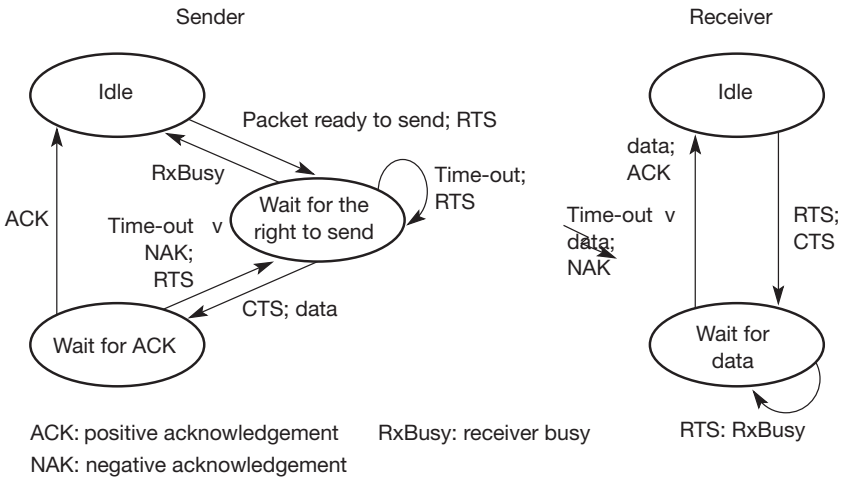

Figure 3.12
Protocol machines for multiple access with collision avoidance

3.4.9 Polling

Where one station is to be heard by all others (e.g., the base station of a mobile phone network or any other dedicated station), **polling** schemes (known from the mainframe/terminal world) can be applied. Polling is a strictly centralized scheme with one master station and several slave stations. The master can poll the slaves according to many schemes: round robin (only efficient if traffic patterns are similar over all stations), randomly, according to reservations (the classroom example with polite students) etc. The master could also establish a list of stations wishing to transmit during a contention phase. After this phase, the station polls each station on the list. Similar schemes are used, e.g., in the Bluetooth wireless LAN and as one possible access function in IEEE 802.11 systems as described in chapter 7.

3.4.10 Inhibit sense multiple access

Another combination of different schemes is represented by **inhibit sense multiple access (ISMA)**. This scheme, which is used for the packet data transmission service Cellular Digital Packet Data (CDPD) in the AMPS mobile phone system, is also known as **digital sense multiple access (DSMA)**. Here, the base station only signals a busy medium via a busy tone (called BUSY/IDLE indicator) on the downlink (see Figure 3.13). After the busy tone stops, accessing the uplink is not coordinated any further. The base station acknowledges successful transmissions, a mobile station detects a collision only via the missing positive acknowledgement. In case of collisions, additional back-off and retransmission mechanisms are implemented. (Salkintzis, 1999)

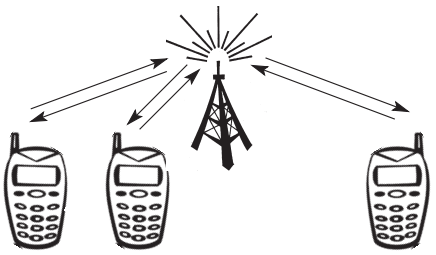

Figure 3.13 Inhibit sense multiple access using a busy tone

3.5 CDMA

Finally, codes with certain characteristics can be applied to the transmission to enable the use of **code division multiplexing (CDM)**. **Code division multiple access (CDMA)** systems use exactly these codes to separate different users in code space and to enable access to a shared medium without interference. The main problem is how to find "good" codes and how to separate the signal from noise generated by other signals and the environment.

Chapter 2 demonstrated how the codes for spreading a signal (e.g., using DSSS) could be used. The code directly controls the chipping sequence. But what is a good code for CDMA? A code for a certain user should have a good autocorre-

lation[2] and should be **orthogonal** to other codes. Orthogonal in code space has the same meaning as in standard space (i.e., the three dimensional space). Think of a system of coordinates and vectors starting at the origin, i.e., in (0, 0, 0).[3] Two vectors are called orthogonal if their inner product is 0, as is the case for the two vectors (2, 5, 0) and (0, 0, 17): (2, 5, 0)*(0, 0, 17) = 0 + 0 + 0 = 0. But also vectors like (3, –2, 4) and (–2, 3, 3) are orthogonal: (3, –2, 4)*(–2, 3, 3) = –6 – 6 + 12 = 0. By contrast, the vectors (1,2,3) and (4,2, –6) are not orthogonal (the inner product is –10), and (1, 2, 3) and (4, 2, –3) are "almost" orthogonal, with their inner product being –1 (which is "close" to zero). This description is not precise in a mathematical sense. However, it is useful to remember these simplified definitions when looking at the following examples where the original code sequences may be distorted due to noise. Orthogonality cannot be guaranteed for initially orthogonal codes.

Now let us translate this into code space and explain what we mean by a good **autocorrelation**. The Barker code (+1, –1, +1, +1, –1, +1, +1, +1, –1, –1, –1), for example, has a good autocorrelation, i.e., the inner product with itself is large, the result is 11. This code is used for ISDN and IEEE 802.11. But as soon as this Barker code is shifted 1 chip further (think of shifting the 11 chip Barker code over itself concatenated several times), the correlation drops to an absolute value of 1. It stays at this low value until the code matches itself again perfectly. This helps, for example, to synchronize a receiver with the incoming data stream. The peak in the matching process helps the receiver to reconstruct the original data precisely, even if noise distorts the original signal up to a certain level.

After this quick introduction to orthogonality and autocorrelation, the following (theoretical) example explains the basic function of CDMA before it is applied to signals:

- Two senders, A and B, want to send data. CDMA assigns the following unique and orthogonal key sequences: key A_k = 010011 for sender A, key B_K = 110101 for sender B. Sender A wants to send the bit A_d = 1, sender B sends B_d = 0. To illustrate this example, let us assume that we code a binary 0 as –1, a binary 1 as +1. We can then apply the standard addition and multiplication rules.
- Both senders spread their signal using their key as chipping sequence (the term 'spreading' here refers to the simple multiplication of the data bit with the whole chipping sequence). In reality, parts of a much longer chipping sequence are applied to single bits for spreading. Sender A then sends the signal $A_s = A_d*A_k$ = +1*(–1, +1, –1, –1, +1, +1) = (–1, +1, –1, –1, +1, +1). Sender B does the same with its data to spread the signal with the code: $B_s = B_d*B_k$ = –1*(+1, +1, –1, +1, –1, +1) = (–1, –1, +1, –1, +1, –1).

---

2 The absolute value of the inner product of a vector multiplied with itself should be large. The inner product of two vectors a and b with $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$ is defined as $a * b = \sum_{i=1}^{n} a_i b_i$.

3 82This example could also be n dimensional.

- Both signals are then transmitted at the same time using the same frequency, so, the signals superimpose in space (analog modulation is neglected in this example). Discounting interference from other senders and environmental noise from this simple example, and assuming that the signals have the same strength at the receiver, the following signal C is received at a receiver: $C = A_s + B_s = (-2, 0, 0, -2, +2, 0)$.
- The receiver now wants to receive data from sender A and, therefore, tunes in to the code of A, i.e., applies A's code for despreading: $C*A_k = (-2, 0, 0, -2, +2, 0)*(-1, +1, -1, -1, +1, +1) = 2 + 0 + 0 + 2 + 2 + 0 = 6$. As the result is much larger than 0, the receiver detects a binary 1. Tuning in to sender B, i.e., applying B's code gives $C*B_k = (-2, 0, 0, -2, +2, 0)*(+1, +1, -1, +1, -1, +1) = -2 + 0 + 0 - 2 - 2 + 0 = -6$. The result is negative, so a 0 has been detected.

This example involved several simplifications. The codes were extremely simple, but at least orthogonal. More importantly, noise was neglected. Noise would add to the transmitted signal C, the results would not be as even with –6 and +6, but would maybe be close to 0, making it harder to decide if this is still a valid 0 or 1. Additionally, both spread bits were precisely superimposed and both signals are equally strong when they reach the receiver. What would happen if, for example, B was much stronger? Assume that B's strength is five times A's strength. Then, $C' = A_s + 5*B_s = (-1, +1, -1, -1, +1, +1) + (-5, -5, +5, -5, +5, -5) = (-6, -4, +4, -6, +6, -4)$. Again, a receiver wants to receive B: $C'*B_k = -6 - 4 - 4 - 6 - 6 - 4 = -30$. It is easy to detect the binary 0 sent by B. Now the receiver wants to receive A: $C'*A_k = 6 - 4 - 4 + 6 + 6 - 4 = 6$. Clearly, the (absolute) value for the much stronger signal is higher (30 compared to 6). While –30 might still be detected as 0, this is not so easy for the 6 because compared to 30, 6 is quite close to zero and could be interpreted as noise. Remember the party example. If one person speaks in one language very loudly, it is of no more use to have another language as orthogonal code – no one can understand you, your voice will only add to the noise. Although simplified, this example shows that power control is essential for CDMA systems. This is one of the biggest problems CDMA systems face as the power has to be adjusted over one thousand times per second in some systems – this consumes a lot of energy.

The following examples summarize the behaviour of CDMA together with the DSSS spreading using orthogonal codes. The examples now use longer codes or key sequences (i.e., longer as a single bit). Code sequences in IS-95, for example, (a mobile phone system that uses CDMA) are $2^{42} - 1$ chips long, the chipping rate is 1228800 chips/s (i.e., the code repeats after 41.425 days). More details about CDMA can be found in Viterbi (1995).

Figure 3.14 shows a sender A that wants to transmit the bits 101. The key of A is shown as signal and binary key sequence A_k. In this example, the binary "0" is assigned a positive signal value, the binary "1" a negative signal value. After spreading, i.e., XORing A_d and A_k, the resulting signal is A_s.

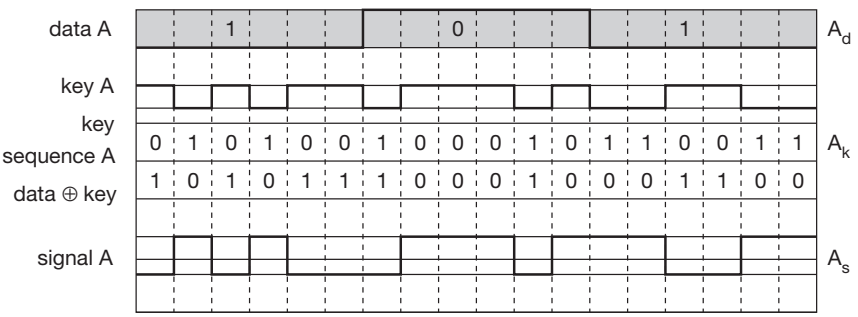

Figure 3.14
Coding and spreading of data from sender A

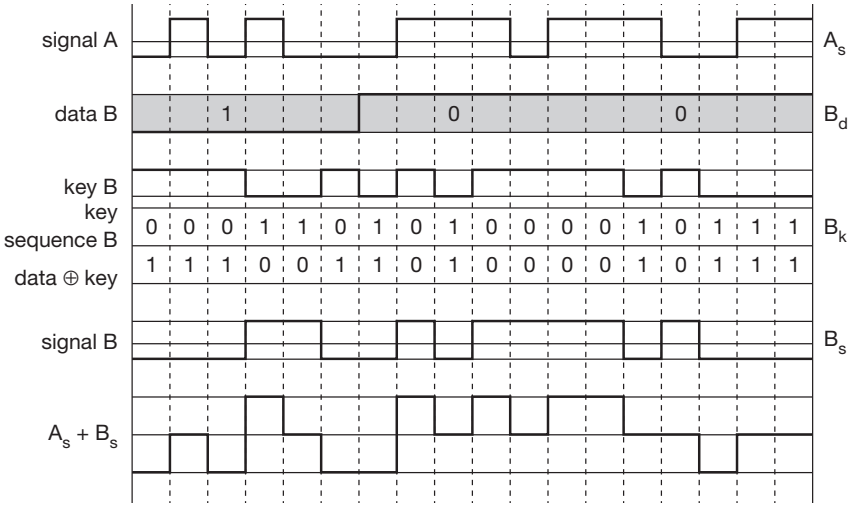

Figure 3.15
Coding and spreading of data from sender B

The same happens with data from sender B, here the bits are 100. The result of spreading with the code is the signal B_s. A_s and B_s now superimpose during transmission (again without noise and both signals having the same strength). The resulting signal is simply the sum $A_s + B_s$ as shown in Figure 3.15.

A receiver now tries to reconstruct the original data from A, A_d. Therefore the receiver applies A's key, A_k, to the received signal and feeds the result into an integrator (see section 2.7.1). The integrator adds the products (i.e., calculates the inner product), a comparator then has to decide if the result is a 0 or a 1 as shown in Figure 3.16. As we can see, although the original signal form is distorted by B's signal, the result is still quite clear.

The same happens if a receiver wants to receive B's data (see Figure 3.17). The comparator can easily detect the original data. Looking at $(A_s + B_s)*B_k$ one can also imagine what could happen if A's signal was much stronger and noise distorted the signal. The little peaks which are now caused by A's signal would

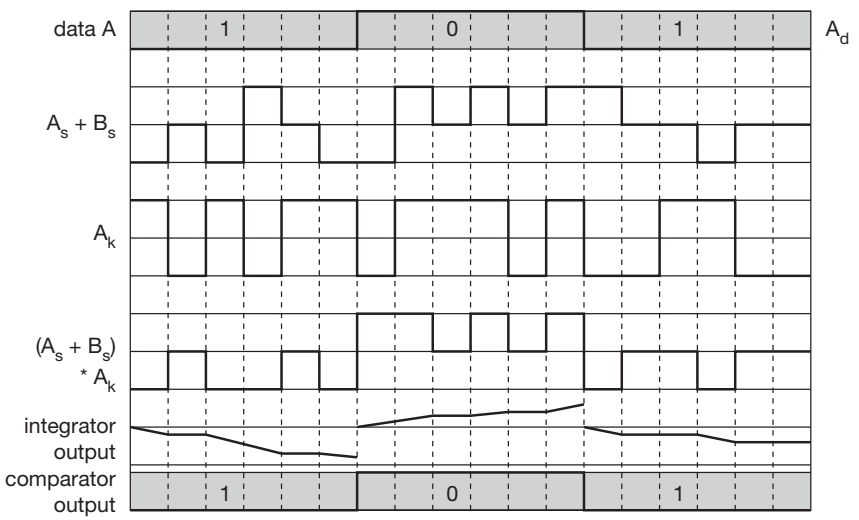

Figure 3.16 Reconstruction of A's data

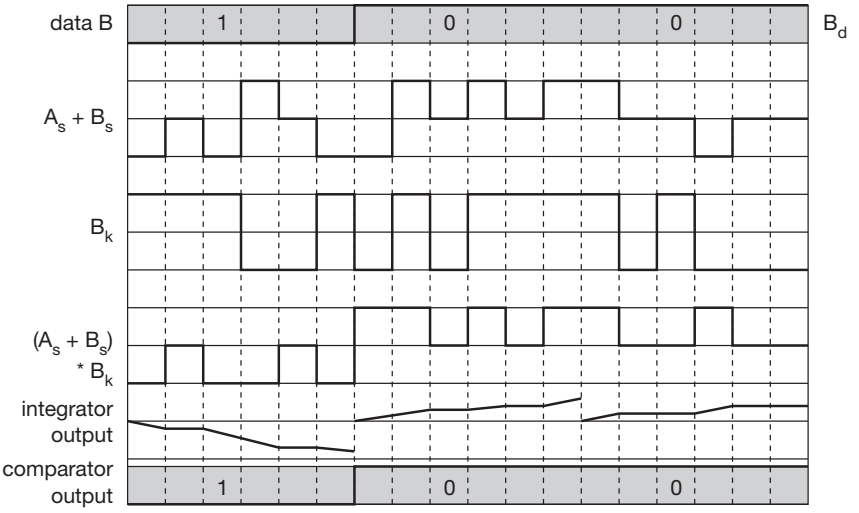

Figure 3.17 Reconstruction of B's data

be much higher, and the result of the integrator would be wrong. If A_k and B_k are perfectly orthogonal and no noise disturbs the transmission, the method works (in theory) for arbitrarily different signal strengths.

Finally, Figure 3.18 shows what happens if a receiver has the wrong key or is not synchronized with the chipping sequence of the transmitter. The integrator still presents a value after each bit period, but now it is not always possible for the comparator to decide for a 1 or a 0, as the signal rather resembles noise. Integrating over noise results in values close to zero. Even if the comparator

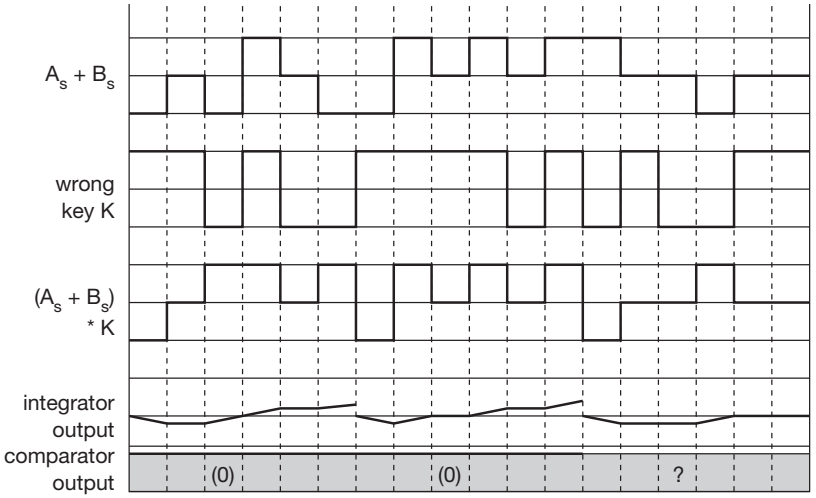

Figure 3.18
Receiving a signal with the wrong key

could detect a clear 1, this could still not reconstruct the whole bit sequence transmitted by a sender. A checksum on layer 2 would detect the erroneous packet. This illustrates CDMA's inherent protection against tapping. It is also the reason for calling the spreading code a key, as it is simultaneously used for encryption on the physical layer.

3.5.1 Spread Aloha multiple access

As shown in the previous section, using different codes with certain properties for spreading data results in a nice and powerful multiple access scheme – namely CDMA. But CDMA senders and receivers are not really simple devices. Communicating with *n* devices requires programming of the receiver to be able to decode *n* different codes (and probably sending with *n* codes, too). For mobile phone systems, a lot of the complexity needed for CDMA is integrated in the base stations. The wireless and mobile devices communicate with the base station only. However, if spontaneous, bursty traffic has to be supported between an arbitrary number of devices, the CDMA technique seems to pose too much overhead. No one wants to program many different spreading codes for, e.g., ad-hoc networks. On the other hand, Aloha was a very simple scheme, but could only provide a relatively low bandwidth due to collisions.

What happens if we combine the spreading of CDMA and the medium access of Aloha or, in other words, what if we use CDMA with only a single code, i.e., without CD? The resulting scheme is called **spread Aloha multiple access (SAMA)** and is a combination of CDMA and TDMA (Abramson, 1996).

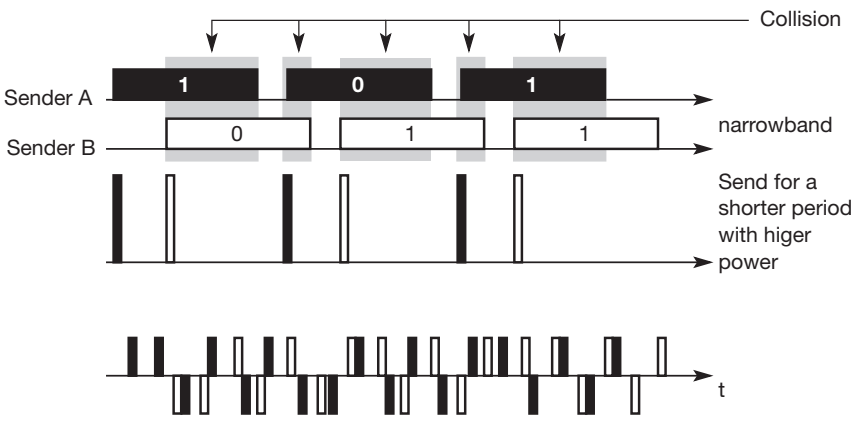

Figure 3.19 Spread Aloha multiple access

SAMA works as follows: each sender uses the same spreading code (in the example shown in Figure 3.19 this is the code 110101).[4] The standard case for Aloha access is shown in the upper part of the figure. Sender A and sender B access the medium at the same time in their narrowband spectrum, so that all three bits shown cause a collision.

The same data could also be sent with higher power for a shorter period as shown in the middle, but now spread spectrum is used to spread the shorter signals, i.e., to increase the bandwidth (spreading factor s = 6 in the example). Both signals are spread, but the chipping phase differs slightly. Separation of the two signals is still possible if one receiver is synchronized to sender A and another one to sender B. The signal of an unsynchronized sender appears as noise. The probability of a 'collision' is quite low if the number of simultaneous transmitters stays below 0.1–0.2s (Abramson, 1996). This also depends on the noise level of the environment. The main problem in using this approach is finding good chipping sequences. Clearly, the code is not orthogonal to itself – it should have a good autocorrelation but, at the same time, correlation should be low if the phase differs slightly. The maximum throughput is about 18 per cent, which is very similar to Aloha, but the approach benefits from the advantages of spread spectrum techniques: robustness against narrowband interference and simple coexistence with other systems in the same frequency bands.

4 Clearly, this is not a good code, for it is much too short. Here, coding is only done per bit, a much longer code could also stretch over many bits.

3.6 Comparison of S/T/F/CDMA

To conclude the chapter, a comparison of the four basic multiple access versions is given in Table 3.1. The table shows the MAC schemes without combination with other schemes. However, in real systems, the MAC schemes always occur in combinations. A very typical combination is constituted by SDMA/TDMA/FDMA as used in IS-54, GSM, DECT, PHS, and PACS phone systems, or the Iridium and ICO satellite systems. CDMA together with SDMA is used in the IS-95 mobile phone system and the Globalstar satellite system (see chapters 4 and 5).

Although many network providers and manufacturers have lowered their expectations regarding the performance of CDMA compared to the early 1980s (due to experiences with the IS-95 mobile phone system) CDMA is integrated into almost all third generation mobile phone systems either as W-CDMA (FOMA, UMTS) or cdma2000 (see chapter 4). CDMA can be used in combination with FDMA/TDMA access schemes to increase the capacity of a cell. In contrast to other schemes, CDMA has the advantage of a soft handover and soft capacity. Handover, explained in more detail in chapter 4, describes the switching from one cell to another, i.e., changing the base station that a mobile station is connected to. Soft handover means that a mobile station can smoothly switch cells. This is achieved by communicating with two base stations at the same time. CDMA does this using the same code and the receiver even benefits from both signals. TDMA/FDMA systems perform a hard handover, i.e., they switch base station and hopping sequences (time/frequency) precisely at the moment of handover. Handover decision is based on the signal strength, and oscillations between base stations are possible.

Soft capacity in CDMA systems describes the fact that CDMA systems can add more and more users to a cell, i.e., there is no hard limit. For TDMA/FDMA systems, a hard upper limit exists – if no more free time/frequency slots are available, the system rejects new users. If a new user is added to a CDMA cell, the noise level rises and the cell shrinks, but the user can still communicate. However, the shrinking of a cell can cause problems, as other users could now drop out of it. Cell planning is more difficult in CDMA systems compared to the more fixed TDMA/FDMA schemes (see chapter 2).

While mobile phone systems using SDMA/TDMA/FDMA or SDMA/CDMA are centralized systems – a base station controls many mobile stations – arbitrary wireless communication systems need different MAC algorithms. Most distributed systems use some version of the basic Aloha. Typically, Aloha is slotted and some reservation mechanisms are applied to guarantee access delay and bandwidth. Each of the schemes has advantages and disadvantages. Simple CSMA is very efficient at low load, MACA can overcome the problem of hidden or exposed terminals, and polling guarantees bandwidth. No single scheme combines all benefits, which is why, for example, the wireless LAN standard IEEE 802.11 combines all three schemes (see section 7.3). Polling is used to set up a time structure via a base station. A CSMA version is used to access the medium during uncoordinated periods, and additionally, MACA can be used to avoid hidden terminals or in cases where no base station exists.

Table 3.1 Comparison of SDMA, TDMA, FDMA, and CDMA mechanisms

Approach	SDMA	TDMA	FDMA	CDMA
Idea	Segment space into cells/sectors	Segment sending time into disjoint time-slots, demand driven or fixed patterns	Segment the frequency band into disjoint sub-bands	Spread the spectrum using orthogonal codes
Terminals	Only one terminal can be active in one cell/one sector	All terminals are active for short periods of time on the same frequency	Every terminal has its own frequency, uninterrupted	All terminals can be active at the same place at the same moment, uninterrupted
Signal separation	Cell structure directed antennas	Synchronization in the time domain	Filtering in the frequency domain	Code plus special receivers
Advantages	Very simple, increases capacity per km^2	Established, fully digital, very flexible	Simple, established, robust	Flexible, less planning needed, soft handover
Disadvantages	Inflexible, antennas typically fixed	Guard space needed (multi-path propagation), synchronization difficult	Inflexible, frequencies are a scarce resource	Complex receivers, needs more complicated power control for senders
Comment	Only in combination with TDMA, FDMA or CDMA useful	Standard in fixed networks, together with FDMA/SDMA used in many mobile networks	Typically combined with TDMA (frequency hopping patterns) and SDMA (frequency reuse)	Used in many 3G systems, higher complexity, lowered expectations; integrated with TDMA/FDMA

3.7 Review exercises

1. What is the main physical reason for the failure of many MAC schemes known from wired networks? What is done in wired networks to avoid this effect?
2. Recall the problem of hidden and exposed terminals. What happens in the case of such terminals if Aloha, slotted Aloha, reservation Aloha, or MACA is used?
3. How does the near/far effect influence TDMA systems? What happens in CDMA systems? What are countermeasures in TDMA systems, what about CDMA systems?
4. Who performs the MAC algorithm for SDMA? What could be possible roles of mobile stations, base stations, and planning from the network provider?
5. What is the basic prerequisite for applying FDMA? How does this factor increase complexity compared to TDMA systems? How is MAC distributed if we consider the whole frequency space as presented in chapter 1?
6. Considering duplex channels, what are alternatives for implementation in wireless networks? What about typical wired networks?
7. What are the advantages of a fixed TDM pattern compared to random, demand driven TDM? Compare the efficiency in the case of several connections with fixed data rates or in the case of varying data rates. Now explain why traditional mobile phone systems use fixed patterns, while computer networks generally use random patterns. In the future, the main data being transmitted will be computer-generated data. How will this fact change mobile phone systems?
8. Explain the term interference in the space, time, frequency, and code domain. What are countermeasures in SDMA, TDMA, FDMA, and CDMA systems?
9. Assume all stations can hear all other stations. One station wants to transmit and senses the carrier idle. Why can a collision still occur after the start of transmission?
10. What are benefits of reservation schemes? How are collisions avoided during data transmission, why is the probability of collisions lower compared to classical Aloha? What are disadvantages of reservation schemes?
11. How can MACA still fail in case of hidden/exposed terminals? Think of mobile stations and changing transmission characteristics.
12. Which of the MAC schemes can give hard guarantees related to bandwidth and access delay?
13. How are guard spaces realized between users in CDMA?
14. Redo the simple CDMA example of section 3.5, but now add random ‘noise’ to the transmitted signal (–2,0,0,–2,+2,0). Add, for example, (1,–1,0,1,0,–1). In this case, what can the receiver detect for sender A and B respectively? Now include the near/far problem. How does this complicate the situation? What would be possible countermeasures?

3.8 References

Abramson, N. (1977) 'The throughput of packet broadcasting channels,' *IEEE Transactions on Communication*, COM-25(1).

Abramson, N. (1996) 'Wideband random access for the last mile,' *IEEE Personal Communications*, 3(6).

Halsall, F. (1996) *Data communications, computer networks and open systems*. Addison-Wesley Longman.

Kleinrock, L., Tobagi, F. (1975) 'Packet switching in radio channels: part 1 – carrier sense multiple-access modes and their throughput-delay characteristics,' *IEEE Transactions on Communications*, COM-23(12).

Salkintzis, A. (1999) 'Packet data over cellular networks: The CDPD approach,' *IEEE Communications Magazine*, 37(6).

Stallings, W. (1997) *Data and computer communications*. Prentice Hall.

Viterbi, A. (1995) *CDMA: principles of spread spectrum communication*. Addison-Wesley Longman.

Willinger, W., Paxson, V. (1998a) 'Where Mathematics meets the Internet,' *Notices of the American Mathematical Society*, 45(8).

Willinger, W, Paxson, V., Taqqu, M. (1998b) 'Self-similarity and Heavy Tails: Structural Modeling of Network Traffic,' *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Adler, Taqqu (eds.), Birkhäuser-Verlag, Boston.

7 Wireless LAN

This chapter presents several wireless local area network (WLAN) technologies. This constitutes a fast-growing market introducing the flexibility of wireless access into office, home, or production environments. In contrast to the technologies described in chapters 4 through 6, WLANs are typically restricted in their diameter to buildings, a campus, single rooms etc. and are operated by individuals, not by large-scale network providers. The global goal of WLANs is to replace office cabling, to enable tetherless access to the internet and, to introduce a higher flexibility for ad-hoc communication in, e.g., group meetings. The following points illustrate some general advantages and disadvantages of WLANs compared to their wired counterparts.

Some **advantages** of WLANs are:

- **Flexibility:** Within radio coverage, nodes can communicate without further restriction. Radio waves can penetrate walls, senders and receivers can be placed anywhere (also non-visible, e.g., within devices, in walls etc.). Sometimes wiring is difficult if firewalls separate buildings (real firewalls made out of, e.g., bricks, not routers set up as a firewall). Penetration of a firewall is only permitted at certain points to prevent fire from spreading too fast.
- **Planning:** Only wireless ad-hoc networks allow for communication without previous planning, any wired network needs wiring plans. As long as devices follow the same standard, they can communicate. For wired networks, additional cabling with the right plugs and probably interworking units (such as switches) have to be provided.
- **Design:** Wireless networks allow for the design of small, independent devices which can for example be put into a pocket. Cables not only restrict users but also designers of small PDAs, notepads etc. Wireless senders and receivers can be hidden in historic buildings, i.e., current networking technology can be introduced without being visible.
- **Robustness:** Wireless networks can survive disasters, e.g., earthquakes or users pulling a plug. If the wireless devices survive, people can still communicate. Networks requiring a wired infrastructure will usually break down completely.

- **Cost:** After providing wireless access to the infrastructure via an access point for the first user, adding additional users to a wireless network will not increase the cost. This is, important for e.g., lecture halls, hotel lobbies or gate areas in airports where the numbers using the network may vary significantly. Using a fixed network, each seat in a lecture hall should have a plug for the network although many of them might not be used permanently. Constant plugging and unplugging will sooner or later destroy the plugs. Wireless connections do not wear out.

But WLANs also have several **disadvantages**:

- **Quality of service:** WLANs typically offer lower quality than their wired counterparts. The main reasons for this are the lower bandwidth due to limitations in radio transmission (e.g., only 1–10 Mbit/s user data rate instead of 100–1,000 Mbit/s), higher error rates due to interference (e.g., 10^{-4} instead of 10^{-12} for fiber optics), and higher delay/delay variation due to extensive error correction and detection mechanisms.
- **Proprietary solutions:** Due to slow standardization procedures, many companies have come up with proprietary solutions offering standardized functionality plus many enhanced features (typically a higher bit rate using a patented coding technology or special inter-access point protocols). However, these additional features only work in a homogeneous environment, i.e., when adapters from the same vendors are used for all wireless nodes. At least most components today adhere to the basic standards IEEE 802.11b or (newer) 802.11a (see section 7.3).
- **Restrictions:** All wireless products have to comply with national regulations. Several government and non-government institutions worldwide regulate the operation and restrict frequencies to minimize interference. Consequently, it takes a very long time to establish global solutions like, e.g., IMT-2000, which comprises many individual standards (see chapter 4). WLANs are limited to low-power senders and certain license-free frequency bands, which are not the same worldwide.
- **Safety and security:** Using radio waves for data transmission might interfere with other high-tech equipment in, e.g., hospitals. Senders and receivers are operated by laymen and, radiation has to be low. Special precautions have to be taken to prevent safety hazards. The open radio interface makes eavesdropping much easier in WLANs than, e.g., in the case of fiber optics. All standards must offer (automatic) encryption, privacy mechanisms, support for anonymity etc. Otherwise more and more wireless networks will be hacked into as is the case already (aka war driving: driving around looking for unsecured wireless networks; WarDriving, 2002).

Many different, and sometimes competing, design goals have to be taken into account for WLANs to ensure their commercial success:

- **Global operation:** WLAN products should sell in all countries so, national and international frequency regulations have to be considered. In contrast to the infrastructure of wireless WANs, LAN equipment may be carried from one country into another – the operation should still be legal in this case.
- **Low power:** Devices communicating via a WLAN are typically also wireless devices running on battery power. The LAN design should take this into account and implement special power-saving modes and power management functions. Wireless communication with devices plugged into a power outlet is only useful in some cases (e.g., no additional cabling should be necessary for the network in historic buildings or at trade shows). However, the future clearly lies in small handheld devices without any restricting wire.
- **License-free operation:** LAN operators do not want to apply for a special license to be able to use the product. The equipment must operate in a license-free band, such as the 2.4 GHz ISM band.
- **Robust transmission technology:** Compared to their wired counterparts, WLANs operate under difficult conditions. If they use radio transmission, many other electrical devices can interfere with them (vacuum cleaners, hairdryers, train engines etc.). WLAN transceivers cannot be adjusted for perfect transmission in a standard office or production environment. Antennas are typically omnidirectional, not directed. Senders and receivers may move.
- **Simplified spontaneous cooperation:** To be useful in practice, WLANs should not require complicated setup routines but should operate spontaneously after power-up. These LANs would not be useful for supporting, e.g., ad-hoc meetings.
- **Easy to use:** In contrast to huge and complex wireless WANs, wireless LANs are made for simple use. They should not require complex management, but rather work on a plug-and-play basis.
- **Protection of investment:** A lot of money has already been invested into wired LANs. The new WLANs should protect this investment by being interoperable with the existing networks. This means that simple bridging between the different LANs should be enough to interoperate, i.e., the wireless LANs should support the same data types and services that standard LANs support.
- **Safety and security:** Wireless LANs should be safe to operate, especially regarding low radiation if used, e.g., in hospitals. Users cannot keep safety distances to antennas. The equipment has to be safe for pacemakers, too. Users should not be able to read personal data during transmission, i.e., encryption mechanisms should be integrated. The networks should also take into account user privacy, i.e., it should not be possible to collect roaming profiles for tracking persons if they do not agree.

- **Transparency for applications:** Existing applications should continue to run over WLANs, the only difference being higher delay and lower bandwidth. The fact of wireless access and mobility should be hidden if it is not relevant, but the network should also support location aware applications, e.g., by providing location information.

The following sections first introduce basic transmission technologies used for WLANs, infra red and radio, then the two basic settings for WLANs: infrastructure-based and ad-hoc, are presented. The three main sections of this chapter present the IEEE standard for WLANs, IEEE 802.11, the European ETSI standard for a high-speed WLAN with QoS support, HiperLAN2, and finally, an industry approach toward wireless personal area networks (WPAN), i.e., WLANs at an even smaller range, called Bluetooth.

7.1 Infra red vs radio transmission

Today, two different basic transmission technologies can be used to set up WLANs. One technology is based on the transmission of infra red light (e.g., at 900 nm wavelength), the other one, which is much more popular, uses radio transmission in the GHz range (e.g., 2.4 GHz in the license-free ISM band). Both technologies can be used to set up ad-hoc connections for work groups, to connect, e.g., a desktop with a printer without a wire, or to support mobility within a small area.

Infra red technology uses diffuse light reflected at walls, furniture etc. or directed light if a line-of-sight (LOS) exists between sender and receiver. Senders can be simple light emitting diodes (LEDs) or laser diodes. Photodiodes act as receivers. Details about infra red technology, such as modulation, channel impairments etc. can be found in Wesel (1998) and Santamaría (1994).

- The main **advantages** of infra red technology are its simple and extremely cheap senders and receivers which are integrated into nearly all mobile devices available today. PDAs, laptops, notebooks, mobile phones etc. have an infra red data association (IrDA) interface. Version 1.0 of this industry standard implements data rates of up to 115 kbit/s, while IrDA 1.1 defines higher data rates of 1.152 and 4 Mbit/s. No licenses are needed for infra red technology and shielding is very simple. Electrical devices do not interfere with infra red transmission.
- **Disadvantages** of infra red transmission are its low bandwidth compared to other LAN technologies. Typically, IrDA devices are internally connected to a serial port limiting transfer rates to 115 kbit/s. Even 4 Mbit/s is not a particularly high data rate. However, their main disadvantage is that infra red is quite easily shielded. Infra red transmission cannot penetrate walls or other obstacles. Typically, for good transmission quality and high data rates a LOS, i.e., direct connection, is needed.

Almost all networks described in this book use **radio** waves for data transmission, e.g., GSM at 900, 1,800, and 1,900 MHz, DECT at 1,880 MHz etc.

- **Advantages** of radio transmission include the long-term experiences made with radio transmission for wide area networks (e.g., microwave links) and mobile cellular phones. Radio transmission can cover larger areas and can penetrate (thinner) walls, furniture, plants etc. Additional coverage is gained by reflection. Radio typically does not need a LOS if the frequencies are not too high. Furthermore, current radio-based products offer much higher transmission rates (e.g., 54 Mbit/s) than infra red (directed laser links, which offer data rates well above 100 Mbit/s. These are not considered here as it is very difficult to use them with mobile devices).
- Again, the main advantage is also a big **disadvantage** of radio transmission. Shielding is not so simple. Radio transmission can interfere with other senders, or electrical devices can destroy data transmitted via radio. Additionally, radio transmission is only permitted in certain frequency bands. Very limited ranges of license-free bands are available worldwide and those that are available are not the same in all countries. However, a lot of harmonization is going on due to market pressure.

Of the three WLAN technologies presented in this chapter, only one (IEEE 802.11) standardized infra red transmission in addition to radio transmission. The other two (HIPERLAN and Bluetooth) rely on radio. The main reason for this are the shielding problems of infra red. WLANs should, e.g., cover a whole floor of a building and not just the one room where LOSs exist. Future mobile devices may have to communicate while still in a pocket or a suitcase so cannot rely on infra red. The big advantage of radio transmission in everyday use is indeed the ability to penetrate certain materials and that a LOS is not required. Many users experience a lot of difficulties adjusting infra red ports of, e.g., mobile phones to the infra red port of their PDA. Using, e.g., Bluetooth is much simpler.

7.2 Infrastructure and ad-hoc networks

Many WLANs of today need an **infrastructure** network. Infrastructure networks not only provide access to other networks, but also include forwarding functions, medium access control etc. In these infrastructure-based wireless networks, communication typically takes place only between the wireless nodes and the access point (see Figure 7.1), but not directly between the wireless nodes.

The access point does not just control medium access, but also acts as a bridge to other wireless or wired networks. Figure 7.1 shows three access points with their three wireless networks and a wired network. Several wireless networks may form one logical wireless network, so the access points together with the fixed network in between can connect several wireless networks to form a larger network beyond actual radio coverage.

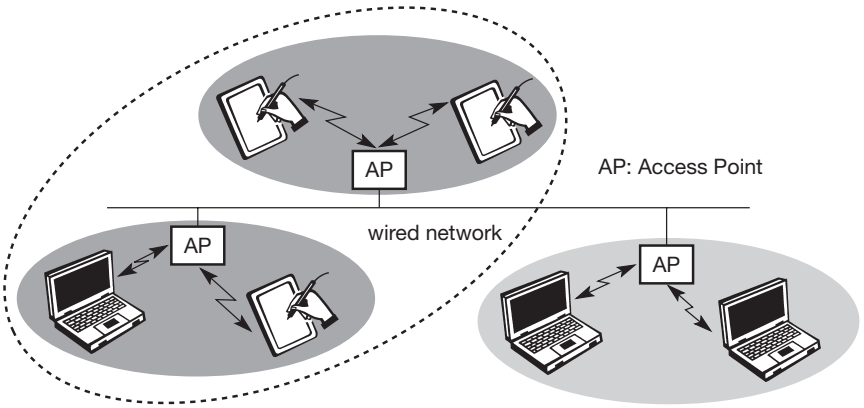

Figure 7.1
Example of three infrastructure-based wireless networks

Typically, the design of infrastructure-based wireless networks is simpler because most of the network functionality lies within the access point, whereas the wireless clients can remain quite simple. This structure is reminiscent of switched Ethernet or other star-based networks, where a central element (e.g., a switch) controls network flow. This type of network can use different access schemes with or without collision. Collisions may occur if medium access of the wireless nodes and the access point is not coordinated. However, if only the access point controls medium access, no collisions are possible. This setting may be useful for quality of service guarantees such as minimum bandwidth for certain nodes. The access point may poll the single wireless nodes to ensure the data rate.

Infrastructure-based networks lose some of the flexibility wireless networks can offer, e.g., they cannot be used for disaster relief in cases where no infrastructure is left. Typical cellular phone networks are infrastructure-based networks for a wide area (see chapter 4). Also satellite-based cellular phones have an infrastructure – the satellites (see chapter 5). Infrastructure does not necessarily imply a wired fixed network.

Ad-hoc wireless networks, however, do not need any infrastructure to work. Each node can communicate directly with other nodes, so no access point controlling medium access is necessary. Figure 7.2 shows two ad-hoc networks with three nodes each. Nodes within an ad-hoc network can only communicate if they can reach each other physically, i.e., if they are within each other's radio range or if other nodes can forward the message. Nodes from the two networks shown in Figure 7.2 cannot, therefore, communicate with each other if they are not within the same radio range.

In ad-hoc networks, the complexity of each node is higher because every node has to implement medium access mechanisms, mechanisms to handle hidden or exposed terminal problems, and perhaps priority mechanisms, to provide a certain quality of service. This type of wireless network exhibits the greatest possible flexibility as it is, for example, needed for unexpected meetings, quick replacements of infrastructure or communication scenarios far away from any infrastructure.

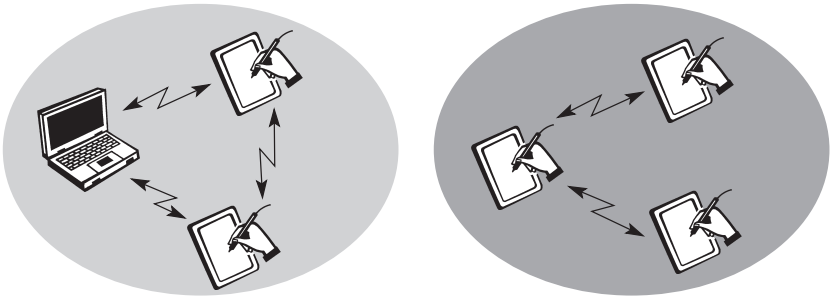

Figure 7.2
Example of two ad-hoc wireless networks

Clearly, the two basic variants of wireless networks (here especially WLANs), infrastructure-based and ad-hoc, do not always come in their pure form. There are networks that rely on access points and infrastructure for basic services (e.g., authentication of access, control of medium access for data with associated quality of service, management functions), but that also allow for direct communication between the wireless nodes.

However, ad-hoc networks might only have selected nodes with the capabilities of forwarding data. Most of the nodes have to connect to such a special node first to transmit data if the receiver is out of their range.

From the three WLANs presented, IEEE 802.11 (see section 7.3) and HiperLAN2 (see section 7.4) are typically infrastructure-based networks, which additionally support ad-hoc networking. However, many implementations only offer the basic infrastructure-based version. The third WLAN, Bluetooth (see section 7.5), is a typical wireless ad-hoc network. Bluetooth focuses precisely on spontaneous ad-hoc meetings or on the simple connection of two or more devices without requiring the setup of an infrastructure.

7.3 IEEE 802.11

The IEEE standard 802.11 (IEEE, 1999) specifies the most famous family of WLANs in which many products are available. As the standard's number indicates, this standard belongs to the group of 802.x LAN standards, e.g., 802.3 Ethernet or 802.5 Token Ring. This means that the standard specifies the physical and medium access layer adapted to the special requirements of wireless LANs, but offers the same interface as the others to higher layers to maintain interoperability.

The primary goal of the standard was the specification of a simple and robust WLAN which offers time-bounded and asynchronous services. The MAC layer should be able to operate with multiple physical layers, each of which exhibits a different medium sense and transmission characteristic. Candidates for physical layers were infra red and spread spectrum radio transmission techniques.

Additional features of the WLAN should include the support of power management to save battery power, the handling of hidden nodes, and the ability to operate worldwide. The 2.4 GHz ISM band, which is available in most countries around the world, was chosen for the original standard. Data rates envisaged for the standard were 1 Mbit/s mandatory and 2 Mbit/s optional.

The following sections will introduce the system and protocol architecture of the initial IEEE 802.11 and then discuss each layer, i.e., physical layer and medium access. After that, the complex and very important management functions of the standard are presented. Finally, this subsection presents the enhancements of the original standard for higher data rates, 802.11a (up to 54 Mbit/s at 5 GHz) and 802.11b (today the most successful with 11 Mbit/s) together with further developments for security support, harmonization, or other modulation schemes.

7.3.1 System architecture

Wireless networks can exhibit two different basic system architectures as shown in section 7.2: infrastructure-based or ad-hoc. Figure 7.3 shows the components of an infrastructure and a wireless part as specified for IEEE 802.11. Several nodes, called **stations (STA_i)**, are connected to **access points (AP)**. Stations are terminals with access mechanisms to the wireless medium and radio contact to

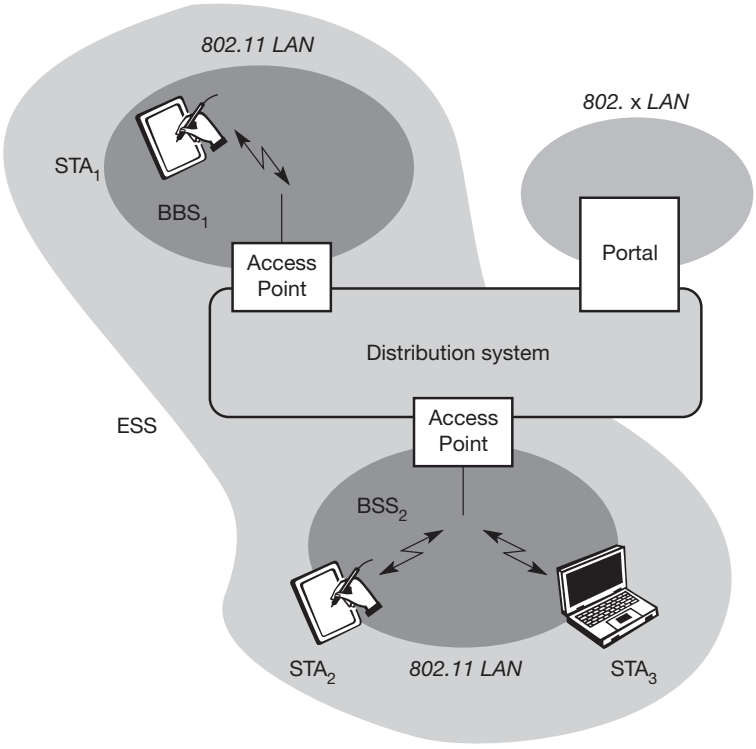

Figure 7.3 Architecture of an infrastructure-based IEEE 802.11

the AP. The stations and the AP which are within the same radio coverage form a **basic service set (BSS_i)**. The example shows two BSSs – BSS_1 and BSS_2 – which are connected via a **distribution system**. A distribution system connects several BSSs via the AP to form a single network and thereby extends the wireless coverage area. This network is now called an **extended service set (ESS)** and has its own identifier, the ESSID. The ESSID is the 'name' of a network and is used to separate different networks. Without knowing the ESSID (and assuming no hacking) it should not be possible to participate in the WLAN. The distribution system connects the wireless networks via the APs with a **portal**, which forms the interworking unit to other LANs.

The architecture of the distribution system is not specified further in IEEE 802.11. It could consist of bridged IEEE LANs, wireless links, or any other networks. However, **distribution system services** are defined in the standard (although, many products today cannot interoperate and needs the additional standard IEEE 802.11f to specify an inter access point protocol, see section 7.3.8).

Stations can select an AP and associate with it. The APs support roaming (i.e., changing access points), the distribution system handles data transfer between the different APs. APs provide synchronization within a BSS, support power management, and can control medium access to support time-bounded service. These and further functions are explained in the following sections.

In addition to infrastructure-based networks, IEEE 802.11 allows the building of ad-hoc networks between stations, thus forming one or more independent BSSs (IBSS) as shown in Figure 7.4. In this case, an IBSS comprises a group of stations using the same radio frequency. Stations STA_1, STA_2, and STA_3 are in $IBSS_1$, STA_4 and STA_5 in $IBSS_2$. This means for example that STA_3 can communicate

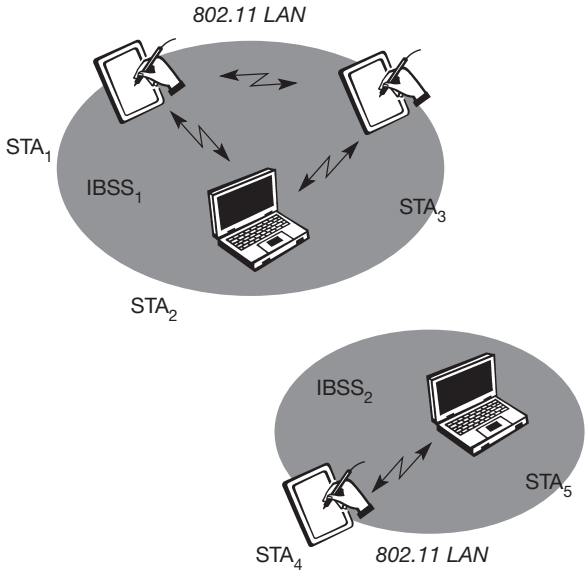

Figure 7.4 Architecture of IEEE 802.11 ad-hoc wireless LANs

directly with STA_2 but not with STA_5. Several IBSSs can either be formed via the distance between the IBSSs (see Figure 7.4) or by using different carrier frequencies (then the IBSSs could overlap physically). IEEE 802.11 does not specify any special nodes that support routing, forwarding of data or exchange of topology information as, e.g., HIPERLAN 1 (see section 7.4) or Bluetooth (see section 7.5).

7.3.2 Protocol architecture

As indicated by the standard number, IEEE 802.11 fits seamlessly into the other 802.x standards for wired LANs (see Halsall, 1996; IEEE, 1990). Figure 7.5 shows the most common scenario: an IEEE 802.11 wireless LAN connected to a switched IEEE 802.3 Ethernet via a bridge. Applications should not notice any difference apart from the lower bandwidth and perhaps higher access time from the wireless LAN. The WLAN behaves like a slow wired LAN. Consequently, the higher layers (application, TCP, IP) look the same for wireless nodes as for wired nodes. The upper part of the data link control layer, the logical link control (LLC), covers the differences of the medium access control layers needed for the different media. In many of today's networks, no explicit LLC layer is visible. Further details like Ethertype or sub-network access protocol (SNAP) and bridging technology are explained in, e.g., Perlman (1992).

The IEEE 802.11 standard only covers the physical layer **PHY** and medium access layer **MAC** like the other 802.x LANs do. The physical layer is subdivided into the **physical layer convergence protocol (PLCP)** and the **physical medium dependent** sublayer **PMD** (see Figure 7.6). The basic tasks of the MAC layer comprise medium access, fragmentation of user data, and encryption. The

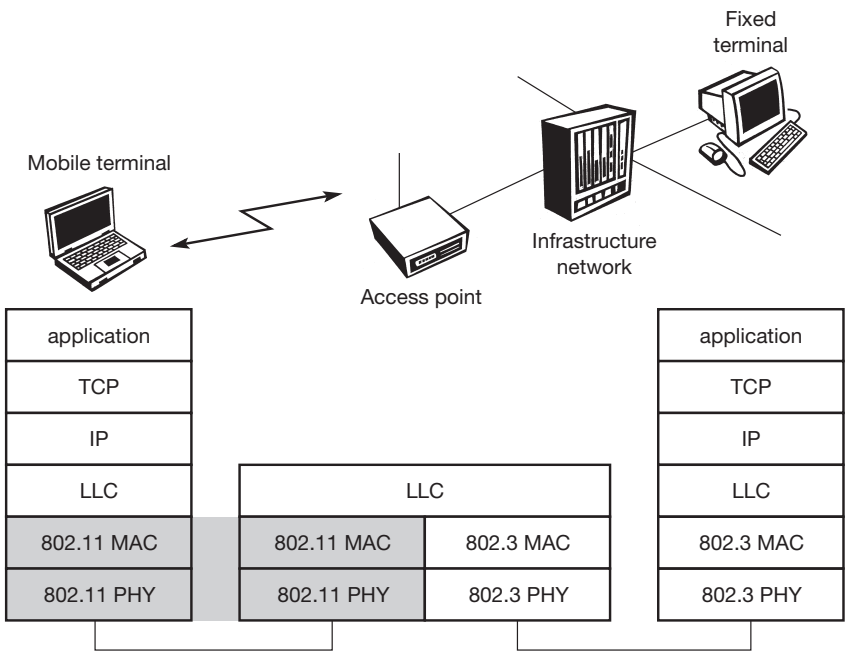

Figure 7.5 IEEE 802.11 protocol architecture and bridging

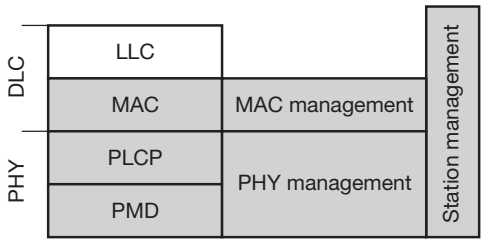

Figure 7.6
Detailed IEEE 802.11 protocol architecture and management

PLCP sublayer provides a carrier sense signal, called clear channel assessment (CCA), and provides a common PHY service access point (SAP) independent of the transmission technology. Finally, the PMD sublayer handles modulation and encoding/decoding of signals. The PHY layer (comprising PMD and PLCP) and the MAC layer will be explained in more detail in the following sections.

Apart from the protocol sublayers, the standard specifies management layers and the station management. The **MAC management** supports the association and re-association of a station to an access point and roaming between different access points. It also controls authentication mechanisms, encryption, synchronization of a station with regard to an access point, and power management to save battery power. MAC management also maintains the MAC management information base (MIB).

The main tasks of the **PHY management** include channel tuning and PHY MIB maintenance. Finally, **station management** interacts with both management layers and is responsible for additional higher layer functions (e.g., control of bridging and interaction with the distribution system in the case of an access point).

7.3.3 Physical layer

IEEE 802.11 supports three different physical layers: one layer based on infra red and two layers based on radio transmission (primarily in the ISM band at 2.4 GHz, which is available worldwide). All PHY variants include the provision of the **clear channel assessment** signal **(CCA)**. This is needed for the MAC mechanisms controlling medium access and indicates if the medium is currently idle. The transmission technology (which will be discussed later) determines exactly how this signal is obtained.

The PHY layer offers a service access point (SAP) with 1 or 2 Mbit/s transfer rate to the MAC layer (basic version of the standard). The remainder of this section presents the three versions of a PHY layer defined in the standard.

7.3.3.1 Frequency hopping spread spectrum

Frequency hopping spread spectrum (FHSS) is a spread spectrum technique which allows for the coexistence of multiple networks in the same area by separating different networks using different hopping sequences (see chapters 2 and 3). The original standard defines 79 hopping channels for North America and Europe, and 23 hopping channels for Japan (each with a bandwidth of 1 MHz

in the 2.4 GHz ISM band). The selection of a particular channel is achieved by using a pseudo-random hopping pattern. National restrictions also determine further parameters, e.g., maximum transmit power is 1 W in the US, 100 mW EIRP (equivalent isotropic radiated power) in Europe and 10 mW/MHz in Japan.

The standard specifies Gaussian shaped FSK (frequency shift keying), GFSK, as modulation for the FHSS PHY. For 1 Mbit/s a 2 level GFSK is used (i.e., 1 bit is mapped to one frequency, see chapter 2), a 4 level GFSK for 2 Mbit/s (i.e., 2 bits are mapped to one frequency). While sending and receiving at 1 Mbit/s is mandatory for all devices, operation at 2 Mbit/s is optional. This facilitated the production of low-cost devices for the lower rate only and more powerful devices for both transmission rates in the early days of 802.11.

Figure 7.7 shows a frame of the physical layer used with FHSS. The frame consists of two basic parts, the PLCP part (preamble and header) and the payload part. While the PLCP part is always transmitted at 1 Mbit/s, payload, i.e. MAC data, can use 1 or 2 Mbit/s. Additionally, MAC data is scrambled using the polynomial $s(z) = z^7 + z^4 + 1$ for DC blocking and whitening of the spectrum. The fields of the frame fulfill the following functions:

- **Synchronization:** The PLCP preamble starts with 80 bit synchronization, which is a 010101... bit pattern. This pattern is used for synchronization of potential receivers and signal detection by the CCA.
- **Start frame delimiter (SFD):** The following 16 bits indicate the start of the frame and provide frame synchronization. The SFD pattern is 0000110010111101.
- **PLCP_PDU length word (PLW):** This first field of the PLCP header indicates the length of the payload in bytes including the 32 bit CRC at the end of the payload. PLW can range between 0 and 4,095.
- **PLCP signalling field (PSF):** This 4 bit field indicates the data rate of the payload following. All bits set to zero (0000) indicates the lowest data rate of 1 Mbit/s. The granularity is 500 kbit/s, thus 2 Mbit/s is indicated by 0010 and the maximum is 8.5 Mbit/s (1111). This system obviously does not accommodate today's higher data rates.
- **Header error check (HEC)**: Finally, the PLCP header is protected by a 16 bit checksum with the standard ITU-T generator polynomial $G(x) = x^{16} + x^{12} + x^5 + 1$.

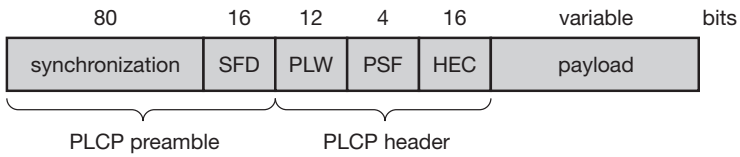

Figure 7.7 Format of an IEEE 802.11 PHY frame using FHSS

7.3.3.2 Direct sequence spread spectrum

Direct sequence spread spectrum (DSSS) is the alternative spread spectrum method separating by code and not by frequency. In the case of IEEE 802.11 DSSS, spreading is achieved using the 11-chip Barker sequence (+1, –1, +1, +1, –1, +1, +1, +1, –1, –1, –1). The key characteristics of this method are its robustness against interference and its insensitivity to multipath propagation (time delay spread). However, the implementation is more complex compared to FHSS.

IEEE 802.11 DSSS PHY also uses the 2.4 GHz ISM band and offers both 1 and 2 Mbit/s data rates. The system uses differential binary phase shift keying (DBPSK) for 1 Mbit/s transmission and differential quadrature phase shift keying (DQPSK) for 2 Mbit/s as modulation schemes. Again, the maximum transmit power is 1 W in the US, 100 mW EIRP in Europe and 10 mW/MHz in Japan. The symbol rate is 1 MHz, resulting in a chipping rate of 11 MHz. All bits transmitted by the DSSS PHY are scrambled with the polynomial $s(z) = z^7 + z^4 + 1$ for DC blocking and whitening of the spectrum. Many of today's products offering 11 Mbit/s according to 802.11b are still backward compatible to these lower data rates.

Figure 7.8 shows a frame of the physical layer using DSSS. The frame consists of two basic parts, the PLCP part (preamble and header) and the payload part. While the PLCP part is always transmitted at 1 Mbit/s, payload, i.e., MAC data, can use 1 or 2 Mbit/s. The fields of the frame have the following functions:

- **Synchronization:** The first 128 bits are not only used for synchronization, but also gain setting, energy detection (for the CCA), and frequency offset compensation. The synchronization field only consists of scrambled 1 bits.
- **Start frame delimiter (SFD):** This 16 bit field is used for synchronization at the beginning of a frame and consists of the pattern 1111001110100000.
- **Signal:** Originally, only two values have been defined for this field to indicate the data rate of the payload. The value 0x0A indicates 1 Mbit/s (and thus DBPSK), 0x14 indicates 2 Mbit/s (and thus DQPSK). Other values have been reserved for future use, i.e., higher bit rates. Coding for higher data rates is explained in sections 7.3.6 and 7.3.7.
- **Service:** This field is reserved for future use; however, 0x00 indicates an IEEE 802.11 compliant frame.
- **Length:** 16 bits are used in this case for length indication of the payload in microseconds.
- **Header error check (HEC):** Signal, service, and length fields are protected by this checksum using the ITU-T CRC-16 standard polynomial.

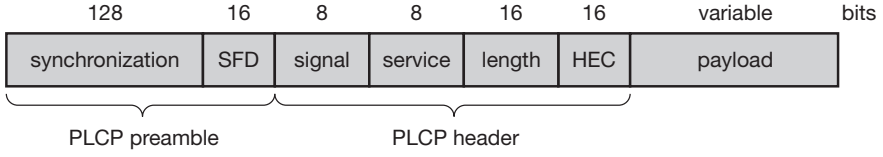

Figure 7.8 Format of an IEEE 802.11 PHY frame using DSSS

7.3.3.3 Infra red

The PHY layer, which is based on infra red (IR) transmission, uses near visible light at 850–950 nm. Infra red light is not regulated apart from safety restrictions (using lasers instead of LEDs). The standard does not require a line-of-sight between sender and receiver, but should also work with diffuse light. This allows for point-to-multipoint communication. The maximum range is about 10 m if no sunlight or heat sources interfere with the transmission. Typically, such a network will only work in buildings, e.g., classrooms, meeting rooms etc. Frequency reuse is very simple – a wall is more than enough to shield one IR based IEEE 802.11 network from another. (See also section 7.1 for a comparison between IR and radio transmission and Wesel, 1998 for more details.) Today, no products are available that offer infra red communication based on 802.11. Proprietary products offer, e.g., up to 4 Mbit/s using diffuse infra red light. Alternatively, directed infra red communication based on IrDA can be used (IrDA, 2002).

7.3.4 Medium access control layer

The MAC layer has to fulfill several tasks. First of all, it has to control medium access, but it can also offer support for roaming, authentication, and power conservation. The basic services provided by the MAC layer are the mandatory **asynchronous data service** and an optional **time-bounded service**. While 802.11 only offers the asynchronous service in ad-hoc network mode, both service types can be offered using an infrastructure-based network together with the access point coordinating medium access. The asynchronous service supports broadcast and multi-cast packets, and packet exchange is based on a 'best effort' model, i.e., no delay bounds can be given for transmission.

The following three basic access mechanisms have been defined for IEEE 802.11: the mandatory basic method based on a version of CSMA/CA, an optional method avoiding the hidden terminal problem, and finally a contention-free polling method for time-bounded service. The first two methods are also summarized as **distributed coordination function (DCF)**, the third method is called **point coordination function (PCF)**. DCF only offers asynchronous service, while PCF offers both asynchronous and time-bounded service but needs an access point to control medium access and to avoid contention. The MAC mechanisms are also called **distributed foundation wireless medium access control (DFWMAC)**.

For all access methods, several parameters for controlling the waiting time before medium access are important. Figure 7.9 shows the three different parameters that define the priorities of medium access. The values of the parameters depend on the PHY and are defined in relation to a **slot** time. Slot time is derived from the medium propagation delay, transmitter delay, and other PHY dependent parameters. Slot time is 50 μs for FHSS and 20 μs for DSSS.

The medium, as shown, can be busy or idle (which is detected by the CCA). If the medium is busy this can be due to data frames or other control frames. During a contention phase several nodes try to access the medium.

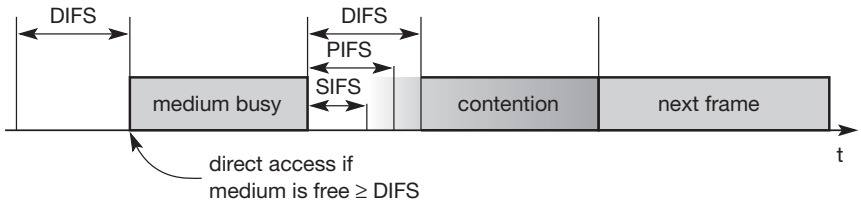

Figure 7.9
Medium access and inter-frame spacing

- **Short inter-frame spacing (SIFS):** The shortest waiting time for medium access (so the highest priority) is defined for short control messages, such as acknowledgements of data packets or polling responses. For DSSS SIFS is 10 μs and for FHSS it is 28 μs. The use of this parameter will be explained in sections 7.3.4.1 through 7.3.4.3.
- **PCF inter-frame spacing (PIFS):** A waiting time between DIFS and SIFS (and thus a medium priority) is used for a time-bounded service. An access point polling other nodes only has to wait PIFS for medium access (see section 7.3.4.3). PIFS is defined as SIFS plus one slot time.
- **DCF inter-frame spacing (DIFS):** This parameter denotes the longest waiting time and has the lowest priority for medium access. This waiting time is used for asynchronous data service within a contention period (this parameter and the basic access method are explained in section 7.3.4.1). DIFS is defined as SIFS plus two slot times.

7.3.4.1 Basic DFWMAC-DCF using CSMA/CA

The mandatory access mechanism of IEEE 802.11 is based on **carrier sense multiple access with collision avoidance** (CSMA/CA), which is a random access scheme with carrier sense and collision avoidance through random backoff. The basic CSMA/CA mechanism is shown in Figure 7.10. If the medium is idle for at least the duration of DIFS (with the help of the CCA signal of the physical layer), a node can access the medium at once. This allows for short access delay under light load. But as more and more nodes try to access the medium, additional mechanisms are needed.

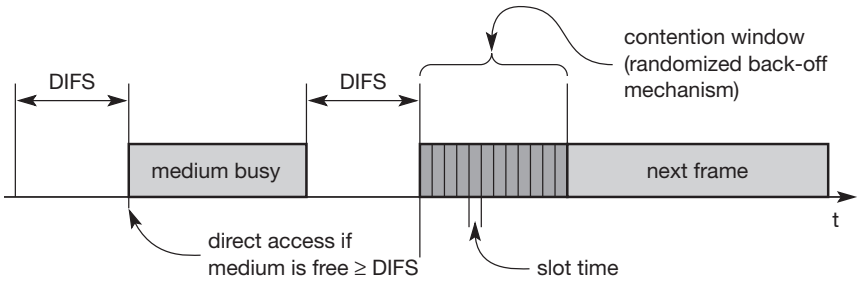

Figure 7.10
Contention window and waiting time

If the medium is busy, nodes have to wait for the duration of DIFS, entering a contention phase afterwards. Each node now chooses a **random backoff time** within a **contention window** and delays medium access for this random amount of time. The node continues to sense the medium. As soon as a node senses the channel is busy, it has lost this cycle and has to wait for the next chance, i.e., until the medium is idle again for at least DIFS. But if the randomized additional waiting time for a node is over and the medium is still idle, the node can access the medium immediately (i.e., no other node has a shorter waiting time). The additional waiting time is measured in multiples of the above-mentioned slots. This additional randomly distributed delay helps to avoid collisions – otherwise all stations would try to transmit data after waiting for the medium becoming idle again plus DIFS.

Obviously, the basic CSMA/CA mechanism is not fair. Independent of the overall time a node has already waited for transmission; each node has the same chances for transmitting data in the next cycle. To provide fairness, IEEE 802.11 adds a **backoff timer.** Again, each node selects a random waiting time within the range of the contention window. If a certain station does not get access to the medium in the first cycle, it stops its backoff timer, waits for the channel to be idle again for DIFS and starts the counter again. As soon as the counter expires, the node accesses the medium. This means that deferred stations do not choose a randomized backoff time again, but continue to count down. Stations that have waited longer have the advantage over stations that have just entered, in that they only have to wait for the remainder of their backoff timer from the previous cycle(s).

Figure 7.11 explains the basic access mechanism of IEEE 802.11 for five stations trying to send a packet at the marked points in time. Station_3 has the first request from a higher layer to send a packet (packet arrival at the MAC SAP). The station senses the medium, waits for DIFS and accesses the medium, i.e., sends the packet. Station_1, station_2, and station_5 have to wait at least until the medium is idle for DIFS again after station_3 has stopped sending. Now all three stations choose a backoff time within the contention window and start counting down their backoff timers.

Figure 7.11 shows the random backoff time of station_1 as sum of bo_e (the elapsed backoff time) and bo_r (the residual backoff time). The same is shown for station_5. Station_2 has a total backoff time of only bo_e and gets access to the medium first. No residual backoff time for station_2 is shown. The backoff timers of station_1 and station_5 stop, and the stations store their residual backoff times. While a new station has to choose its backoff time from the whole contention window, the two old stations have statistically smaller backoff values. The older values are on average lower than the new ones.

Now station_4 wants to send a packet as well, so after DIFS waiting time, three stations try to get access. It can now happen, as shown in the figure, that two stations accidentally have the same backoff time, no matter whether remaining or newly chosen. This results in a collision on the medium as shown, i.e., the trans-

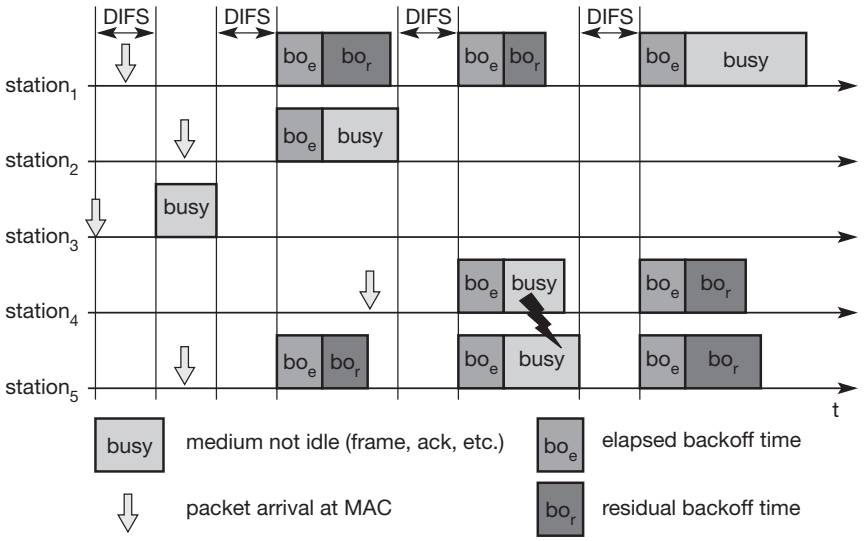

Figure 7.11
Basic DFWMAC–DCF with several competing senders

mitted frames are destroyed. Station$_1$ stores its residual backoff time again. In the last cycle shown station$_1$ finally gets access to the medium, while station$_4$ and station$_5$ have to wait. A collision triggers a retransmission with a new random selection of the backoff time. Retransmissions are not privileged.

Still, the access scheme has problems under heavy or light load. Depending on the size of the contention window (CW), the random values can either be too close together (causing too many collisions) or the values are too high (causing unnecessary delay). The system tries to adapt to the current number of stations trying to send.

The contention window starts with a size of, e.g., $CW_{min} = 7$. Each time a collision occurs, indicating a higher load on the medium, the contention window doubles up to a maximum of, e.g., $CW_{max} = 255$ (the window can take on the values 7, 15, 31, 63, 127, and 255). The larger the contention window is, the greater is the resolution power of the randomized scheme. It is less likely to choose the same random backoff time using a large CW. However, under a light load, a small CW ensures shorter access delays. This algorithm is also called **exponential backoff** and is already familiar from IEEE 802.3 CSMA/CD in a similar version.

While this process describes the complete access mechanism for broadcast frames, an additional feature is provided by the standard for unicast data transfer. Figure 7.12 shows a sender accessing the medium and sending its data. But now, the receiver answers directly with an **acknowledgement (ACK)**. The receiver accesses the medium after waiting for a duration of SIFS so no other station can access the medium in the meantime and cause a collision. The other stations have to wait for DIFS plus their backoff time. This acknowledgement ensures the correct reception (correct checksum CRC at the receiver) of a frame on the MAC layer, which is especially important in error-prone environments

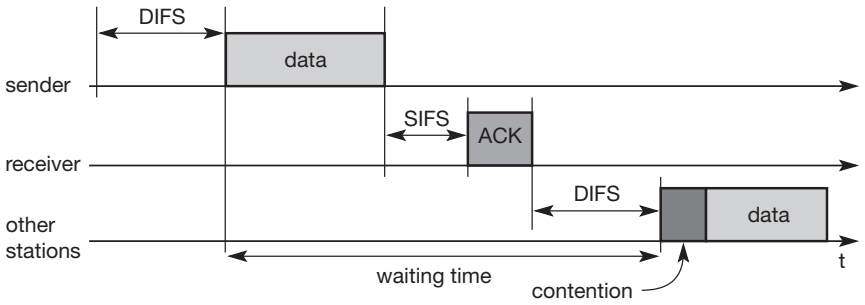

Figure 7.12
IEEE 802.11 unicast data transfer

such as wireless connections. If no ACK is returned, the sender automatically retransmits the frame. But now the sender has to wait again and compete for the access right. There are no special rules for retransmissions. The number of retransmissions is limited, and final failure is reported to the higher layer.

7.3.4.2 DFWMAC-DCF with RTS/CTS extension

Section 3.1 discussed the problem of hidden terminals, a situation that can also occur in IEEE 802.11 networks. This problem occurs if one station can receive two others, but those stations cannot receive each other. The two stations may sense the channel is idle, send a frame, and cause a collision at the receiver in the middle. To deal with this problem, the standard defines an additional mechanism using two control packets, RTS and CTS. The use of the mechanism is optional; however, every 802.11 node has to implement the functions to react properly upon reception of RTS/CTS control packets.

Figure 7.13 illustrates the use of RTS and CTS. After waiting for DIFS (plus a random backoff time if the medium was busy), the sender can issue a **request to send (RTS)** control packet. The RTS packet thus is not given any higher priority compared to other data packets. The RTS packet includes the receiver of the data transmission to come and the duration of the whole data transmission. This duration specifies the time interval necessary to transmit the whole data frame and the acknowledgement related to it. Every node receiving this RTS now has to set its **net allocation vector (NAV)** in accordance with the duration field. The NAV then specifies the earliest point at which the station can try to access the medium again.

If the receiver of the data transmission receives the RTS, it answers with a **clear to send (CTS)** message after waiting for SIFS. This CTS packet contains the duration field again and all stations receiving this packet from the receiver of the intended data transmission have to adjust their NAV. The latter set of receivers need not be the same as the first set receiving the RTS packet. Now all nodes within receiving distance around sender and receiver are informed that they have to wait more time before accessing the medium. Basically, this mechanism reserves the medium for one sender exclusively (this is why it is sometimes called a virtual reservation scheme).

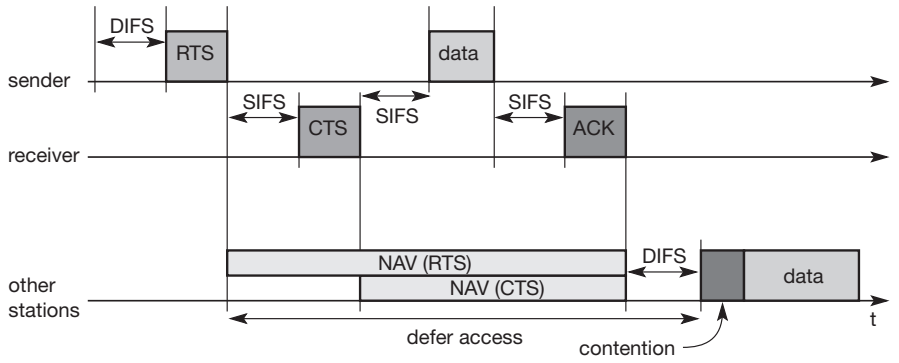

Figure 7.13
IEEE 802.11 hidden node provisions for contention-free access

Finally, the sender can send the data after SIFS. The receiver waits for SIFS after receiving the data packet and then acknowledges whether the transfer was correct. The transmission has now been completed, the NAV in each node marks the medium as free and the standard cycle can start again.

Within this scenario (i.e., using RTS and CTS to avoid the hidden terminal problem), collisions can only occur at the beginning while the RTS is sent. Two or more stations may start sending at the same time (RTS or other data packets). Using RTS/CTS can result in a non-negligible overhead causing a waste of bandwidth and higher delay. An RTS threshold can determine when to use the additional mechanism (basically at larger frame sizes) and when to disable it (short frames). Chhaya (1996) and Chhaya (1997) give an overview of the asynchronous services in 802.11 and discuss performance under different load scenarios.

Wireless LANs have bit error rates in transmission that are typically several orders of magnitude higher than, e.g., fiber optics. The probability of an erroneous frame is much higher for wireless links assuming the same frame length. One way to decrease the error probability of frames is to use shorter frames. In this case, the bit error rate is the same, but now only short frames are destroyed and, the frame error rate decreases.

However, the mechanism of fragmenting a user data packet into several smaller parts should be transparent for a user. The MAC layer should have the possibility of adjusting the transmission frame size to the current error rate on the medium. The IEEE 802.11 standard specifies a **fragmentation** mode (see Figure 7.14). Again, a sender can send an RTS control packet to reserve the medium after a waiting time of DIFS. This RTS packet now includes the duration for the transmission of the first fragment and the corresponding acknowledgement. A certain set of nodes may receive this RTS and set their NAV according to the duration field. The receiver answers with a CTS, again including the duration of the transmission up to the acknowledgement. A (possibly different) set of receivers gets this CTS message and sets the NAV.

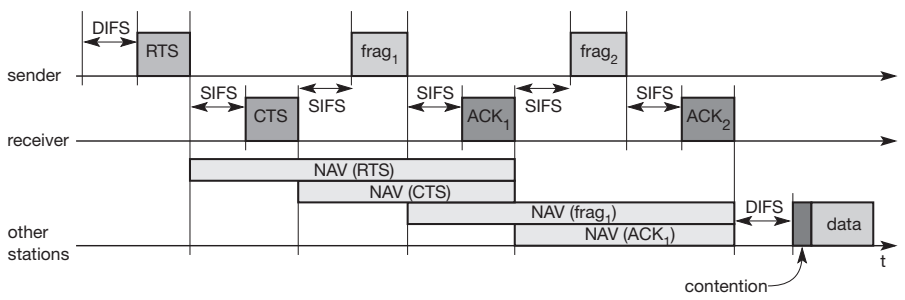

Figure 7.14 IEEE 802.11 fragmentation of user data

As shown in Figure 7.13, the sender can now send the first data frame, $frag_1$, after waiting only for SIFS. The new aspect of this fragmentation mode is that it includes another duration value in the frame $frag_1$. This duration field reserves the medium for the duration of the transmission following, comprising the second fragment and its acknowledgement. Again, several nodes may receive this reservation and adjust their NAV. If all nodes are static and transmission conditions have not changed, then the set of nodes receiving the duration field in $frag_1$ should be the same as the set that has received the initial reservation in the RTS control packet. However, due to the mobility of nodes and changes in the environment, this could also be a different set of nodes.

The receiver of $frag_1$ answers directly after SIFS with the acknowledgement packet ACK_1 including the reservation for the next transmission as shown. Again, a fourth set of nodes may receive this reservation and adjust their NAV (which again could be the same as the second set of nodes that has received the reservation in the CTS frame).

If $frag_2$ was not the last frame of this transmission, it would also include a new duration for the third consecutive transmission. (In the example shown, $frag_2$ is the last fragment of this transmission so the sender does not reserve the medium any longer.) The receiver acknowledges this second fragment, not reserving the medium again. After ACK_2, all nodes can compete for the medium again after having waited for DIFS.