4

Physical layer

Objectives of this Chapter

This chapter is devoted to the physical layer, that is, those functions and components of a sensor node that mediate between the transmission and reception of wireless waveforms and the processing of digital data in the remaining node, including the higher-layer protocol processing.

It is a commonly acknowledged truth that the properties of the transmission channel and the physical-layer shape significant parts of the protocol stack. The first goal of this chapter is therefore to provide the reader with a basic understanding of some fundamental concepts related to digital communications over wireless channels.

The second important goal is to explain how the specific constraints of wireless sensor networks (regarding, for example, energy and node costs) in turn shape the design of modulation schemes and transceivers. The reader should get an understanding on some of the fundamental trade-offs regarding transmission robustness and energy consumption and how these are affected by the power-consumption properties of transceiver components.

Chapter Outline

4.1 Introduction

The physical layer is mostly concerned with modulation and demodulation of digital data; this task is carried out by so-called **transceivers**. In sensor networks, the challenge is to find modulation schemes and transceiver architectures that are simple, low cost, but still robust enough to provide the desired service.

Protocols and Architectures for Wireless Sensor Networks H. Karl and A. Willig

The first part of this chapter explains the most important concepts regarding wireless channels and digital communications (over wireless channels); its main purpose is to provide appropriate notions and to give an insight into the tasks involved in transmission and reception over wireless channels. We discuss some simple modulation schemes as well.

In the second part, we discuss the implications of the specific requirements of wireless sensor networks, most notably the scarcity of energy, for the design of transceivers and transmission schemes.

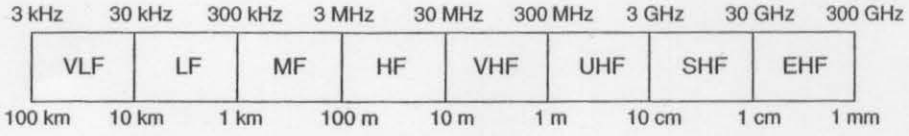
4.2 Wireless channel and communication fundamentals

This section provides the necessary background on wireless channels and digital communication over these. This is by no means an exhaustive discussion; it should just provide enough background and the most important notions to understand the energy aspects involved. Wireless channels are discussed in some more detail in references [124, 335, 620, 682, 744], some good introductory books on digital communication in general are references [772], [661], and more specific for wireless communications and systems are references [682, 848].

In wireless channels, electromagnetic waves propagate in (nearly) free space between a transmitter and a receiver. Wireless channels are therefore an **unguided medium**, meaning that signal propagation is not restricted to well-defined locations, as is the case in wired transmission with proper shielding.

4.2.1 Frequency allocation

For a practical wireless, RF-based system, the carrier frequency has to be carefully chosen. This carrier frequency determines the propagation characteristics – for example, how well are obstacles like walls penetrated – and the available capacity. Since a single frequency does not provide any capacity, for communication purposes always a finite portion of the electromagnetic spectrum, called a **frequency band**, is used. In radio-frequency (RF) communications, the range of usable radio frequencies in general starts at the Very Low Frequency (VLF) range and ends with the Extremely High Frequency (EHF) range (Figure 4.1). There is also the option of **infrared** or **optical** communications, used, for example, in the "Smart Dust" system [392]. The infrared spectrum is between wavelengths of 1 mm (corresponding to 300 GHz[1]) and 2.5 μm (120 THz), whereas the optical range ends at 780 nm (≈385 THz).



VLF = Very low frequency
LF = Low frequency
MF = Medium frequency
HF = High frequency
VHF = Very high frequency
UHF = Ultrahigh frequency
SHF = Super high frequency
EHF = Extremely high frequency

Figure 4.1 Electromagnetic spectrum – radio frequencies

[1] Assuming that the speed of light is 300,000,000 m/s.

Table 4.1 Some of the ISM bands

Frequency	Comment
13.553–13.567 MHz	
26.957–27.283 MHz	
40.66–40.70 MHz	
433–464 MHz	Europe
902–928 MHz	Only in the Americas
2.4–2.5 GHz	Used by WLAN/WPAN technologies
5.725–5.875 GHz	Used by WLAN technologies
24–24.25 GHz	

The choice of a frequency band is an important factor in system design. Except for ultrawideband technologies (see Section 2.1.4), most of today's RF-based systems work at frequencies below 6 GHz. The range of radio frequencies is subject to **regulation** to avoid unwanted interference between different users and systems. Some systems have special licenses for reserved bands; for example, in Europe, the GSM system can exclusively use the GSM 900 (880–915 MHz) and GSM 1800 (1710–1785 MHz) bands.[2] There are also licensefree bands, most notably the Industrial, Scientific, and Medical (ISM) bands, which are granted by the ITU for private and unlicensed use subject to certain restrictions regarding transmit power, power spectral density, or duty cycle. Table 4.1 lists some of the ISM frequency bands. Working in an unlicensed band means that one can just go to a shop, buy equipment, and start to transmit data without requiring any permission from the government/frequency allocation body. It is not surprising that these bands are rather popular, not only for sensor networks but also for/in other wireless technologies. For example, the 2.4-GHz ISM band is used for IEEE 802.11, Bluetooth, and IEEE 802.15.4.

Some considerations in the choice of frequency are the following:

- In the public ISM bands, any system has to live with interference created by other systems (using the same or different technologies) in the same frequency band, simply because there is no usage restriction. For example, many systems share the 2.4-GHz ISM band, including IEEE 802.11b [466, 467], Bluetooth [318, 319], and the IEEE 802.15.4 WPAN [468] – they **coexist** with each other in the same band. Therefore, all systems in these bands have to be robust against interference from other systems with which they cannot explicitly coordinate their operation. Coexistence needs to be approached both on the physical and the MAC layer [154, 359, 360, 469]. On the other hand, requesting allocation of some exclusive spectrum for a specific sensor network application from the competent regulatory organizations is a time consuming and likely futile endeavor.
- An important parameter in a transmission system is the **antenna efficiency**, which is defined as the ratio of the **radiated power** to the total input power to the antenna; the remaining power is dissipated as heat. The small form factor of wireless sensor nodes allows only small antennas. For example, radio waves at 2.4 GHz have a wave length of 12.5 cm, much longer than the intended dimensions of many sensor nodes. In general, it becomes more difficult to construct efficient antennas as the ratio of antenna dimension to wavelength decreases. As the efficiency decreases, more energy must be spent to achieve a fixed radiated power. These problems are discussed in some detail in reference [115, Chap. 8].

[2] http://www.gsmworld.com/technology/spectrum/frequencies.shtml

4.2.2 Modulation and demodulation

When digital computers communicate, they exchange **digital data**, which are essentially sequences of **symbols**, each symbol coming from a finite alphabet, the **channel alphabet**. In the process of **modulation**, (groups of) symbols from the channel alphabet are mapped to one of a finite number of **waveforms** of the same finite length; this length is called the **symbol duration**. With two different waveforms, a **binary modulation** results; if the size is $m \in \mathbb{N}, m > 2$, we talk about m-ary modulation. Some common cases for the symbol alphabet are binary data (the alphabet being $\{0, 1\}$) or bipolar data ($\{-1, 1\}$) in spread-spectrum systems.

When referring to the "speed" of data transmission/modulation, we have to distinguish between the following parameters:

Symbol rate The **symbol rate** is the inverse of the symbol duration; for binary modulation, it is also called **bit rate**.

Data rate The **data rate** is the rate in bit per second that the modulator can accept for transmission; it is thus the rate by which a user can transmit binary data. For binary modulation, bit rate and data rate are the same and often the term bit rate is (sloppily) used to denote the data rate.

For m-ary modulation, the data rate is actually given as the symbol rate times the number of bits encoded in a single waveform. For example, if we use 8-ary modulation, we can associate with each waveform one of eight possible groups of three bits and thus the bit rate is three times the symbol rate. The fundamentals of modulation and several modulation schemes are discussed in textbooks on digital communications, for example, references [78, 661, 772].

Modulation is carried out at the transmitter. The receiver ultimately wants to recover the transmitted symbols from a received waveform. The mapping from a received waveform to symbols is called **demodulation**. Because of noise, attenuation, or interference, the received waveform is a distorted version of the transmitted waveform and accordingly the receiver cannot determine the transmitted symbol with certainty. Instead, the receiver decides for the wrong symbol with some probability, called the **symbol error rate**. For digital data represented by bits, the notion of **bit error rate** (BER) is even more important: it describes the probability that a bit delivered to a higher layer is incorrect. If binary modulation is used, bit error probability and symbol error probability are the same; in case of m-ary modulation they can differ: even if a symbol is demodulated incorrectly, the delivered group of bits might be correct at some places (as long as the SNR is not too low, it is often acceptable to assume that an incorrect symbol maps to only a single incorrect bit). All upper layers are primarily interested in the bit error probability.

The most common form of modulation is the so-called **bandpass modulation**, where the information signal is modulated onto a periodic carrier wave of comparably high frequency [772, Chap. 3]. The spectrum used by bandpass modulation schemes is typically described by a **center frequency** f_c and a **bandwidth** B, and most of the signal energy can be found in the frequency range $\left[f_c - \frac{B}{2}, f_c + \frac{B}{2}\right]$.[3] The carrier is typically represented as a cosine wave, which is uniquely determined by amplitude, frequency, and phase shift.[4] Accordingly, the modulated signal $s(t)$ can, in general, be represented as:

$$s(t) = A(t) \cdot \cos(\omega(t) + \phi(t)),$$

[3] For theoretical reasons, it is not possible to have perfectly band-limited digital signals; there is always some minor signal energy leaking into neighboring frequency bands. For example, the spectrum occupied by a rectangular pulse can be described by a function similar to $\sin(x)/x$, which has nonzero values almost everywhere.

[4] There are three main advantages of bandpass modulation over digital baseband modulation like, for example, pulse modulation: it is technically comparably easy to generate sinusoids; one does not need to build huge antennas to transmit a 5-kHz data signal efficiently, and by choice of nonoverlapping bands, multiple users can transmit in parallel, which would not be possible in case of baseband modulation.

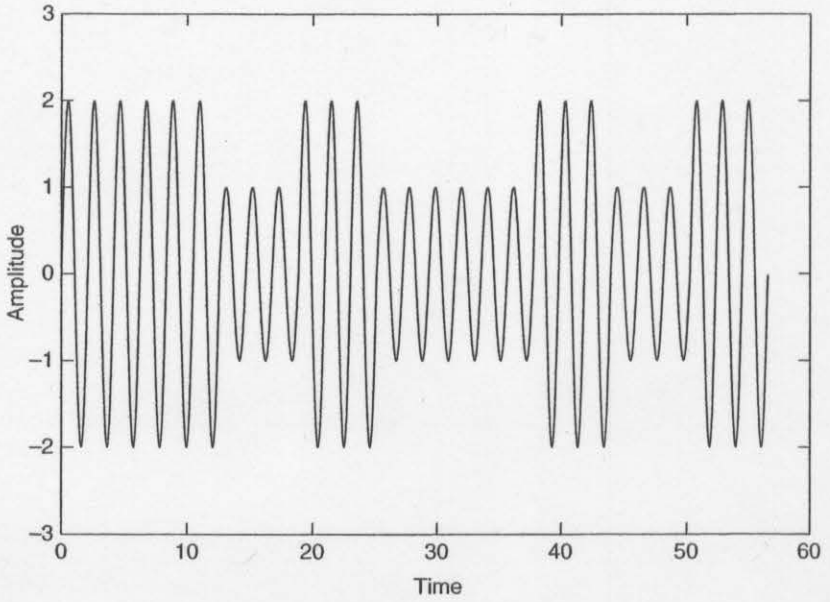


Figure 4.2 Amplitude shift keying (ASK) example

where $A(t)$ is the time-dependent amplitude, $\omega(t)$ is the time-dependent frequency, and $\phi(t)$ is the phase shift. Accordingly, there are three fundamental modulation types: Amplitude Shift Keying (ASK), Phase Shift Keying (PSK) and Frequency Shift Keying (FSK), which can be used as they are or in combination.

In ASK, the waveforms $s_i(\cdot)$ for the different symbols are chosen as:

$$s_i(t) = \sqrt{\frac{2E_i(t)}{T}} \cdot \cos\left[\omega_0 t + \phi\right],$$

where ω_0 is the center frequency, ϕ is an arbitrary constant initial phase, and $E_i(t)$ is constant over the symbol duration $[0, T]$ and assumes one of m different levels. The particular form of the amplitude $\sqrt{\frac{2E_i(t)}{T}}$ is a convention; it displays explicitly the **symbol energy** E. An example for ASK modulation is shown in Figure 4.2, where the binary data string 110100101 is modulated, using $E_0(t) = 1$ and $E_1(t) = 2$ for all t to represent logical zeros and ones. A special case of ASK modulation is a scheme with a binary channel alphabet where zeros are mapped to no signal at all, $E_0(t) = 0$, and $E_1(t) = 1$ for all t. Since it corresponds to switching off the transmitter, it is called On-Off-Keying (OOK).

In PSK, we have:

$$s_i(t) = \sqrt{\frac{2E}{T}} \cdot \cos\left[\omega_0 t + \phi_i(t)\right],$$

where ω_0 is the center frequency, E is the symbol energy, and $\phi_i(t)$ is one of m different constant values describing the phase shifts. The same binary data as in the ASK example is shown using PSK in Figure 4.3. Two popular PSK schemes are BPSK and QPSK; they are used, for example,

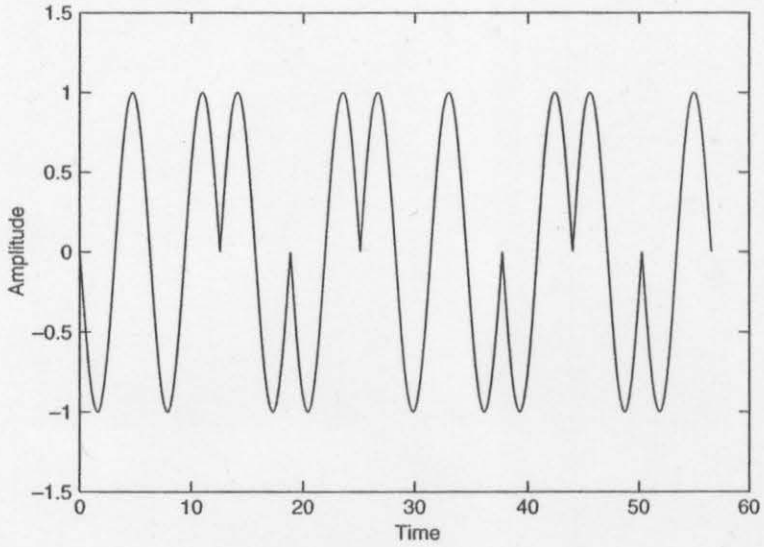


Figure 4.3 Phase shift keying (PSK) example

for the 1-Mbps and 2-Mbps modulations in IEEE 802.11 [467]. In BPSK, phase shifts of zero and π are used and in QPSK, four phase shifts of 0, $\frac{\pi}{2}$, π and $\frac{3\pi}{2}$ are used.[5]

In FSK, we have:

$$s_i(t) = \sqrt{\frac{2E}{T}} \cdot \cos\left[\omega_i(t) \cdot t + \phi\right],$$

where $\omega_i(t)$ is one of n different frequencies, E is the symbol energy, and ϕ is some constant initial phase. Figure 4.4 repeats the above example with FSK modulation.

Clearly, these basic types can be mixed. For example, Quadrature Amplitude Modulation (QAM) combines amplitude and phase modulation, using two different amplitudes and two different phases to represent two bits in one symbol.

4.2.3 Wave propagation effects and noise

Waveforms transmitted over wireless channels are subject to several physical phenomena that all *distort* the originally transmitted waveform at the receiver. This distortion introduces uncertainty at the receiver about the originally encoded and modulated data, resulting ultimately in bit errors.

Reflection, diffraction, scattering, doppler fading

The basic wave propagation phenomena [682, Chap. 3] are:

Reflection When a waveform propagating in medium A hits the boundary to another medium B and the boundary layer between them is smooth, one part of the waveform is reflected

[5] More precisely, IEEE 802.11 uses Differential Binary Phase Shift Keying (DBPSK) and Differential Quaternary Phase Shift Keying (DQPSK). In these differential versions, the information is not directly encoded in the phase of a symbol's waveform, but in the difference between phases of two subsequent symbols' waveforms.

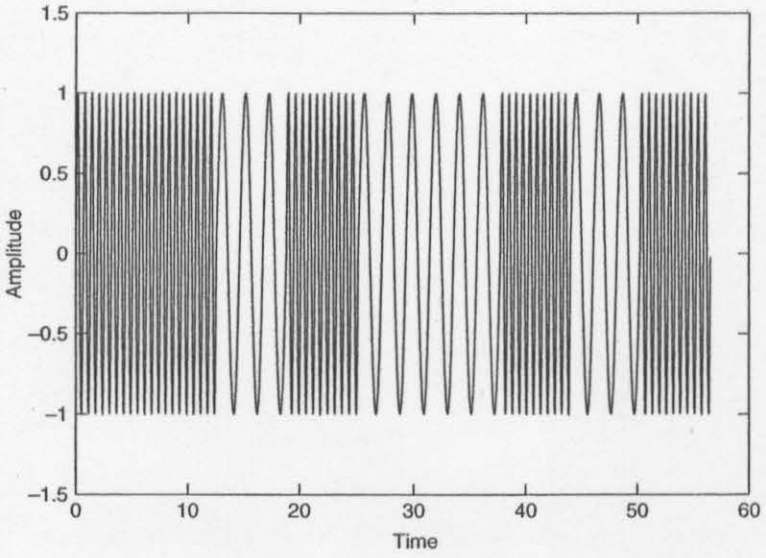


Figure 4.4 Frequency shift keying (FSK) example

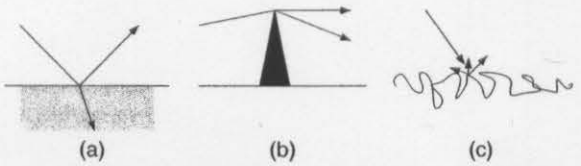


Figure 4.5 Illustration of wave propagation phenomena

back into medium A, another one is transmitted into medium B, and the rest is absorbed (Figure 4.5(a)). The amount of reflected/transmitted/absorbed energy depends on the materials and frequencies involved.

Diffraction By Huygen's principle, all points on a wavefront can be considered as sources of a new wavefront. If a waveform hits a sharp edge, it can by this token be propagated into a shadowed region (Figure 4.5(b)).

Scattering When a waveform hits a rough surface, it can be reflected multiple times and diffused into many directions (Figure 4.5(c)).

Doppler fading When a transmitter and receiver move relative to each other, the waveforms experience a shift in frequency, according to the Doppler effect. Too much of a shift can cause the receiver to sample signals at wrong frequencies.

Radio antennas radiate their signal into all directions at (nearly) the same strength, or they have a preferred direction characterized by a beam. In the first case, we have **omnidirectional antennas**,and in the second, we speak of **directed antennas**. In either case, it is likely that not only a single but multiple copies of the same signal would reach the receiver over different paths with different path lengths and attenuation (Figure 4.6), where a direct path or **Line Of Sight (LOS) path** and a reflected, or **Non line Of Sight (NLOS) path** are shown.

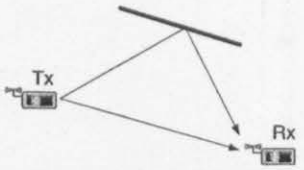


Figure 4.6 Multipath propagation

The signal at the receiver is therefore a superposition of multiple and delayed copies of the same signal. A signal actually occupies a certain spectrum, which can be represented by Fourier techniques. The different signal copies have different relative delays, which translate for each frequency component of the signal into different relative phase shifts at the receiver. Depending on the relative phase shift of the signal components, *destructive* or *constructive* interference can occur. If the channel treats all frequency components of a signal in "more or less the same way" (i.e., their amplitudes at the receiver are strongly correlated [682, Sec. 5.4]), we have **frequency-nonselective fading**, also often called **flat fading**; otherwise, we have a **frequency-selective channel**. The frequency (non-)selectivity of a channel is closely related to its **time dispersion** or **delay spread**, more exactly to the **RMS delay spread** value.[6] The **coherence bandwidth** captures, for a given propagation environment, the range of frequencies over which a channel can be considered flat; it is defined as the inverse of the RMS delay spread times a constant factor. A channel is a flat fading channel if the full signal bandwidth is smaller than the coherence bandwidth.

For wireless sensor networks with their small transmission ranges (leading to small RMS delay spread) and their comparably low symbol rates, it is reasonable to assume flat fading channels.

When transmitter and receiver move relatively to each other, the number and relative phase offset of the multiple paths changes over time and the received signal strength can fluctuate on the order of 30–40 dB within short time; this is called **fast fading** or **multipath fading**. Depending on the relative speed, the fluctuations occur at timescales of tens to hundreds of milliseconds.[7]

The importance of fading is its impact on the receiver. Since any receiver needs a minimum signal strength to have a chance for proper demodulation, a fade with its resulting drop in received signal strength is a source of errors. When the signal strength falls below this threshold because of fast fading, this is called a **deep fade**. When judging fast fading channels, specifically the rate at which the signal falls below this threshold (the **level-crossing rate**) and the duration of the deep fades are important. Qualitatively, fading channels tend to show **bursty errors**, that is, symbol errors tend to occur in clusters separated by errorfree periods.

Another source of errors (predominantly) caused by multipath propagation is InterSymbol Interference (ISI): When the transmitter transmits its symbols back-to-back, the presence of multiple paths with different delays can lead to a situation where waveforms belonging to some symbol s_t and reaching the receiver on an Line Of Sight (LOS) path overlap with delayed copies of previously sent symbols $s_{t-1}, s_{t-2}, \ldots$. The severity of ISI depends on the relationship between the symbol duration and the RMS delay spread.

[6] To characterize time dispersion of a multipath channel, the channel impulse response can be used: The transmitter emits a very short pulse and the receiver records the incoming pulses and their signal strength. The first received pulse corresponds to the shortest path and all subsequent pulses are from longer paths and likely attenuated. The time difference between the delayed pulses and the reference pulse are called **excess delays**, the **mean excess delay** is defined as the weighted average of the excess delays (using the pulse amplitudes as weights), and the **RMS delay spread** (root mean square) is the standard deviation of the weighted excess delays [682, Chap. 5].

[7] Example: For 2.4 GHz, the wavelength is 12.5 cm, and accordingly a change of 6.25 cm in the path length difference of two paths suffices to move from amplification (constructive interference) to cancellation (destructive interference) or vice versa.

Path loss and attenuation

Wireless waveforms propagating through free space are subject to a distance-dependent loss of power, called **path loss**. The received power at a distance of $d \geq d_0$ m between transmitter and receiver is described by the **Friis free-space equation** (compare reference [682, p.107], reflections are not considered):

$$\begin{aligned} P_{\text{rcvd}}(d) &= \frac{P_{\text{tx}} \cdot G_t \cdot G_r \cdot \lambda^2}{(4\pi)^2 \cdot d^2 \cdot L} \\ &= \frac{P_{\text{tx}} \cdot G_t \cdot G_r \cdot \lambda^2}{(4\pi)^2 \cdot d_0^2 \cdot L} \cdot \left(\frac{d_0}{d}\right)^2 = P_{\text{rcvd}}(d_0) \cdot \left(\frac{d_0}{d}\right)^2, \end{aligned} \tag{4.1}$$

where P_{tx} is the transmission power, G_t and G_r are the **antenna gains**[8] of transmitter and receiver, d_0 is the so-called **far-field distance**, which is a reference distance[9] depending on the antenna technology, $d \geq d_0$ is the distance between transmitter and receiver, λ is the **wavelength** and $L \geq 1$ summarizes losses through transmit/receive circuitry. Note that this equation is only valid for $d \geq d_0$. For environments other than free space, the model is slightly generalized:

$$P_{\text{rcvd}}(d) = P_{\text{rcvd}}(d_0) \cdot \left(\frac{d_0}{d}\right)^{\gamma}, \tag{4.2}$$

where γ is the **path-loss exponent**, which typically varies between 2 (free-space path loss) and 5 to 6 (shadowed areas and obstructed in-building scenarios [682, Table 4.2]). However, even values $\gamma < 2$ are possible in case of constructive interference. The path loss is defined as the ratio of the radiated power to the received power $\frac{P_{\text{tx}}}{P_{\text{rcvd}}(d)}$ and, starting from Equation 4.2, can be expressed in decibel as:

$$\text{PL}(d)[\text{dB}] = \text{PL}(d_0)[\text{dB}] + 10\gamma \log_{10}\left(\frac{d}{d_0}\right) \tag{4.3}$$

This is the so-called **log-distance** path loss model. $\text{PL}(d_0)[\text{dB}]$ is the known path loss at the reference distance.

We can draw some first conclusions from this equation. First, the received power depends on the frequency: the higher the frequency, the lower the received power. Second, the received power depends on the distance according to a power law. For example, assuming a path-loss exponent of 2, a node at a distance of $2d$ to some receiver must spent four times the energy of a node at distance d to the same receiver, to reach the same level of received power P_{rcvd}. Since, in general, the bit/symbol error rate at the receiver is a monotone function of the received power P_{rcvd}, higher frequencies or larger distances must be compensated by an appropriate increase in transmitted power to maintain a specified P_{rcvd} value. This will be elaborated further on in the following sections of this chapter.

An extension of the log-distance path-loss model takes the presence of obstacles into account. In the so-called **lognormal fading**, the deviations from the log-distance models due to obstacles

[8] Antenna gain: For directional antennas, this gives the ratio of the received power in the main direction to what would have been received from an isotropic/omnidirectional antenna (using the same transmit power).

[9] d_0 is for cellular systems with large coverage in the range of 1 km; for short range systems like WLANs, it is in the range of 1 m [682, p. 139].

are modeled as a multiplicative lognormal random variable. Equivalently, the received power can be expressed in dB as:

$$\mathrm{PL}(d)[\mathrm{dB}] = \mathrm{PL}(d_0)[\mathrm{dB}] + 10\gamma \log_{10}\left(\frac{d}{d_0}\right) + X_\sigma[\mathrm{dB}], \tag{4.4}$$

where X_σ is a zero-mean Gaussian random variable with variance σ^2, also called the **shadowing variance**.

Significant variations in the distance between transmitter and receiver or the movement beyond obstacles lead to variations of the long-term mean signal strength at the receiver. Movements and "distance hops" happen at timescales of (tens of) seconds to minutes and the variations are accordingly referred to as **slow fading**.

Besides path loss, there is often also **attenuation**. Most signals are not transmitted in a vacuum but in some media, for example, air, cables, liquids, and so on. In outdoor scenarios, there may also be fog or rain. These media types introduce additional, frequency-dependent signal attenuation. However, since attenuation obeys also a power law depending on the distance, it is only rarely modeled explicitly but accounted for in the path-loss exponent of the log-distance model.

Noise and interference

In general, **interference** refers to the presence of any unwanted signals from external (w.r.t. transmitter and receiver) sources, which obscure or mask a signal. These signals can come from other transmitters sending in the same band at the same time (**multiple access interference**) or from other devices like microwave ovens radiating in the same frequency band. In **co-channel interference**, the interference sources radiates in the same or in an overlapping frequency band as the transmitter and receiver node under consideration. In **adjacent-channel interference**, the interferer works in a neighboring band. Either the interferer leaks some signal energy into the band used by transmitter and receiver or the receiver has imperfect filters and captures signals from neighboring bands.

An important further phenomenon is **thermal noise** or simply **noise**. It is caused by thermal motions of electrons in any conducting media, for example, amplifiers and receiver/transmitter circuitry. Within the context of digital receivers, noise is typically measured by the single-sided noise Power Spectral Density (PSD)[10] N_0 given by [772, Sec. 4]:

$$N_0 = K \cdot T \left[\frac{\text{Watts}}{\text{Hertz}}\right]$$

where K is Boltzmanns constant ($\approx 1.38 \cdot 10^{-23}$ J/K) and T is the so-called system temperature in Kelvin. The thermal noise is **additive**, that is, the received signal $r(t)$ can be represented as a sum of the transmitted signal $s(t)$ (as it arrives at the receiver after path loss, attenuation, scattering, and so forth) and the noise signal $n(t)$:

$$r(t) = s(t) + n(t) \tag{4.5}$$

and furthermore this noise is **Gaussian**, that is, $n(t)$ has a Gaussian/normal distribution with zero mean and finite variance σ^2 for all t. A very important property of Gaussian noise is that its PSD can be assumed constant (with value $N_0/2$ over all frequencies of practical interest). A process with constant PSD is also called **white noise**. Hence, thermal noise is also often referred to as Additive White Gaussian Noise (AWGN).

[10] Technically, the PSD of a wide-sense-stationary random process $n(t)$ is the Fourier transform of the process's autocorrelation function; intuitively, the PSD describes the distribution of a signal's power in the frequency domain.

Symbols and bit errors

The symbol/bit error probability depends on the actual modulation scheme and on the ratio of the power of the received signal (P_{rcvd}) to the noise and interference power. When only AWGN is considered, this ratio is called Signal-to-Noise Ratio (SNR) and is given in decibel as:

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{rcvd}}}{N_0} \right)$$

where N_0 is the noise power and P_{rcvd} is the average received signal power. When other sources of interference are considered, too, often the Signal to Interference and Noise Ratio (SINR) is important:

$$\text{SINR} = 10 \log_{10} \left(\frac{P_{\text{rcvd}}}{N_0 + \sum_{i=1}^{k} I_i} \right)$$

where N_0 is the noise power and I_i is the power received from the i-th interferer.

The SINR describes the power that arrives at the receiver and is thus related to the symbols sent over the channel. In the end, the symbols are not relevant; the data bits are. To correctly demodulate and decode an arriving bit, the energy per such a bit E_b in relation to the noise energy N_0 is relevant. This ratio E_b/N_0 has a close relationship to the SNR (or SINR, when interference is treated as noise) [772, Sec. 3.7]:

$$\frac{E_b}{N_0} = \text{SNR} \cdot \frac{1}{R} = \frac{P_{\text{rcvd}}}{N_0} \cdot \frac{1}{R} \tag{4.6}$$

where R is the bit rate. It will be useful later on in this chapter to look also at the bandwidth W occupied by the modulated signal and to use the **bandwidth efficiency** $\eta_{\text{BW}} = \frac{R}{W}$ (in bit/s/Hz) as a measure of a modulation scheme's efficiency. This can be used to rewrite Equation 4.6 as:

$$\frac{E_b}{N_0} = \frac{P_{\text{rcvd}}}{N_0} \cdot \frac{1}{\eta_{\text{BW}} \cdot W}. \tag{4.7}$$

An important distinction not directly concerning modulation but concerning the receiver is the one between **coherent detection** and **noncoherent detection**. In coherent detection, the receiver has perfect phase and frequency information, for example, learned from preambles or synchronization sequences (see also Section 4.2.6). In general, coherent receivers are much more complex than noncoherent ones, but need lower signal-to-noise ratios to achieve a given target Bit-Error Rate (BER).

If we prescribe a desired maximum BER, we can, for many modulation schemes, determine some minimum SNR needed to achieve this BER on an AWGN channel. To illustrate this, we show in Figure 4.7 the BER versus the ratio E_b/N_0 given in decibel for coherently detected binary PSK and binary FSK. The qualitative behavior of such BER versus E_b/N_0 is the same for all popular modulation types. For example, with BPSK, the E_b/N_0 ratio must be larger than 4 dB to reach a BER of at least 10^{-3}. The noise power is fixed, so we have to tune the received power P_{rcvd} to achieve the desired SNR. For given antennas, this can only be achieved by increasing the radiated power at the transmitter P_{tx}; compare Equation 4.1. An alternative is clearly to use better modulation schemes.

The choice of modulation schemes for wireless sensor networks is discussed in Section 4.3.2.

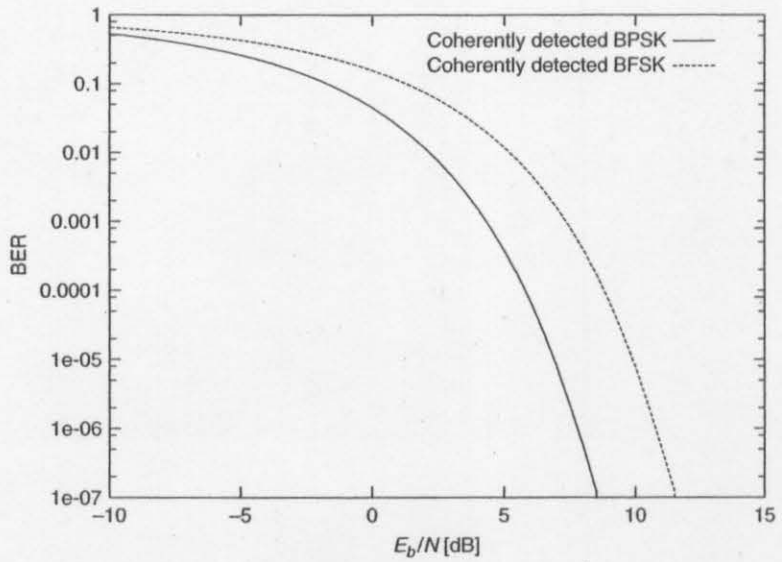


Figure 4.7 Bit error rate for coherently detected binary PSK and FSK

4.2.4 Channel models

For investigation of modulation or error control schemes, models for wireless channels are needed [36]. Because of the apparent complexity of real wireless channels, mostly *stochastic* models are used, which replace complex and tedious modeling of propagation environments by random variables. At the lowest level, such models work on the level of waveforms, describing the received signal. "Higher", more abstract models describe the statistics of symbol or bit errors or even of packet errors. These models are more amenable for investigation of network protocols, where often thousands or millions of packets are transmitted.

Signal models

We have already seen one waveform model, the AWGN model, having a constant SNR. As a reminder, this model expresses the received signal $r(t)$ as:

$$r(t) = s(t) + n(t),$$

where $s(t)$ is the transmitted signal and $n(t)$ is white Gaussian noise. One important property of this model is that the SNR is constant throughout. The simplicity of this model eases theoretical analysis; however, it is not appropriate to model time-varying channels like fading channels.

There are other popular models, specifically for frequency-nonselective fading channels [80]. These models assume that the SNR is a random variable, fluctuating from symbol to symbol or from block to block [79]. In the **Rayleigh fading** model, it is assumed that there is no LOS path. Instead, a large number of signal copies with stochastically independent signal amplitudes of the same mean value overlap at the receiver. By virtue of the central limit theorem, it can be shown that the amplitude of the resulting signal has a Rayleigh distribution, whereas the phase is uniformly distributed in $[0, 2\pi]$. A second popular model is the **Rice fading** model, which makes the same

assumptions as the Rayleigh fading model, but additionally a strong LOS component is present. Such a fading channel together with AWGN can be represented as:

$$r(t) = R \cdot e^{i\theta} \cdot s(t) + n(t)$$

where again $n(t)$ is white Gaussian noise and $R \cdot e^{i\theta}$ is a Gaussian random variable such that R has a Rice or Rayleigh probability density function.

Digital models

In the AWGN channel, each transmitted symbol is erroneous with a certain fixed error probability, and errors of subsequent symbols are independent. If these two conditions hold true and if in addition the error probability does not depend on the symbol value, we have a Binary Symmetric Channel (BSC) [180].

There have been several efforts to find good stochastic models for (Rayleigh) fading channels on the bit/symbol level. These models try to capture the tendency of fading channels to have bursty errors. Often, such channels are modeled as **Markov chains** with the states of the chain corresponding to different channel "quality levels". For example, the popular two-state **Gilbert–Elliot model** [231, 290] describes the alternation between deep fades and good periods in a fading channel. WANG and MOAYERI [858] discuss how the parameters of an N-state Markov chain describing the received signal level can be derived under Rayleigh fading assumptions from simple physical parameters like wavelength, relative speed of the nodes, and others. A more general class of models, which has also often been used, are Hidden Markov Models (HMMs); see, for instance, reference [241, 834].

WSN-specific channel models

One design constraint of wireless sensor networks is the intention to use small transmission power (and consequently the radiated power) – on the order of 1 dBm [855] – with the hope to save energy by leveraging multihop communication. The choice of a small transmit power has several consequences for the channel characteristics:

- By the Friis equation (Equation 4.1), a small transmit power implies a small range.
- Having a small transmission range means that the rms delay spread will be in the range of nanoseconds [682, Table 5.1], which is small compared to symbol durations in the order of milli- or microseconds. Since in addition the data rates are moderate, it is reasonable to expect frequency nonselective fading channels with noise [762] and a low-to-negligible degree of ISI. Accordingly, no special provisions against ISI like equalizers are needed.

SOHRABI et al. [779] present measurements of the near-ground propagation conditions for a 200-MHz frequency band between 800 MHz and 1000 MHz in various environments. These measurements comprise the path-loss exponents γ, shadowing variance σ^2, the reference path loss $\mathrm{PL}(d_0)$[dB] at $d_0 = 1$ m and the coherence bandwidth. The measurement sites under consideration include parking lots, hallways, engineering buildings and plant fences, covering distances between 1 and 30 m. Mobility was not considered. The average path-loss exponents (the average is formed over the range of frequencies), the average shadowing variance, and the ranges of the reference path loss $\mathrm{PL}(d_0)$[dB] are quoted in Table 4.2. It is interesting to note that the average path-loss exponents can range from $\gamma = 1.9$ up to $\gamma = 5$. It is also interesting to note that already at a distance of 1 m the signal has lost between 30 and 50 dB. The coherence bandwidth depends strongly on the environment as well as on the distance; with increasing distance, the coherence

Table 4.2 Average path-loss exponents, shadowing variance, and range of path loss at reference distances for near-ground measurements in 800–1000 MHz [779]

Location	Average of γ	Average of σ^2 [dB]	Range of PL(1m)[dB]
Engineering building	1.9	5.7	[−50.5, −39.0]
Apartment hallway	2.0	8.0	[−38.2, −35.0]
Parking structure	3.0	7.9	[−36.0, −32.7]
One-sided corridor	1.9	8.0	[−44.2, −33.5]
One-sided patio	3.2	3.7	[−39.0, −34.2]
Concrete canyon	2.7	10.2	[−48.7, −44.0]
Plant fence	4.9	9.4	[−38.2, −34.5]
Small boulders	3.5	12.8	[−41.5, −37.2]
Sandy flat beach	4.2	4.0	[−40.8, −37.5]
Dense bamboo	5.0	11.6	[−38.2, −35.2]
Dry tall underbrush	3.6	8.4	[−36.4, −33.2]

bandwidth decreases, but is for many scenarios in the range of 50 MHz and beyond. Accordingly, low-bandwidth channels in this frequency range can be considered as frequency nonselective. Other references propose path-loss values in the range of $\gamma = 4$ [245, 648]. In reference [563], the parameters PL(1m)[dB] $= -30$ and $\gamma = 3.5$ are used to model transmission using the μAMPS-1 nodes (2.4 GHz, 1 Mbps FSK transceiver).

4.2.5 Spread-spectrum communications

In spread-spectrum systems [293, 297, 557], the bandwidth occupied by the transmitted waveforms is much larger than what would be really needed to transmit the given user data.[11] The user signal is *spreaded* at the transmitter and *despreaded* at the receiver. By using a wideband signal, the effects of narrowband noise/interference are reduced. Spread-spectrum systems offer an increased robustness against multipath effects but pay the price of a more complex receiver operation compared to conventional modulation schemes.

The two most popular kinds of spread-spectrum communications are Direct Sequence Spread Spectrum (DSSS) and Frequency Hopping Spread Spectrum (FHSS).

Direct sequence spread spectrum

In Direct Sequence Spread Spectrum (DSSS), the transmission of a data bit of duration t_b is replaced by transmission of a finite **chip sequence** $\mathbf{c} = c_1 c_2 \ldots c_n$ with $c_i \in \{0, 1\}$ if the user bit is a logical one, or $\overline{c_1 c_2} \ldots \overline{c_n}$ if it is a logical zero ($\overline{c_i}$ is the logical inverse of c_i). Each chip c_i has duration $t_c = t_b/n$, where n is called the **spreading factor** or **gain**. Each chip is then modulated with a digital modulation scheme like BPSK or QPSK. Since the spectrum occupied by a digital signal is roughly inverse of the symbol duration, the spectrum of the chip sequence is much wider than the spectrum the user data signal would require in case of direct modulation. The intention is that the chip duration becomes smaller than the average or RMS delay spread value and the channel becomes, thus, frequency selective. Therefore, when multipath fading is present, a chip sequence $\mathbf{c}$ coming from an LOS and a delayed copy $\mathbf{c}$ (of the same chip sequence overlap, and the delay

[11] Information theorists would say that the Fourier bandwidth (describing the occupied spectrum) is much larger than the Shannon bandwidth (describing the number of dimensions of the signal space used per second) [542].

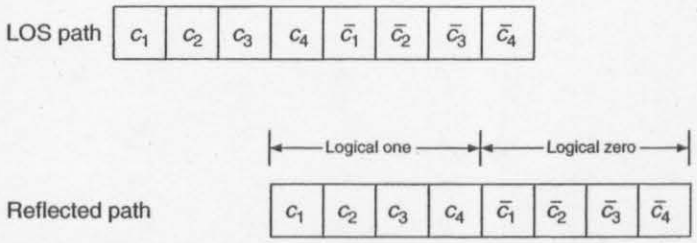


Figure 4.8 Direct sequence spread-spectrum example

difference (the **lag**) between these amounts to more than one chip duration. This is exploited by proper design of the chip sequences: these are **pseudorandom sequences** chosen such that the autocorrelation between a chip sequence and a lagged version of itself has a peak for lag zero and almost vanishes for all nonzero lags.

To explain this, consider the example shown in Figure 4.8. Both a direct LOS path and a reflected path are present, with the lag corresponding to three chip durations. The direct LOS chip sequence is given by $\mathbf{c} = c_1 c_2 \ldots c_n$ followed by $\mathbf{c}^{-1} = \overline{c_1 c_2} \ldots \overline{c_n}$, whereas the chip sequence from the reflected path starts with a lag of three chips. Somewhat simplified, the operation of the receiver can be described as follows (coherent matched filter receiver): Let us assume that the receiver is synchronized to the direct LOS path. It compares the incoming chip sequence with the well-known reference sequence $\mathbf{c}$ by computing the inner product (term-wise multiplication and final summation in terms of modulo-2 operations). If the received sequence is the same as $\mathbf{c}$, then this operation yields the value n, if the incoming chip sequence is $\mathbf{c}^{-1}$, then the result is $-n$. By proper choice of the chip sequence, it can be achieved that the inner product formed between the chip sequence and a shifted/lagged version of it assumes absolute values smaller than n. For example, the 11-chip **Barker sequence** used in IEEE 802.11 [467] assumes for all shifted versions only the values -1, 0, or 1.[12] Delayed copies distort the direct signal in the same way as AWGN does. Thus, DSSS increases robustness against multipath effects.

However, there are also downsides. First, receivers must be properly synchronized with the transmitter, and second, there is the issue of management of chip sequences. In systems like IEEE 802.11 with DSSS Physical Layer (PHY) or IEEE 802.15.4, there is only a single chip sequence used by all nodes. Proper measures at the MAC level must be taken to avoid collisions. It is also possible to assign different chip sequences or **codes** to different users, which then can transmit in parallel and create only minor distortion to each other. Such an approach is called Code Division Multiple Access (CDMA) and is used, for example, in UMTS [847]. However, immediately the question how codes are assigned to nodes ("code management") comes up.

Frequency hopping spread spectrum

In Frequency Hopping Spread Spectrum (FHSS) systems like Bluetooth [318, 319] and the (outdated) FHSS version of IEEE 802.11, the available spectrum is subdivided into a number of equal-sized **subbands** or **channels** (not to be confused with the physical channels discussed above); Bluetooth and IEEE 802.11 divide their spectrum in the 2.4-GHz range into 78 subbands 1-MHz wide. The user data is always transmitted within one channel at a time; its bandwidth is thus limited. All nodes in a network hop synchronously through the channels according to a prespecified

[12] When the inner product of a chip sequence with a shifted version of itself assumes "large" values only for lag zero, but comparably small values for all other lags, it is also called **nearly orthogonal**.

schedule. This way, a channel currently in a deep fade is left at some point in time and the nodes switch to another, hopefully, good channel. Different networks can share the same geographic area by using (mostly) nonoverlapping hopping schedules.

As an example, the FHSS version of IEEE 802.11 hops with 2.5 Hz and many packets can be transmitted before the next hop. In Bluetooth, the hopping frequency is 1.6 kHz and at most one packet can be transmitted before the next hop. Packets can have lengths corresponding to one, three, or five hops. During a longer packet, hopping is suppressed – the packet is transmitted at the same frequency. Once a packet is finished, the systems continues with the frequency it would have reached if the long packet had been absent.

4.2.6 Packet transmission and synchronization

The MAC layer above the physical layer uses **packets** or **frames** as the basic unit of transmission.[13] From the perspective of the MAC layer, such a frame has structure; for the transceiver, however, it is just a block of bits. Transceivers perform the functions of modulation and demodulation along with associated high- and intermediate-frequency processing, typically in hardware, and provide an interface to the physical layer. They are discussed in Section 2.2.4.

The receiver must know certain properties of an incoming waveform to make any sense of it and to detect a frame, including its frequency, phase, start and end of bits/symbols, and start and end of frames [772, Chap. 8], [286]. What is the root of this **synchronization problem**? The generation of sinusoidal carriers and of local clocks (with respect to which symbol times are expressed) involves **oscillators** of a certain **nominal frequency**. However, because of production inaccuracies, temperature differences, aging effects, or any of several other reasons, the *actual frequency* of oscillators deviates from the nominal frequency. This **drift** is often expressed in **parts per million (ppm)** and gives the number of additional or missing oscillations a clock makes in the amount of time needed for one million oscillations at the nominal rate. As a rule of thumb, the cheaper the oscillator, the more likely are larger drifts.

To compensate this drift, the receiver has to *learn* about the frequency or time base of the transmitter. The receiver has to extract synchronization information from the incoming waveform. An often-found theme for such approaches is the distinction between **training** (or **acquisition**) and **tracking** phases. Frames are equipped with a well-known **training sequence** that allows the receiver to learn about the detailed parameters of the transmitter, for example, its clock rate – the receive can "train" its parameters. This training sequence is often placed at the beginning of frames (for example, in IEEE 802.11 [467] or IEEE 802.15.4 [468]), but sometimes it is placed in the middle (e.g. in GSM [848]). In the first case we speak of a **preamble**, and in the second case of a midamble. In either case, the training sequence imposes some overhead. As an example, in IEEE 802.15.4, the preamble consists of 32 zero bits.

After the receiver has successfully acquired initial synchronization from the training sequence, it enters a tracking mode, continuously readjusting its local oscillator.

Important synchronization problems are:

Carrier synchronization The receiver has to learn the frequency and, for coherent detection schemes, also the phase of the signal. A frequency drift can be caused by oscillators or by Doppler shift in case of mobile nodes. One way to achieve frequency synchronization is to let the transmitter occasionally send packets with known spectral shape and to let the receiver scan some portion of the spectrum around the nominal frequency band for this shape; for example, in the GSM system, special **frequency correction bursts** are used to

[13] In OSI terminology, this would be MAC PDUs. In fact, packets and frames are two words for the same thing; however, the word frame tends to be used more often when discussing lower layers.

this end [848, Chap. 3]. The phase varies typically much faster than the frequency; accordingly, phase synchronization must be done more often than frequency synchronization [286]. Phase synchronization can be avoided in noncoherent detection schemes but at the price of a higher BER at the same transmit power.

Bit/symbol synchronization Having acquired carrier synchronization, the receiver must determine both the symbol duration as well as the start and end of symbols to demodulate them successfully. The continuous readjustment in the tracking phase requires sufficient "stimuli" indicating symbol bounds. This can be explained with the example of OOK, where logical zeros are modulated as the absence of any carrier. If a long run of zeros occurs in the data, the receiver clock gets no stimulus for readjustment and may drift away from the transmitter clock, this way adding spurious symbols or skipping symbols. For example, for the RF Monolithics TR1000 transceiver used in the Mica motes, more than four consecutive zero or one bits should be avoided [351]. This situation can be avoided by choosing coding schemes with a sufficient number of logical ones, by bit-stuffing techniques, or by **scrambling** where the data stream is shifted through a linear-feedback shift register. The scrambling technique is, for example, applied in IEEE 802.11 and no extra symbols have to be sent. The other schemes incur some overhead symbols.

Frame synchronization The receiver of a frame must be able to detect where the frame starts and where it ends, that is, the frame bounds. Frame synchronization assumes that bit/symbol synchronization is already acquired. There are several techniques known for framing [327], including time gaps, length fields, usage of special flag sequences along with bit-stuffing techniques to avoid the occurrence of these sequences in the packet data, and others. One technique to mark the start of a frame is the approach of IEEE 802.15.4, where the preamble is immediately followed by a well-known Start Frame Delimiter (SFD). This SFD is part of the physical layer header, not of the data part, and thus no measures to avoid the SFD pattern in the data part have to be taken.

Let us discuss a simple example (Figure 4.9). In the Mica motes [351], one option for modulation is OOK. Accordingly, bits are represented by two transmission power levels: a power level of zero corresponds to a logical zero, whereas a nonzero power level corresponds to a logical one (ignoring the noise floor). A packet consists of a preamble, a start frame delimiter, and a data part. A long idle period on the medium is interpreted as boundary between packets. Within such a long idle period, the receiver of a packet needs to sample the medium for activity only occasionally. The time between samples must be smaller than the preamble length not to miss it, but large enough

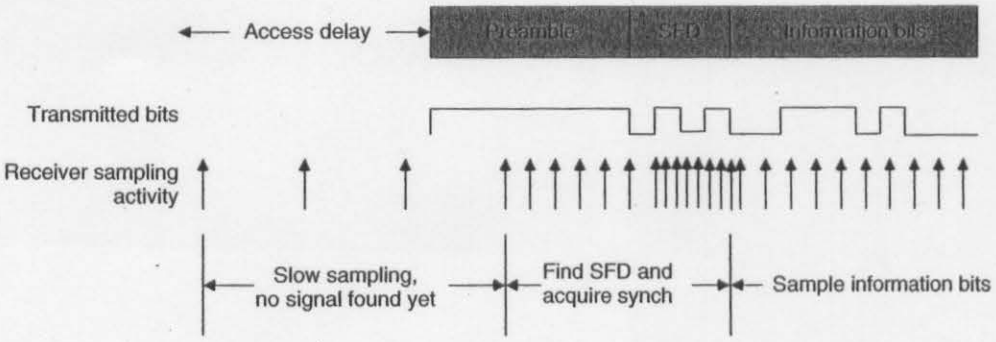


Figure 4.9 Example for sampling and synchronization (adapted from reference [351, Fig. 5])

to keep the energy costs induced by sampling. When sampling reveals activity in the channel, its frequency is increased to find the end of the preamble and to derive the length of a transmitted bit from the SFD. Once this information is determined, the receiver samples the medium in the mid of the data bits. To avoid the presence of long idle periods in the data part and misinterpretation as packet boundary, the length of runs of zeros (and ones) must be bounded, for example, by four. This has to be achieved by proper transformation of the user data.

4.2.7 Quality of wireless channels and measures for improvement

As opposed to wired channels, wireless channels often have a poorer quality in terms of bit/symbol error rate. The actual channel quality depends on many factors, including frequency, distance between transmitter and receiver, and their relative speed, propagation environment (number of paths and their respective attenuation), technology, and much more. Consequently, there is no such thing as "the" wireless channel. Many measurements of error rates have appeared in the literature; two of them are references [13, 223].

A great deal of work has been devoted to improve transmission quality on wireless channels, working on the physical as well as on higher layers and in many cases not taking energy concerns or other constraints specific for wireless sensor networks into account. Some of the mechanisms developed are the following:

Optimization of transmission parameters The choice of modulation scheme as well as the choice of radiated power (within legal constraints) can influence the BER significantly. Another control knob is the choice of packet sizes and the structure of packets. This is discussed in Chapter 6.

Diversity mechanisms All **diversity techniques** [682, Chap. 7], [625] seek to obtain and exploit statistically independent (or at least uncorrelated) replicas of the same signal. Simply speaking, it is hoped that even if one replica is in a deep fade and delivers symbol errors, another replica is currently good. The receiver tries to pick the best of all replicas. In **explicit diversity** schemes, the multiple copies are explicitly created by the transmitter, by sending the same packet over another frequency, during another time slot, or sending it into another spatial direction. In **implicit diversity** schemes, the signal is sent only once, but multiple copies are created *in the channel* through multipath propagation. In either case, the receiver needs mechanisms to take advantage of the multiple copies. One simple example is the so-called **receive diversity**, where the receiver is equipped with two or more appropriately spaced antennas and the receiver combines the different signals (e.g. by so-called **selection combining**: pick the signal with the best quality; or by **maximum ratio combining**: sum up all signals, weighted by their quality). Receive diversity works best when the signals at the two antennas are independent or at least uncorrelated. As a rule of thumb, this can be achieved with an antenna spacing of at least 40–50 % of the wavelength [682, Chap. 5].

Equalization **Equalization** techniques [682, Chap. 7], [660] are useful to combat InterSymbol Interference (ISI). Equalization works as follows: The transmitter sends a well-known symbol pattern/waveform, the so-called **training sequence**. The equalizer at the receiver works in two modes: **training** and **tracking**. During the training phase, the equalizer analyzes the received version of the well-known pattern, learns the mode of distortion, and computes an algorithm for "inverting" the distortion. In the tracking phase, the remaining packet is analyzed by applying the inversion algorithm to it and the equalizer continually readjusts the inversion algorithm. Equalization requires some signal processing at the receiver and the channel is assumed to be stationary during the packet transmission time. As a side effect, the training sequence can also be used to acquire **bit synchronization**.

Forward error correction (FEC) The transmitter accepts a stream or a block of user data bits or source bits, adds suitable redundancy, and transmits the result to the receiver. Depending on the amount and structure of the redundancy, the receiver might be able to correct some bit/symbol errors. It is known that AWGN channels have a higher capacity than Rayleigh fading channels and many coding schemes achieve better BER performance on AWGN than on fading channels with their bursty errors [79]. The operation of **interleaving** applies a permutation operation to a block of bits, hoping to distribute bursty errors smoothly and letting the channel "look" like an AWGN channel. FEC is discussed in some more detail in Section 6.2.3.

ARQ The basic idea of ARQ protocols [322, 511] can be described as follows: The transmitter prepends a header and appends a checksum to a data block. The resulting packet is then transmitted. The receiver checks the packet's integrity with the help of the checksum and provides some feedback to the transmitter regarding the success of packet transmission. On receiving negative feedback, the transmitter performs a retransmission. ARQ protocols are discussed in Section 6.2.2.

4.3 Physical layer and transceiver design considerations in WSNs

So far, we have discussed the basics of the PHY without specific reference to wireless sensor networks. Some of the most crucial points influencing PHY design in wireless sensor networks are:

- Low power consumption.
- As one consequence: small transmit power and thus a small transmission range.
- As a further consequence: low duty cycle. Most hardware should be switched off or operated in a low-power standby mode most of the time.
- Comparably low data rates, on the order of tens to hundreds kilobits per second, required.
- Low implementation complexity and costs.
- *Low degree of mobility.*
- A small form factor for the overall node.

In this section, we discuss some of the implications of these requirements.

In general, in sensor networks, the challenge is to find modulation schemes and transceiver architectures that are simple, low-cost but still robust enough to provide the desired service.

4.3.1 Energy usage profile

The choice of a small transmit power leads to an energy consumption profile different from other wireless devices like cell phones. These pivotal differences have been discussed in various places already but deserve a brief summary here.

First, the radiated energy is small, typically on the order of 0 dBm (corresponding to 1 mW). On the other hand, the overall transceiver (RF front end and baseband part) consumes much more energy than is actually radiated; WANG et al. [855] estimate that a transceiver working at frequencies beyond 1 GHz takes 10 to 100 mW of power to radiate 1 mW. In reference [115, Chap. 3], similar numbers are given for 2.4-GHz CMOS transceivers: For a radiated power of 0 dBm, the transmitter uses actually 32 mW, whereas the receiver uses even more, 38 mW. For the Mica motes, 21 mW are consumed in transmit mode and 15 mW in receive mode [351]. These numbers coincide well

with the observation that many practical transmitter designs have efficiencies below 10 % [46] at low radiated power.

A second key observation is that for small transmit powers the transmit and receive modes consume more or less the same power; it is even possible that reception requires more power than transmission [670, 762]; depending on the transceiver architecture, the idle mode's power consumption can be less or in the same range as the receive power [670]. To reduce average power consumption in a low-traffic wireless sensor network, keeping the transceiver in idle mode all the time would consume significant amounts of energy. Therefore, it is important to put the transceiver into sleep state instead of just idling. It is also important to explicitly include the received power into energy dissipation models, since the traditional assumption that receive energy is negligible is no longer true.

However, there is the problem of the **startup energy/startup time**, which a transceiver has to spend upon waking up from sleep mode, for example, to ramp up phase-locked loops or voltage-controlled oscillators. During this startup time, no transmission or reception of data is possible [762]. For example, the μAMPS-1 transceiver needs a startup time of 466 µs and a power dissipation of 58 mW [561, 563]. Therefore, going into sleep mode is unfavorable when the next wakeup comes fast. It depends on the traffic patterns and the behavior of the MAC protocol to schedule the transceiver operational state properly. If possible, not only a single but multiple packets should be sent during a wakeup period, to distribute the startup costs over more packets. Clearly, one can attack this problem also by devising transmitter architectures with faster startup times. One such architecture is presented in reference [855].

A third key observation is the relative costs of communications versus computation in a sensor node. Clearly, a comparison of these costs depends for the communication part on the BER requirements, range, transceiver type, and so forth, and for the computation part on the processor type, the instruction mix, and so on. However, in [670], a range of energy consumptions is given for Rockwell's WIN nodes, UCLA's WINS NG 2.0 nodes, and the MEDUSA II nodes. For the WIN nodes, 1500 to 2700 instructions can be executed per transmitted bit, for the MEDUSA II nodes this ratio ranges from 220:1 up to 2900:1, and for the WINS NG nodes, it is around 1400:1. The bottom line is that computation is cheaper than communication!

4.3.2 Choice of modulation scheme

A crucial point is the choice of modulation scheme. Several factors have to be balanced here: the required and desirable data rate and symbol rate, the implementation complexity, the relationship between radiated power and target BER, and the expected channel characteristics.

To maximize the time a transceiver can spend in sleep mode, the transmit times should be minimized. The higher the data rate offered by a transceiver/modulation, the smaller the time needed to transmit a given amount of data and, consequently, the smaller the energy consumption.

A second important observation is that the power consumption of a modulation scheme depends much more on the symbol rate than on the data rate [115, Chap. 3]. For example, power consumption measurements of an IEEE 802.11b Wireless Local Area Network (WLAN) card showed that the power consumption depends on the modulation scheme, with the faster Complementary Code Keying (CCK) modes consuming more energy than DBPSK and DQPSK; however, the relative differences are below 10 % and all these schemes have the same symbol rate. It has also been found that for the μAMPS-1 nodes the power consumption is insensitive to the data rate [762].

Obviously, the desire for "high" data rates at "low" symbol rates calls for m-ary modulation schemes. However, there are trade-offs:

- m-ary modulation requires more complex digital and analog circuitry than 2-ary modulation [762], for example, to parallelize user bits into m-ary symbols.

Table 4.3 Bandwidth efficiency η_{BW} and E_b/N_0[dB] required at the receiver to reach a BER of 10^{-6} over an AWGN channel for m-ary orthogonal FSK and PSK (adapted from reference [682, Chap. 6])

m	2	4	8	16	32	64
m-ary PSK:η_{BW}	0.5	1.0	1.5	2.0	2.5	3.0
m-ary PSK:E_b/N_0	10.5	10.5	14.0	18.5	23.4	28.5
m-ary FSK:η_{BW}	0.40	0.57	0.55	0.42	0.29	0.18
m-ary FSK:E_b/N_0	13.5	10.8	9.3	8.2	7.5	6.9

- Many m-ary modulation schemes require for increasing m an increased E_b/N_0 ratio and consequently an increased radiated power to achieve the same target BER; others become less and less bandwidth efficient. This is exemplarily shown for coherently detected m-ary FSK and PSK in Table 4.3, where for different values of m, the achieved bandwidth efficiencies and the E_b/N_0 required to achieve a target BER of 10^{-6} are displayed. However, in wireless sensor network applications with only low to moderate bandwidth requirements, a loss in bandwidth efficiency can be more tolerable than an increased radiated power to compensate E_b/N_0 losses.
- It is expected that in many wireless sensor network applications most packets will be short, on the order of tens to hundreds of bits. For such packets, the startup time easily dominates overall energy consumption, rendering any efforts in reducing the transmission time by choosing m-ary modulation schemes irrelevant.

Let us explore the involved trade-offs a bit further with the help of an example.

Example 4.1 (Energy efficiency of m-ary modulation schemes) Our goal is to transmit data over a distance of $d = 10\,\text{m}$ at a target BER of 10^{-6} over an AWGN channel having a path-loss exponent of $\gamma = 3.5$ (corresponding to the value determined in reference [563]). We compare two families of modulations: coherently detected m-ary PSK and coherently detected orthogonal m-ary orthogonal FSK. For these two families we display in Table 4.3, the bandwidth efficiencies η_{BW} and the E_b/N_0 in dB required at the receiver to reach a BER of 10^{-6} over an AWGN channel.

From the discussion in Section 4.2.3, the relationship between E_b/N_0 and the received power at a distance d is given as:

$$\begin{aligned}\frac{E_b}{N_0} &= \text{SNR} \cdot \frac{1}{R} = \frac{P_{\text{rcvd}}(d)}{N_0} \cdot \frac{1}{R} \\ &= \frac{1}{N_0 \cdot R} \cdot \frac{P_{\text{tx}} \cdot G_t \cdot G_r \cdot \lambda^2}{(4\pi)^2 \cdot d_0^{\gamma} \cdot L} \cdot \left(\frac{d_0}{d}\right)^{\gamma},\end{aligned} \tag{4.8}$$

which can be easily solved for P_{tx} given a required E_b/N_0 value and data rate R. We denote the solution as $P_{\text{tx}}\left(\frac{E_b}{N_0}, R\right)$. One example: From Table 4.3 we obtain that 16-PSK requires an E_b/N_0 of 18.5 dB to reach the target BER. When fixing the parameters $G_t = G_r = L = 1$, $\lambda = 12.5\,\text{cm}$ (according to a 2.4 GHz transceiver), reference distance $d_0 = 1\,\text{m}$, distance $d = 10\,\text{m}$, a data rate of $R = 1\,\text{Mbps}$, and a noise level of $N_0 = -180\,\text{dB}$ this corresponds to $P_{\text{tx}}\,(18.5\,\text{dB}, R) \approx 2.26\,\text{mW}$.

We next utilize a transceiver energy consumption model developed in references [762, 855] that incorporates startup energy and transmit energy. In this model, it is assumed that during

the startup time mainly a frequency synthesizer is active, consuming energy P_{FS}, while during the actual waveform transmission power is consumed by the frequency synthesizer, the modulator (using P_{MOD}), and the radiated energy $P_{tx}(\cdot,\cdot)$. The power amplifier is not explicitly considered. Using reference [855], we assume $P_{FS} = 10\,\text{mW}$, $P_{MOD} = 2\,\text{mW}$ and a symbol rate of $B = 1$ M symbols/sec. The duration of the startup time is T_{start}. For the case of binary modulation, we assume the following energy model:

$$E_{binary}\left(\frac{E_b}{N_0}, B\right) = P_{FS} \cdot T_{start} + \left(P_{MOD} + P_{FS} + P_{tx}\left(\frac{E_b}{N_0}, B\right)\right) \cdot \frac{n}{B},$$

where n is the number of data bits to transmit in a packet. For the case of m-ary modulation, it is assumed that the power consumption of the modulator and the frequency synthesizer are increased by some factors $\alpha \geq 1$, $\beta \geq 1$, such that the overall energy expenditure is:

$$E_{m\text{-ary}}\left(\frac{E_b}{N_0}, B \cdot \log_2 m\right) = \beta \cdot P_{FS} \cdot T_{start} + \left(\alpha \cdot P_{MOD} + \beta \cdot P_{FS} + P_{tx}\left(\frac{E_b}{N_0}, B \cdot \log_2 m\right)\right) \cdot \frac{n}{B \cdot \log_2(m)}.$$

Accepting the value $\beta = 1.75$ from reference [855] for both PSK and FSK modulation, one can evaluate the ratio $\frac{E_{m\text{-ary}}(\cdot,\cdot)}{E_{binary}(\cdot,\cdot)}$ to measure the energy advantage or disadvantage of m-ary modulation over binary modulation. As an example, we show this ratio in Figure 4.10 for varying $m \in \{4, 8, 16, 32, 64\}$, with $\alpha = 2.0$, a startup time of 466 µs, and two different packet sizes, 100 bits and 2000 bits. The two upper curves correspond to a packet size of 100 bits; the two lower curves correspond to the packet size of 2000 bits. Other results obtained with a shorter startup time of 100 µs or $\alpha = 3.0$ look very similar. One can see that for large packet sizes m-ary FSK modulation is favorable, since the actual packet transmission times are shortened and furthermore the required E_b/N_0 *decreases* for increasing m, at the expense of a reduced bandwidth efficiency, which translates into a wider required spectrum (the FSK scheme is orthogonal FSK). For m-ary PSK, only certain values of m give an energy advantage; for larger m the increased E_b/N_0 requirements outweigh the gains due to reduced transmit times. For small packet sizes, the binary modulation schemes are more energy efficient for both PSK and FSK, because the energy costs are dominated by the startup time. If one reduces β to $\beta = 1$ (assuming no extra energy consumption of the frequency synthesizer due to m-ary modulation), then m-ary modulation would, for all parameters under consideration, be truly better than binary modulation. The results presented in reference [855] indicate that the advantage of m-ary modulation increases as the startup time decreases. For shorter startup times also the packet lengths required to make m-ary modulation pay out are smaller.

Can we conclude from this that it is favorable to use large packets? Unfortunately, the answer is: it depends. As we will see in Chapter 6, longer packets at the same bit error rate and without employing error-correction mechanisms lead to higher packet error rates, which in turn lead to retransmitted packets, easily nullifying the energy gains of choosing m-ary modulation. A careful joint consideration of modulation and other schemes for increasing transmission robustness (FEC or ARQ schemes) is needed.

But it can be beneficial to transmit multiple short packets during a single wakeup period, thus achieving a lower relative influence of the startup costs per packet [562].

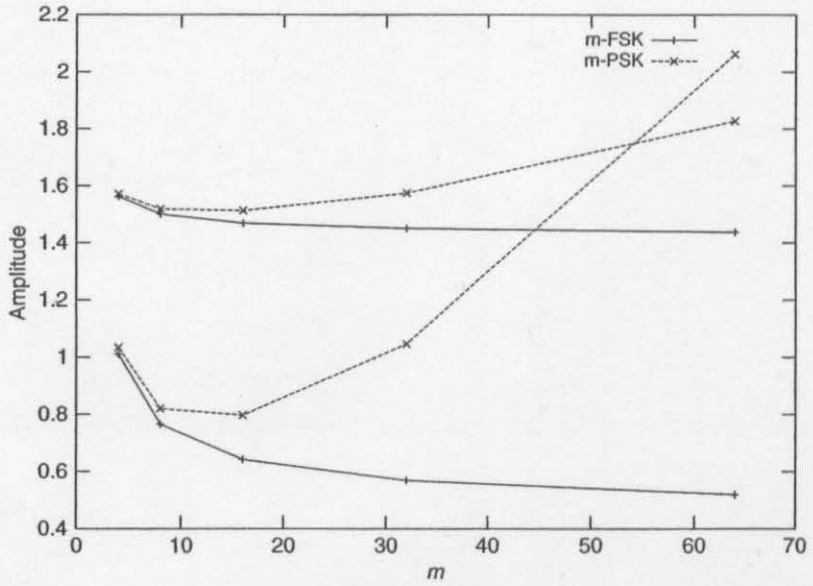


Figure 4.10 Comparison of the energy consumption of m-ary FSK/PSK to binary FSK/PSK for $\alpha = 2.0$ and startup time of 466 µs.

Clearly, this example provides only a single point in the whole design space. The bottom line here is that the choice of modulation scheme depends on several interacting aspects, including technological factors (in the example: α, β), packet size, target error rate, and channel error model (in reference [855], a similar example is carried out for the case of Rayleigh fading). The optimal decision would have to properly balance the modulation scheme and other measures to increase transmission robustness, since these also have energy costs:

- With retransmissions, entire packets have to be transmitted again.
- With FEC coding, more bits have to be sent and there is additional energy consumption for coding and decoding. While coding energy can be neglected, and the receiver needs significant energy for the decoding process [563]. This is especially cumbersome if the receiver is a power-constrained node. Coding and retransmission schemes are discussed in more detail in Chapter 6.
- The cost of increasing the radiated power [855] depends on the efficiency of the power amplifier (compare Section 2.2.4), but the radiated power is often small compared to the overall power dissipated by the transceiver, and additionally this drives the PA into a more efficient regime.[14]

In [670], a similar analysis as in our example has been carried out for m-ary QAM. Specifically, the energy-per-bit consumption (defined as the overall energy consumption for transmitting a packet of n bits divided by n) of different m-ary QAM modulation schemes has been investigated for different packet sizes, taking startup energy and the energy costs of power amplifiers as well as PHY and MAC packet overheads explicitly into account. For the particular setup used in this

[14] Of course, one disadvantage of using an increased transmit power is an increased interference for other transmissions and thus a decreased overall network capacity. However, this plays no role during low-load situations, which prevail in wireless sensor networks – unless event storms or other correlated traffic models are present.

investigation, 16-QAM seems to be the optimum modulation schemes for all different sizes of the user data.

4.3.3 Dynamic modulation scaling

Even if it is possible to determine the optimal scheme for a given combination of BER target, range, packet sizes and so forth, such an optimum is only valid for short time; as soon as one of the constraints changes, the optimum can change, too. In addition, other constraints like delay or the desire to achieve high throughput can dictate to choose higher modulation schemes.

Therefore, it is interesting to consider methods to *adapt* the modulation scheme to the current situation. Such an approach, called **dynamic modulation scaling**, is discussed in reference [738]. In particular, for the case of m-ary QAM and a target BER of 10^{-5}, a model has been developed that uses the symbol rate B and the number of levels per symbol m as parameters. This model expresses the energy required per bit and also the achieved delay per bit (the inverse of the data rate), taking into account that higher modulation levels need higher radiated energy. Extra startup costs are not considered. Clearly, the bit delay decreases for increasing B and m. The energy per bit depends much more on m than on B. In fact, for the particular parameters chosen, it is shown that both energy per bit and delay per bit are minimized for the maximum symbol rate. With modulation scaling, a packet is equipped with a delay constraint, from which directly a minimal required data rate can be derived. Since the symbol rate is kept fixed, the approach is to choose the smallest m that satisfies the required data rate and which thus minimizes the required energy per bit. Such delay constraints can be assigned either explicitly or implicitly. One approach explored in the paper is to make the delay constraint depend on the packet backlog (number of queued packets) in a sensor node: When there are no packets present, a small value for m can be used, having low energy consumption. As backlog increases, m is increased as well to reduce the backlog quickly and switch back to lower values of m. This modulation scaling approach has some similarities to the concept of **dynamic voltage scaling** discussed in Section 2.2.2.

4.3.4 Antenna considerations

The desired small form factor of the overall sensor nodes restricts the size and the number of antennas. As explained above, if the antenna is much smaller than the carrier's wavelength, it is hard to achieve good antenna efficiency, that is, with ill-sized antennas one must spend more transmit energy to obtain the same radiated energy.

Secondly, with small sensor node cases, it will be hard to place two antennas with suitable distance to achieve receive diversity. As discussed in Section 4.2.7, the antennas should be spaced apart at least 40–50 % of the wavelength used to achieve good effects from diversity. For 2.4 GHz, this corresponds to a spacing of between 5 and 6 cm between the antennas, which is hard to achieve with smaller cases.

In addition, radio waves emitted from an antenna close to the ground – typical in some applications – are faced with higher path-loss coefficients than the common value $\alpha = 2$ for free-space communication. Typical attenuation values in such environments, which are also normally characterized by obstacles (buildings, walls, and so forth), are about $\alpha = 4$ [245, 648].

Moreover, depending on the application, antennas must not protrude from the casing of a node, to avoid possible damage to it. These restrictions, in general, limit the achievable quality and characteristics of an antenna for wireless sensor nodes.

Nodes randomly scattered on the ground, for example, deployed from an aircraft, will land in random orientations, with the antennas facing the ground or being otherwise obstructed. This can lead to nonisotropic propagation of the radio wave, with considerable differences in the strength of the emitted signal in different directions. This effect can also be caused by the design of an

antenna, which often results in considerable differences in the spatial propagation characteristics (so-called lobes of an antenna).

Antenna design is an issue in itself and is well beyond the scope of this book. Some specific considerations on antenna design for wireless sensor nodes are discussed in [115, Chap. 8].

4.4 Further reading

Jointly optimizing coding and modulation BIGLIERI et al. [79] consider coding and modulation from an information-theoretic perspective for different channel models, including the AWGN, flat fading channels and block fading channels. Specifically, the influence of symbol-by-symbol power control at the transmitter in the presence of channel-state information such that deep fades are answered with higher output powers ("channel inversion"), of receiver diversity and interleaving and of coding schemes with unequal protection (i.e., user bits of different importance are encoded differently) on the channel capacity are discussed. One particularly interesting result is that the capacity of a Rayleigh fading channel with power control can be higher than the capacity of an AWGN channel with the same average radiated power.

DSSS in WSN Some efforts toward the construction of DSSS transceivers for wireless sensor networks with their space and power constraints are described in references [155, 280, 281]. In addition, MYERS et al. [580] discuss low-power spread-spectrum transceivers for IEEE 802.11.

Energy efficiency in GSM Reducing energy consumption is an issue not only in wireless sensor networks but also in other types of systems, for example, cellular systems. For the interested: advanced signal processing algorithms for reducing power consumption of GSM transceivers are discussed in references [525].