Chapter 2
Optical Communication—A High-Level Perspective

As suggested in Chap. 1, the need for low-cost optical communication systems is omnipresent. Before discussing the design of an optical receiver in detail in the next chapters, it is important to understand the basic concepts of communication systems in general and the most significant aspects of an optical system in particular. An overview is given in this chapter.

2.1 The Communication Model

The objective of communication is to transport a message between two or more entities. One of these entities generates the data to be transported and is called the *transmitter*. A *channel* is needed to transport the message to the addressee, the *receiver*. This is shown schematically in Fig. 2.1. The transmitter consists of a driver and a light source. The driver conditions the digital data so that it can be applied to the light source, which is typically a *laser diode*. In this way, the signal is converted to the optical domain which is necessary to be guided through the optical fiber. At the other end of the fiber, the light signal is converted back to the electrical domain by a photodiode. However, the signal which is generated by the photodiode cannot be interpreted as digital data because typically it is greatly attenuated and distorted during transportation through the fiber. Therefore, the receiver needs to perform three operations before the data can be properly understood: *re-amplification*, *re-shaping* and *re-timing*. Because of this, such a receiver is sometimes called a *3R-receiver*.

A *transceiver* is the combination of a transmitter and a receiver in a single package. Mostly, such a transceiver is implemented in a *small form-factor pluggable* (SFP) format. An example is shown in Fig. 2.2. In this manner, a full-duplex communication link can be established by means of two half-duplex links in parallel.

The communication model of Fig. 2.1 can be used in a vast variety of application domains. Some of them were discussed in Chap. 1. Although they do not always fully comply with the presented model, the basic concept remains the same. For example, in an optical pick-up unit for a CD, DVD or Blu-ray disc, light is modulated by the disc which can therefore be considered as the transmitter. Moreover, the channel exists from air and a lens system instead of an optical fiber.

F. Tavernier, M. Steyaert, *High-Speed Optical Receivers with Integrated Photodiode in Nanoscale CMOS,* Analog Circuits and Signal Processing 5,
DOI 10.1007/978-1-4419-9925-2_2, © Springer Science+Business Media, LLC 2011

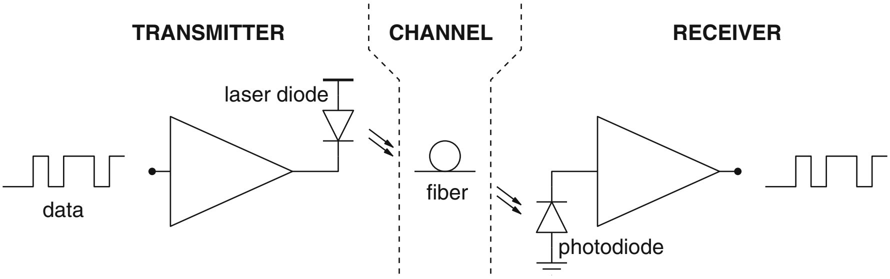
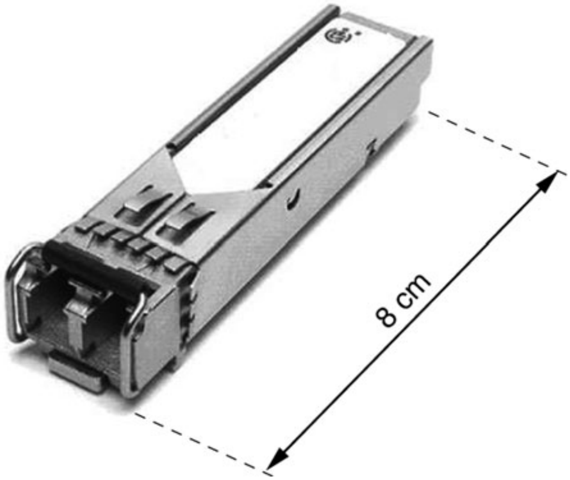


Fig. 2.1 A simple optical communication system, consisting of a transmitter, a channel and a receiver. The laser diode at the transmitter converts the electrical signal into modulated light while the photodiode at the receiver does the opposite

Fig. 2.2 An SFP optical transceiver; one fiber input is used to send data, the other fiber input is used to receive data



2.2 Properties of Binary Data

2.2.1 Random Binary Data

One of the common aspects for almost all optical communication systems is the type of data signals which are used. Logical ONEs and ZEROs typically modulate the amplitude of the transmitted lightwave. This will be clarified in Sect. 2.3. An example waveform of a binary data signal is shown in Fig. 2.3a. A high amplitude of the waveform corresponds with a logical ONE whereas a low amplitude corresponds with a logical ZERO. The duration of a single bit, the *bit period*, is represented by T_b. The reciprocal of T_b is called the *bit rate*. Since they are random in nature, a ONE and a ZERO occur typically with an equal probability. The waveform in Fig. 2.3a is

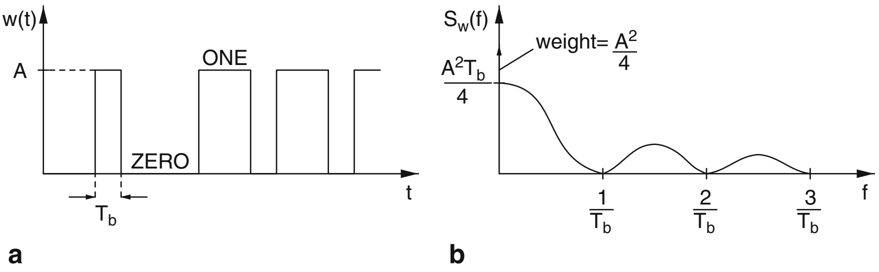


Fig. 2.3 Example of a logical binary signal and its accompanying spectrum. **a** Waveform. **b** Spectrum

mathematically described as:

$$w(t) = \sum_i b_i p(t - iT_b), \tag{2.1}$$

where $b_i = 0$ for a ZERO and $b_i = 1$ for a ONE. The pulse shape, $p(t)$, is defined as follows:

$$p(t) = \begin{cases} A & \text{if } 0 \le t \le T_b \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

The *power spectral density* (PSD) is normally used to analyze the frequency domain behavior of a signal. For a waveform that can be described with (2.1) and for a ONE and a ZERO having an equal probability to appear, the PSD is given by [Cou02]:

$$S_w(f) = \frac{|P(f)|^2}{4T_b}\left(1 + \frac{\delta(f)}{T_b}\right), \tag{2.3}$$

where $P(f)$ represents the Fourier transform of $p(t)$ and $\delta(f)$ is the *Dirac delta function*. If $p(t)$ is defined by (2.2), its Fourier transform is given by [Cou02]:

$$P(f) = AT_b \frac{\sin \pi f T_b}{\pi f T_b} e^{-j\pi f T_b}. \tag{2.4}$$

Using (2.4), (2.3) can be rewritten as:

$$S_w(f) = \frac{A^2 T_b}{4}\left(\frac{\sin \pi f T_b}{\pi f T_b}\right)^2 \left(1 + \frac{\delta(f)}{T_b}\right) \tag{2.5}$$

This spectral behavior is shown in Fig. 2.3b. Obviously, most of the signal power is contained in the frequency band below $f = 1/T_b$. Because the average of the data signal (2.1) is not zero, an impulse is present at DC. Interestingly, there is no spectral component at $1/T_b$, the bit rate of the signal. From now on the term *spectrum* is loosely used instead of PSD.

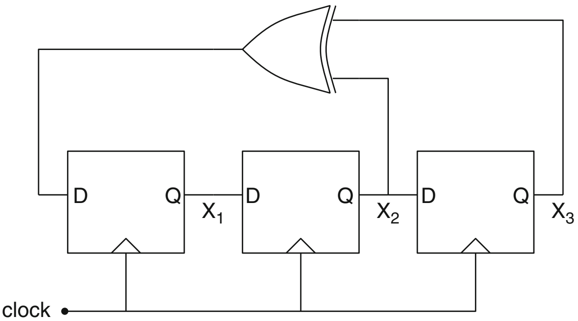
This type of data signal is known as *unipolar non-return-to-zero* (NRZ) signaling. It is unipolar because the signal is not symmetrical with respect to the zero level. This is actually the case when the data modulates the power of a lightwave: the power cannot drop below zero. As the amplitude remains constant during the entire bit period, the signal is said to be NRZ. A *return-to-zero* (RZ) waveform would return to the zero level for the second half of the bit interval. Compared to a NRZ signal, the spectrum of a RZ signal is contained in a frequency band that is twice as wide. This has an effect on the required bandwidth of the receiver which will be explained in Sect. 2.5. On the other hand, an impulse at $1/T_b$ is present in the spectrum of a RZ signal, which simplifies the re-timing task of a 3R-receiver which is performed by the *clock and data recovery* (CDR) system.

2.2.2 Pseudo-random Binary Data

In contrast to a truly random binary data stream, a *pseudo-random binary sequence* (PRBS) is completely predictable. Such a PRBS is typically produced by a *linear feedback shift register* (LFSR), of which an example is shown in Fig. 2.4a. An LFSR has a number of clocked memory elements and a linear feedback loop around (some of) them. The linear feedback is provided by means of the *exclusive OR* (XOR) gate which is in fact the only linear function of single bits.

The topology of Fig. 2.4a passes cyclically through a predefined number of states. Suppose the initial state is 111. At the next clock cycle, this initial state changes to 011. The succession of states for this topology is represented in Fig. 2.4b. After the last state, the circuit returns to its initial state and the process starts all over again.

The signal X_1 forms a PRBS as do X_2 and X_3, which are delayed versions of X_1. The length of the generated pattern is determined by the number of memory elements in the LFSR. For three memory elements, as is the case in Fig. 2.4, this



a

state	X_1	X_2	X_3
1	1	1	1
2	0	1	1
3	0	0	1
4	1	0	0
5	0	1	0
6	1	0	1
7	1	1	0

return to state 1

b

Fig. 2.4 Generation of a PRBS with a length of $2^3 - 1$. **a** LFSR. **b** State table

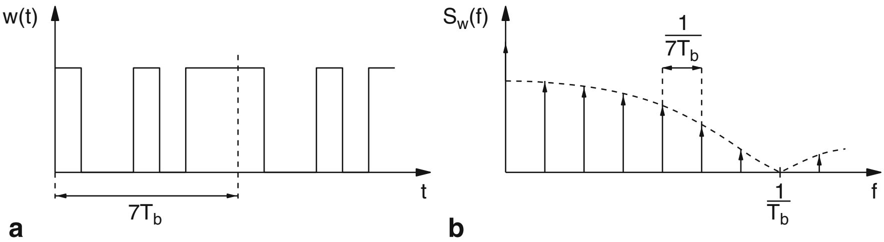


Fig. 2.5 PRBS as generated by the topology of Fig. 2.4a. **a** Waveform. **b** Spectrum

length is $2^3 - 1 = 7$. It can also be noticed that the maximum run length[1] equals 3. The number of ONEs and ZEROs differs by only one, so the pattern is said to be DC-balanced. In general, for an LFSR with m memory elements, the length of the generated PRBS is $2^m - 1$, the maximum run length equals m and the pattern is DC-balanced. A time domain representation of X_1, as generated by the topology in Fig. 2.4a, is shown in Fig. 2.5a.

The spectrum of a PRBS is fundamentally different from that of a truly random signal. Due to its repetitive character it contains only impulses. For the PRBS of Fig. 2.5a, the repetition rate equals $1/7T_b$. In the frequency domain, this is translated into a series of impulses with a mutual spacing of $1/7T_b$ which is shown in Fig. 2.5b. The height of every impulse is determined by an envelope that is formed by the spectrum of a single bit with a bit period T_b [Raz03].

A PRBS is typically used to scramble the actual data bits that need to be transmitted. This is done to eliminate a long run of ONEs or ZEROs which would cause the CDR to lose track of the clock signal and by doing so impede its proper operation. Another reason why a data signal is scrambled is to make sure that the spectrum of the actual transmitted signal is more or less constant in time. As a result, it is less dependent upon the actual message that is transmitted. This is advantageous to prevent interference with adjacent channels. On the contrary, a PRBS is almost always used to simulate and measure optical communication circuits because it is difficult to generate completely random binary waveforms. The PRBS is then used as a fictitious data signal. Typical run lengths for this vary between $2^5 - 1$ and $2^{31} - 1$.

2.3 The Laser Diode

At the transmitter side of an optical communication link, the data is converted to the optical domain by means of a laser diode. In contrast to a *light-emitting diode* (LED), a laser diode can be modulated at much higher speeds. Moreover, a laser diode can be viewed as a monochromatic light source as its output light energy is heavily concentrated around one particular wavelength. Furthermore, it produces an extremely

[1] The maximum number of consecutive ONEs or ZEROs.

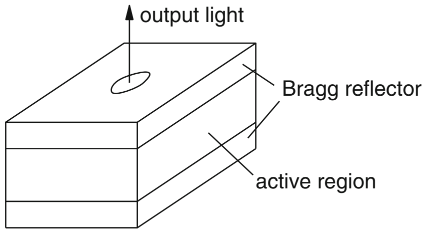


Fig. 2.6 Structural drawing of a VCSEL

focused and coherent[2] light beam that can travel a long distance before diverging. In the field of data communication, a *vertical-cavity surface-emitting laser* (VCSEL) is mostly preferred because of its low production cost. It is a semiconductor laser that emits its beam perpendicular to the top surface instead of from a side of the structure as in a conventional edge-emitting semiconductor laser. A structural drawing of a VCSEL is shown in Fig. 2.6. The light is generated inside the active region and is reflected back and forth between the *distributed Bragg reflectors* (DBR)[3]. Part of the generated light is able to escape through a hole in the upper reflector. The wavelength of the generated light of commercially available VCSELs is generally around 850 nm, which is short in comparison with other available types of lasers. They are mostly used for short communication links.

Electrically, a laser diode can be considered as a forward biased semiconductor diode. The relationship between the current through the laser, I_{laser}, and its optical output power is represented in Fig. 2.7. Up to the *threshold current*, I_{th}, the laser diode emits only incoherent light, just like a LED. If the current rises above this threshold, the lasing operation manifests itself. The optical output power P_{light} increases linearly with I_{laser}. The rate at which it rises is determined by $\delta P_{light}/\delta I_{laser}$, the *slope efficiency*. In contrast to a VCSEL, a LED has a zero threshold current and a lower slope efficiency.

Two techniques are available to generate a modulated optical signal: *external modulation* or *direct modulation*. When external modulation[4] is applied, the laser is actually a *continuous wave* (CW) laser that emits its light always at full power. The modulation process is performed by an external optical shutter that is open for a ONE and closed for a ZERO. A *Mach-Zehnder interferometer* is mostly used for this task. In a direct modulation scheme, the laser itself is turned on and off for a ONE and a ZERO respectively. Compared to external modulation, it is a simpler, more compact and cheaper modulation technique. For this reason, direct modulation is used throughout this work.

[2] Coherent light is light where all the photons have a definite phase relation to each other. This is due to the stimulated emission inside a laser.

[3] A DBR is a structure formed from multiple layers of alternating materials with different refractive indices. Its function is to reflect only light within a very narrow wavelength-band.

[4] External modulation is applied in systems that are less cost sensitive. It is able to output higher quality optical signals at higher bit rates.

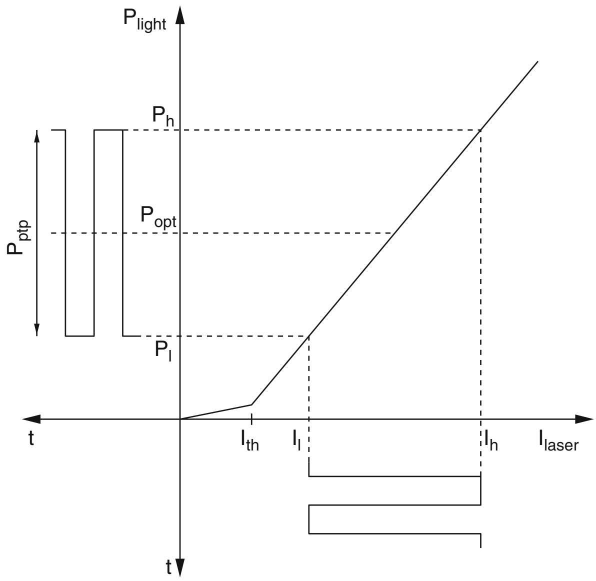


Fig. 2.7 Typical input-output characteristics of a VCSEL; the operation principle of direct modulation is also shown

In Fig. 2.7, the principle of direct modulation is shown. The current through the laser is modulated with the data and, as a consequence, the output light power varies between two levels: P_l and P_h for a ZERO and a ONE respectively. It can be noticed that the current through the laser is always larger than I_{th}. This is to minimize the turn-on delay t_{on} when switching from a ZERO to a ONE which can be approximated as follows [Säc05]:

$$t_{on} = \tau_c \ln \frac{I_h - I_l}{I_h - I_{th}}, \tag{2.6}$$

where τ_c represents the carrier lifetime which is typically around 3 ns. Equation (2.6) reveals that the turn-on delay decreases for increasing values of I_l. It drops to zero if I_l equals or exceeds I_{th}. A second reason why the laser should always be on is to minimize the jitter due to *relaxation oscillation*[5] [Raz03].

A faster operating speed is not the only consequence of choosing I_l above the threshold current of the laser diode. The other effect is a decreased optical signal swing for the same total transmitted power because the light is never turned off completely. This can be quantified by the *extinction ratio* r_e of the modulated laser diode:

$$r_e = \frac{P_h}{P_l}. \tag{2.7}$$

[5] Relaxation oscillation is an oscillation based upon the behavior of a system to return to equilibrium after being perturbed.

The average optical power for a ZERO and a ONE having an equal probability is given by:

$$P_{opt} = \frac{P_l + P_h}{2}. \tag{2.8}$$

This value is also shown in Fig. 2.7. It is clarified in Sect. 2.5.3 that the difference between P_h and P_l determines the quality of the received signal. This difference is called the peak-to-peak signal power P_{ptp}:

$$P_{ptp} = P_h - P_l. \tag{2.9}$$

Part of the average optical power P_{opt} is needed to bias the laser diode in its optimum operating region. The part that is effectively used to describe the signal is called the average signal power P_{sig}. For a ZERO and a ONE having the same probability, it is defined as follows:

$$P_{sig} = \frac{P_h - P_l}{2}. \tag{2.10}$$

The *power penalty* (*PP*) relates the average power to the average signal power:

$$P_{opt} = P_{sig} \cdot \underbrace{\frac{r_e + 1}{r_e - 1}}_{PP} \tag{2.11}$$

$$P_{sig}\ [\text{dB}] = P_{opt}\ [\text{dB}] - PP\ [\text{dB}]. \tag{2.12}$$

The relation between the extinction ratio and the power penalty is presented in Fig. 2.8. The higher the extinction ratio, the more efficient the power is used because a smaller portion of the total power is used to bias the laser diode.

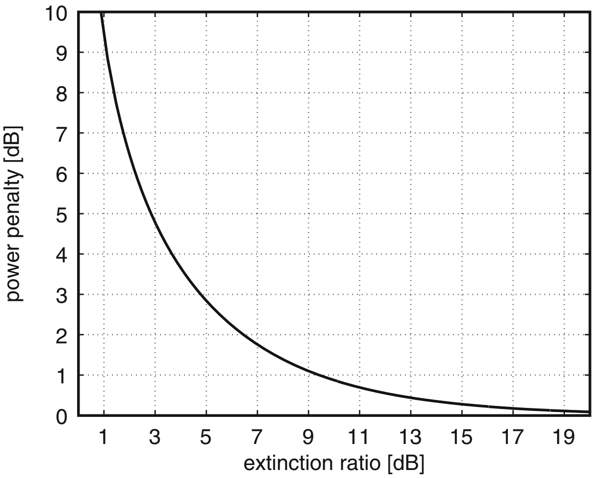


Fig. 2.8 The power penalty versus the extinction ratio

2.4 Optical Fiber

2.4.1 Silica Fiber

An optical fiber is used to guide the modulated light from the transmitter to the receiver. In Fig. 2.9, the typical structure of an optical fiber is shown. Three layers can be distinguished: *core*, *cladding* and *jacket*. The core guides the light while the cladding is needed to keep all the light inside the core. This principle is called *total internal reflection* and is fulfilled only when the angle of incidence is larger than a certain *critical angle*. The jacket protects the vulnerable structure of core and cladding. Usually, core and cladding are made from silica (SiO_2). To enable total internal reflection, the cladding should have a slightly lower *refractive index* than that of the core.

Two silica fiber types can be distinguished: *single-mode fiber* (SMF) and *multi-mode fiber* (MMF). The main structural difference between both is the diameter of the core which is between 8 and 10 µm for SMF and 50 or 62.5 µm for MMF. The cladding has a diameter of 125 µm in both fiber types. Due to the smaller core diameter of SMF, only one *mode*[6] can propagate through it. By contrast, several hundred modes can propagate in a MMF, each having a different path.

Loss A typical loss spectrum of silica fiber is plotted in Fig. 2.10. Two mechanisms can clearly be distinguished: *absorption* and *scattering*. Absorption is caused by the property of atoms and molecules to resonate at a specific frequency. In the infrared region[7] *ionic polarization* is the predominant effect, whereas in the ultraviolet region it is *electronic polarization* [Ina00]. The second type of losses, namely scattering losses, is caused by impurities in the core material. Scattered light waves leave the core if they do not fulfill the criterion of total internal reflection. This effect is called *Rayleigh scattering*. The scattering loss is proportional to

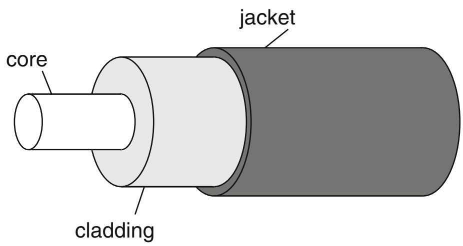
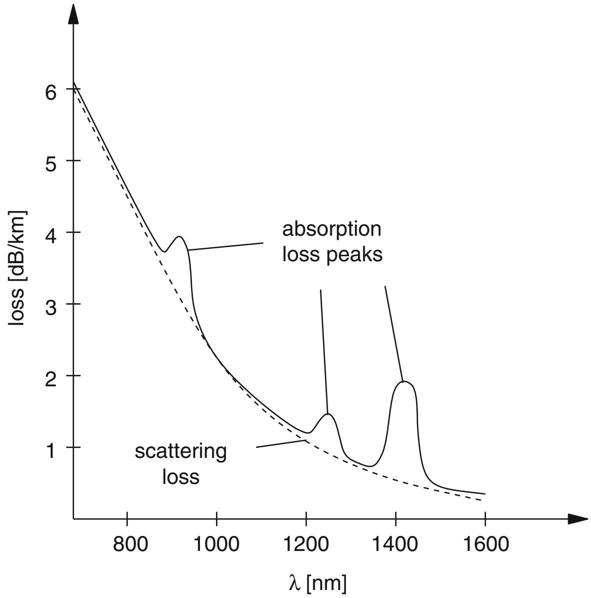


Fig. 2.9 Structure of an optical fiber

[6] There can be many modes traveling along a waveguide at a single frequency. Each mode has a different electric and magnetic field configuration. Ideally, there is no interaction between different modes [Olv92].

[7] Infrared radiation is electromagnetic radiation with a wavelength between 0.7 and 300 µm. Ultraviolet radiation is defined between 1 and 400 nm. Light with a wavelength in between the infrared and ultraviolet region is visible light.

Fig. 2.10 Typical loss spectrum of silica fiber



$1/\lambda^4$ [Raz03]. In order to encounter minimal loss, the wavelength should be approximately 1.5 μm. As will be discussed in Chap. 3, it is impossible to use a wavelength longer than 1 μm when the photodiode is integrated in silicon. Generally, because fiber loss increases linearly with the length of the fiber, it limits the distance between the transmitter and the receiver.

Dispersion Fiber dispersion also increases with the length of the fiber. However, it limits the distance between the transmitter and the receiver as well as the bit rate. There are three important dispersion mechanisms: *modal dispersion* (MD), *chromatic dispersion* (CHD) and *polarization-mode dispersion* (PMD).

The different paths in a MMF all have different path lengths. The direct path is the shortest way to the receiver and consequently arrives first, the wave that travels on the second shortest path arrives second and so on. This variation in the propagation delay between different modes causes the light pulses to spread out in the time domain which introduces *intersymbol interference* (ISI) and limits the bit rate. Due to the nature of a SMF that permits only a single mode in its core to propagate, this type of fiber is not subjected to MD.

CHD is a dispersion mechanism that affects both MMF and SMF. It is caused by a different propagation speed of light with different wavelengths (or colors). No CHD is present when light with a single wavelength is transmitted. However, every real laser diode emits a certain spectrum of wavelengths, though heavily concentrated in a very narrow band. On top of this, the modulation process of light with a single wavelength creates a spectrum in which a lot of different wavelengths are present. CHD is often quantified by the *dispersion parameter* D which is plotted in Fig. 2.11. Note that the minimal CHD is found at the band around a wavelength of 1300 nm.

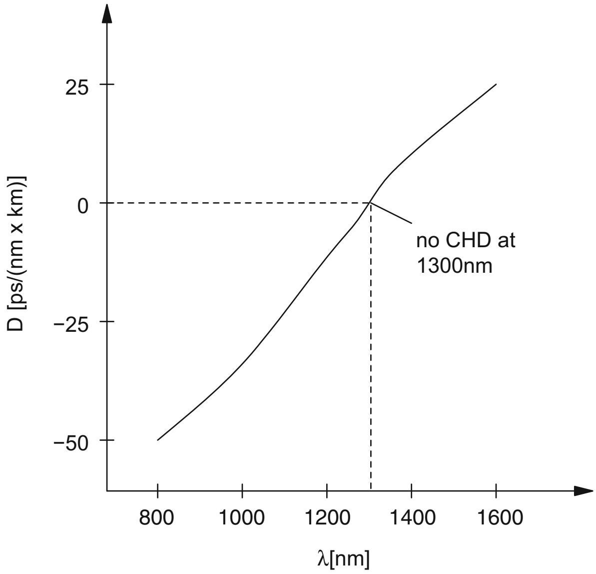


Fig. 2.11 Typical chromatic dispersion spectrum of silica fiber

To calculate the difference between the time of arrival of the fastest and the slowest wavelength for a transmitter with a spectral linewidth of $\Delta\lambda$ and for a fiber length L, the following formula can be used [Säc05]:

$$\Delta T = |D| \cdot \Delta\lambda \cdot L. \tag{2.13}$$

Because CHD is a linear phenomenon, its effect can be removed with the help of a *dispersion compensating fiber* (DCF).

Finally, PMD is caused by mechanical imperfections of the fiber such as a non-circular core or asymmetrical mechanical stress. This results in a difference in propagation speed between the horizontal and vertical polarization modes. Due to recent developments in fiber production, the delay between both modes can be kept below 1 ps for a fiber length of 100 km. Its effect can therefore be neglected in most cases.

2.4.2 Plastic Fiber

In contrast to the omnipresent silica fiber, *plastic optical fiber* (POF) is made out of plastic. It is the preferred choice in the growing market for consumer applications because an optical communication system making use of POF instead of silica fiber has a significantly lower cost thanks to the fiber, associated connectors and installation that are all inexpensive. This is a consequence of the large core diameter of POF, which is typically 1 mm. On top of this, POF has a high mechanical flexibility, which makes it an ideal solution to be used in cars (MOST[8]) or in houses (POF-PLUS[9]).

Compared to silica fiber, POF is characterized by a higher loss (up to 3 dB/m) which is caused by the less pure core material with relatively many impurities. Also, the bit rates that can be transported with POF are much lower than what can be obtained with a traditional fiber due to its higher dispersion. It is therefore predominantly applied in situations where the distance between transmitter and receiver is short and the needed bit rate is not extremely high.

Typically, the core material of POF is polymethyl methacrylate (PMMA) whereas fluoridated polymers are used in the cladding.

2.5 Optical Receiver Fundamentals

Due to fiber loss, the light signal at the receiver is greatly attenuated. Moreover, the signal is distorted owing to the dispersive character of the channel. The function of an optical receiver is to convert this attenuated and distorted light signal into digital data. The block diagram of such a receiver is shown in Fig. 2.12. The *photodiode* (Chap. 3) converts the received photons into an electrical current. In this work, the photodiode is integrated in the same silicon substrate that is also used to integrate the rest of the receiver chain. The *transimpedance amplifier* (TIA) (Chap. 4) amplifies the small photocurrent. The *equalizer* (Chap. 5) enhances the bandwidth of the receiver and more specifically that of the integrated photodiode. In the *post amplifier* (Chap. 6), the signal swing is enlarged to a rail-to-rail level. A *clock and*

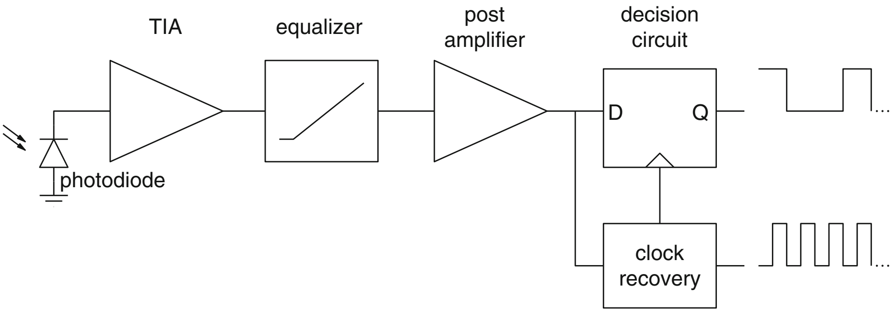


Fig. 2.12 Block diagram of an optical receiver; it provides the conversion of the attenuated modulated light signal at the receiver end of the optical fiber into a digital data stream with accompanying clock signal

[8] The media oriented systems transport (MOST) is a recent standard intended to connect multimedia devices with POF in cars.

[9] POF-PLUS is a European project that focuses on developing new photonic components and transmission technologies for large core POF systems, aiming at the unprecedented implementation of tens of Gbit/s transmission over this medium [POF10].

data recovery (CDR) system extracts the clock signal and retimes the data signal. The CDR is not treated in this work. Designing such a receiver implies making a lot of decisions. Therefore, a concise discussion about the most important trade-offs is given below.

2.5.1 Bandwidth Versus Bit Rate

The finite *bandwidth* of a system generally implies two effects: *low-pass filtering* and *high-pass filtering*. Whereas the former effect is always present in any real circuit implementation, the latter one can be absent, depending on the circuit topology. This is shown in Fig. 2.13 for the case where both effects are present. By definition, the bandwidth of a system is the frequency span within which the magnitude of the transfer function of the system is flat. The corner frequencies f_{hp} and f_{lp} are located where the magnitude is 3 dB lower than in the middle of the frequency band of interest. If $f_{hp} = 0$ Hz, the system is called a *baseband* system while it is called *passband* if $f_{hp} > 0$ Hz.

The receiver should be a baseband system if a truly random bit sequence is to be transmitted because this kind of signal has frequency components down to DC (Fig. 2.3b). On the contrary, when a PRBS is used as data pattern, the system can be passband if the corner frequency of the high-pass filtering is below the lowest frequency component in the PRBS (Fig. 2.5b)[10]. For a PRBS with a bit interval T_b and a maximum run length m, the condition is as follows:

$$f_{hp} < \frac{1}{(2^m - 1)T_b}. \tag{2.14}$$

For a PRBS with a maximum run length of 7 at a bit rate of 1 Gbit/s ($T_b = 1$ ns), f_{hp} should be below 7.81 MHz. However, even if condition (2.14) is fulfilled, *baseline wander* can be noticed. This is shown in Fig. 2.14a. The output level immediately follows the input level up to A but after course of time, the output level starts

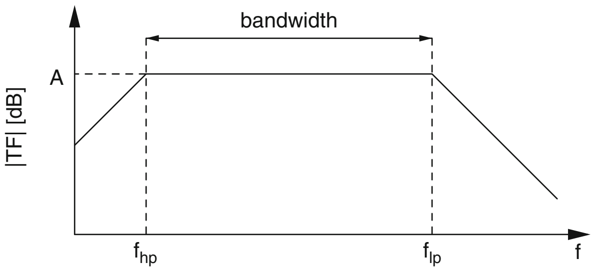


Fig. 2.13 Magnitude versus frequency of a system's transfer function on a logarithmic horizontal scale; the bandwidth is defined as the frequency span between f_{hp} and f_{lp}

[10] If the PRBS is not DC-balanced, an impulse at DC appears in the frequency domain (Fig. 2.5b). However, no information is contained in this impulse. Therefore, it can be removed without loss of signal integrity.

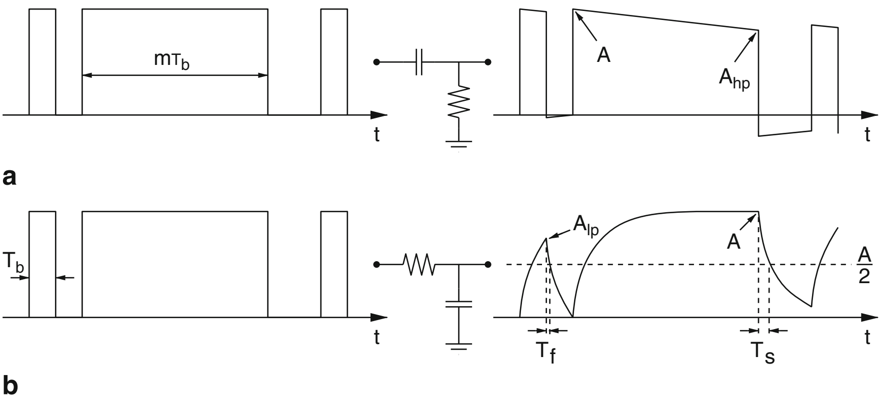


Fig. 2.14 Effect of a finite bandwidth on random digital data. **a** High-pass filtering. **b** Low-pass filtering

to droop significantly towards A_{hp}. The effect is most visible for the longest consecutive run of ONEs or ZEROs. If first-order high-pass filtering is assumed, the amount of baseline wander can be predicted as follows:

$$\frac{A_{hp}}{A} = e^{\frac{-mT_b}{\tau_{hp}}}, \tag{2.15}$$

where τ_{hp} is the time constant of the equivalent first-order high-pass filter. In the example from above, suppose $f_{hp} = 7.81$ MHz ($\tau_{hp} = 20.4$ ns). It follows from (2.15) that $A_{hp} = 0.710 \cdot A$. The effect of high-pass filtering can also be seen as ISI because the output value of a particular bit is dependent upon its predecessors.

In contrast to high-pass filtering which only appears for certain topologies, low-pass filtering is always present in any real circuit implementation. As can be seen in Figs. 2.3b and 2.5b, the spectrum of a random or pseudo-random data pattern contains components up to an infinite frequency. It is important to notice that those components decrease in magnitude with increasing frequency. According to (2.5), the rate at which this happens is proportional to $1/f^2$. In Fig. 2.14b the effect of low-pass filtering on random digital data is shown. Because the actual response in a single bit interval depends on the previous bits, this effect can again be seen as ISI. In contrast to high-pass filtering, however, the effect is worst for a ZERO or ONE following a long run of ONEs or ZEROs, respectively. In this case, the output level is only able to rise up to A_{lp} whereas it would reach A if enough time were available. If first-order low-pass filtering is assumed, the relation between both levels can be quantified as follows:

$$\frac{A_{lp}}{A} = 1 - e^{\frac{-T_b}{\tau_{lp}}}, \tag{2.16}$$

where τ_{lp} is the time constant of the equivalent first-order low-pass filter. For a PRBS at a bit rate of 1 Gbit/s ($T_b = 1$ ns) and a corner frequency $f_{lp} = 500$ MHz ($\tau_{lp} = 318$ ps), it follows from (2.16) that A_{lp} reaches A within 95.7%. Because the same effect manifests itself for a ONE to ZERO transition also, the actual vertical signal swing is only 91.4% of A.

A second effect of low-pass filtering, which can also be seen in Fig. 2.14b, is the variation of the threshold crossing dependent on the previous bits. This crossing can be fast, after T_f if a ZERO follows a single ONE or if a ONE follows a single ZERO. However, if a ZERO follows a long run of ONEs or if a ONE follows a long run of ZEROs, the crossing takes longer (T_s). This effect is known as *data-dependent jitter* (DDJ). The amount of DDJ can be determined with the following formula [Raz03]:

$$\frac{T_s - T_f}{T_b} = \frac{\tau_{lp}}{T_b} \cdot e^{\frac{-T_b}{\tau_{lp}}}. \tag{2.17}$$

The amount of DDJ can be calculated to be 1.38% of the bit interval for a PRBS at 1 Gbit/s ($T_b = 1$ ns) and a corner frequency $f_{lp} = 500$ MHz. To conclude, low-pass filtering affects the vertical (amplitude) scale as well as the horizontal (time) scale.

From now on, the term *bandwidth* is used as an alternative appellation for f_{lp} because in most practical implementations $f_{hp} \ll f_{lp}$.

2.5.2 Noise Versus Bandwidth

The objective of every receiver is to be able to process very small input signals. Therefore, the noise it generates should be sufficiently low. It is shown in Sect. 2.5.3 that a direct relation exists between the amount of noise generated by the receiver and the *bit error ratio* (BER) at the output of it. The quality of the output signal is thus not only affected by the finite bandwidth of the system (Sect. 2.5.2), but also by the amount of noise generated by it.

Every resistor, diode and transistor generates noise. The noise of a resistor is especially *white noise* which is generated by thermal agitation of charge carriers inside the resistor material. While white noise is dominant for a resistor, some *pink noise* is also present. The voltage noise spectral density of a resistor with resistance R is given by [San06]:

$$\overline{dV_{n,R}^2} = \underbrace{4kTR \cdot df}_{\text{white noise}} + \underbrace{V_R^2 \frac{K_R R_\square}{A_R} \cdot \frac{df}{f}}_{\text{pink noise}}, \tag{2.18}$$

where k is the *Boltzmann's constant*[11], T is the temperature of the resistor, V_R is the DC voltage over the resistor, K_R is a fitted parameter for a specific kind of resistor, $R_\square$ is the resistance per square[12] and A_R is the area of the resistor.

[11] Boltzmann's constant ($k = 1.38 \cdot 10^{-23}$ J/K) is a physical constant relating energy at the particle level with temperature at the bulk level.

[12] In integrated circuit technology, resistors are often described in terms of their resistance per square which is the resistance of a square of the resistor material. The resistance per square is not dependent on the size of the square. This is a convenient parameter in integrated circuits because it is essentially a two-dimensional environment.

A diode also generates both types of noise components. The white noise is in this case *shot noise* which is not dependent on temperature. The pink noise is again determined by the area of the device. The total current noise spectral density of a diode that conducts a current I_D is given by:

$$\overline{dI_{n,D}^2} = 2qI_D \cdot df + I_D \frac{K_D}{A_D} \cdot \frac{df}{f}, \tag{2.19}$$

where q is the elementary charge, K_D is a fitted parameter for the specific kind of diode and A_D is the area of the diode. The white and pink noise can again be distinguished clearly. In addition to this current noise, a diode also has a series resistance with its accompanying noise source. However, this noise contribution is not considered in the remainder of this book.

The predominant white noise component of a MOSFET is channel noise. Because the channel is resistive, the noise mechanism is thermal and consequently white. Other white (thermal) noise components in a MOSFET are generated by the gate resistance R_g, the source resistance R_s and the substrate resistance R_b. Because the current in a MOSFET flows close to the surface where a lot of defects in the crystal structure are present, a MOSFET also generates a considerable amount of pink noise. The gate referred noise voltage spectral density of a MOSFET is given by:

$$\overline{dV_{n,T}^2} = 4kTR_{\mathit{eff}} \cdot df + \frac{K_T}{WLC_{ox}^2} \cdot \frac{df}{f}, \tag{2.20}$$

where K_T represents a fitted parameter for a specific transistor type in a specific process technology, W and L are the gate width and length respectively and C_{ox} is the gate oxide capacitance per unit area. R_{eff} is the equivalent resistance that generates the same amount of white noise as the channel, gate, source and bulk together. Because all these noise sources are not correlated, the effective resistance can be calculated as follows [San06]:

$$R_{\mathit{eff}} = \frac{\gamma}{g_m} + R_g + R_s + R_b(n-1)^2, \tag{2.21}$$

where g_m is the transconductance of the transistor and the parameter n has a value between 1.2 and 1.5. The terms corresponding to the gate, source and bulk resistance can be distinguished clearly. The first term corresponds to the channel which has an equivalent noise resistance of γ/g_m. The parameter γ equals 2/3 for long-channel transistors but increases due to velocity saturation in short-channel devices. From now on, the parameter γ is supposed to be equal to 4/3, a value which is valid for a 130 nm CMOS technology [San06].

A circuit consists of a number of noise generating components which all contribute to the total output noise. To calculate this output noise, the transfer function of every noise source to the output needs to be determined. However, the total output noise voltage spectral density of a circuit depends on the configuration of the input as shown in Fig. 2.15. If the input is left open or is short-circuited, the respective output noise

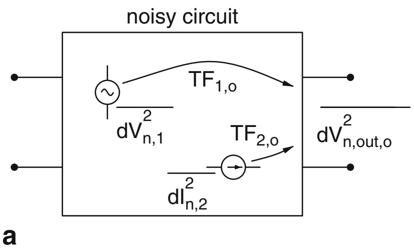


Fig. 2.15 The output noise of a circuit consisting of several internal noise sources. **a** The output noise if the input is left open. **b** The output noise when the input is short-circuited

voltage spectral densities $\overline{dV^2_{n,out,o}}$ and $\overline{dV^2_{n,out,s}}$ can be calculated with the following formulas:

$$\overline{dV^2_{n,out,o}} = \overline{dV^2_{n,1}} \cdot |TF_{1,o}|^2 + \overline{dI^2_{n,2}} \cdot |TF_{2,o}|^2 + \cdots, \tag{2.22}$$

$$\overline{dV^2_{n,out,s}} = \overline{dV^2_{n,1}} \cdot |TF_{1,s}|^2 + \overline{dI^2_{n,2}} \cdot |TF_{2,s}|^2 + \cdots, \tag{2.23}$$

where $TF_{n,o}$ and $TF_{n,s}$ denote the transfer function of the n$^{\text{th}}$ noise source to the output when the input is left open or is short-circuited respectively. Another way to represent the circuit noise is by its equivalent input noise. In this case, the circuit is assumed to be completely noiseless and the real output noise is assumed to be generated by an external voltage and current source as shown in Fig. 2.16 for a circuit where the input is connected to a source impedance Z_s. It can be proven that this can be done for every linear circuit [Raz01]. The equivalent input noise voltage and current spectral densities are calculated as follows:

$$\overline{dV^2_{n,in}} = \frac{\overline{dV^2_{n,out,s}}}{|TF_V|^2}, \tag{2.24}$$

$$\overline{dI^2_{n,in}} = \frac{\overline{dV^2_{n,out,o}}}{|TF_I|^2}, \tag{2.25}$$

where TF_V and TF_I represent the transfer function for an input voltage signal and for an input current signal respectively. Obviously, the input noise current should be

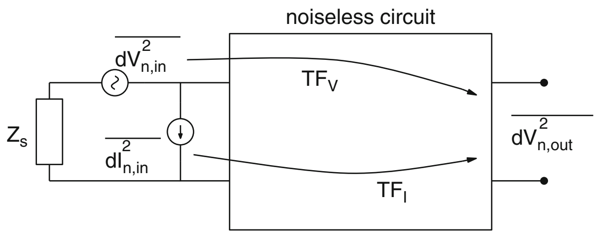


Fig. 2.16 The equivalent input noise of a circuit when the input is connected to a source impedance Z_s

calculated by taking into account $\overline{dV_{n,out,s}^2}$ alone because the voltage source at the input has no effect if the input is left open as in Fig. 2.15a. Equivalently, the input current noise is calculated by taking into account $\overline{dV_{n,out,o}^2}$ alone because the current source at the input has no effect if the input is short-circuited as in Fig. 2.15b. If the magnitude of Z_s in Fig. 2.16 has an infinite value, the voltage noise source $\overline{dV_{n,in}^2}$ has no effect while the noise current $\overline{dI_{n,in}^2}$ flows completely through the input nodes of the circuit, generating by doing so the total output noise. On the contrary, the current noise source is short-circuited if the magnitude of Z_s is zero. For a finite magnitude of Z_s, both the voltage and the current source generate part of the total output noise.

The noise calculations from above determine only the spectral densities of the noise. To obtain the *root mean square* (RMS) value of the output noise, the output noise spectral density of Fig. 2.16 needs to be integrated over the entire frequency spectrum:

$$V_{n,RMS} = \sqrt{\int_0^{\infty} \overline{dV_{n,out}^2}}. \tag{2.26}$$

This operation is visualized in Fig. 2.17 where the shaded area represents the squared integrated output noise voltage. Due to the physical limitations of every circuit implementation, every noise spectrum eventually drops to zero. Therefore, the integral of (2.26) always leads to a finite value[13].

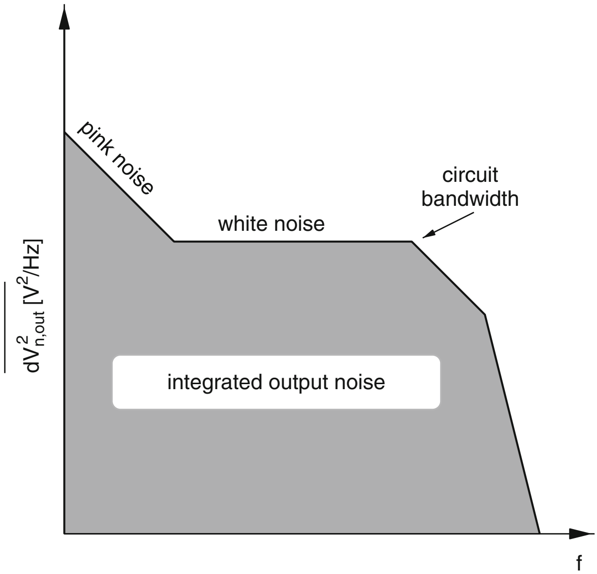


Fig. 2.17 Integration of the output noise voltage spectral density gives the total squared output noise voltage. Note that both the horizontal and the vertical axes have a logarithmic scale

[13] In the case of pink noise, the integral of (2.26) results in an infinitely high RMS value. Consequently, noise can only be pink over a limited range of frequencies, excluding 0 Hz.

It can be derived from (2.26) that the output noise voltage $V_{n,RMS}$ increases if the circuit bandwidth increases. To illustrate this, assume that a circuit contains only one resistor with resistance R_{noise} that generates only white noise with a noise spectral density of $\overline{dV^2_{n,R_{noise}}} = 4kTR_{noise}$. Assume also that the transfer function to the output for the noise of this resistor is a first-order low-pass filter with a time constant $\tau_{circuit}$. The RMS noise voltage at the output can be calculated as follows:

$$\begin{aligned} V_{n,RMS} &= \sqrt{\int_0^\infty \overline{dV^2_{n,R_{noise}}} \cdot \left| \frac{1}{1 + j2\pi f \tau_{circuit}} \right|^2} \\ &= \sqrt{4kTR_{noise} \int_0^\infty \frac{1}{1 + (2\pi f \tau_{circuit})^2} df} \\ &= \sqrt{\frac{kTR_{noise}}{\tau_{circuit}}}. \end{aligned} \tag{2.27}$$

If $R_{noise} = 50\ \Omega$ and $f_{circuit} = 1$ GHz ($\tau_{circuit} = 159$ ps), (2.27) results in an RMS output noise voltage of 36.1 μV_{RMS} at room temperature (300 K). If the bandwidth is increased to 2 GHz, the RMS output noise voltage rises with a factor of $\sqrt{2}$ to 51.0 μV_{RMS}.

In Sect. 2.5.1 it has been shown that the bandwidth should be as high as possible to minimize the degradation of the signal quality in terms of signal swing and DDJ. However, in this section it is found that a higher circuit bandwidth results in a larger RMS noise voltage, which in its turn degrades the quality of the output signal. While a high bandwidth seems advantageous if the input signal swing is large so that noise is of subordinate importance, a smaller bandwidth is better to cope with very weak input signals that are easily affected by circuit noise. In conclusion, the objective in choosing an appropriate receiver bandwidth is to make it as small as possible while still retaining enough vertical signal swing and repressing DDJ. In Sect. 2.5.1 it has been concluded that for a bit rate of 1 Gbit/s and a bandwidth of 500 MHz, the DDJ consumes only 1.38% of the bit interval and the vertical signal swing is 91.4% of that of a receiver with an infinite bandwidth. Therefore, a bandwidth of B/2 Hz seems sufficient not to introduce too much signal distortion for a data stream at a bit rate of B bit/s. In an actual implementation, the receiver is very likely to have additional poles above the circuit bandwidth. These poles increase the DDJ and reduce the signal swing further. A bandwidth of 0.6–0.7 $\cdot$ B Hz is therefore assumed to be optimal to receive data at a bit rate of B bit/s [Säc05].

2.5.3 Bit Error Ratio Versus Noise

As discussed in Sect. 2.5.2, the RMS noise voltage depends on the bandwidth of the system. However, it is not clear yet what the effect of this noise voltage is on the BER. To examine this, the circuit is assumed to have a bandwidth which is high enough not to distort the shape of the signal. Consequently, the signal transitions are performed infinitely fast, as visualized in Fig. 2.18. In the absence of noise, only two discrete values are possible for the signal amplitude: V_h for a ONE and V_l for a ZERO. This is shown in Fig. 2.18a where the data and the conditional *probability density function* (PDF) of its amplitude are plotted. When a ZERO or a ONE is received, the amplitude is always V_l or V_h respectively. The conditional chance of this event is therefore 1 in both cases. In the receiver, the decision whether a ZERO or a ONE is received is based upon the following condition:

$$\text{if } V \le V_t \Rightarrow \text{ZERO} \tag{2.28}$$

$$\text{if } V > V_t \Rightarrow \text{ONE.} \tag{2.29}$$

V_t is called the decision threshold voltage. In the absence of noise, there is no chance of making a wrong decision (2.28)–(2.29) if the decision threshold V_t is located anywhere between V_h and V_l. As a result, the corresponding BER is 0.

Things change quite drastically when noise is taken into account. This situation is visualized in Fig. 2.18b. The signal transitions still happen infinitely fast, so noise has no influence on the decision threshold crossing time. Between the signal transitions, the amplitude is now not well-defined any more because the noise voltage is added to the signal. To simplify the analysis, only white noise is considered here. The PDF

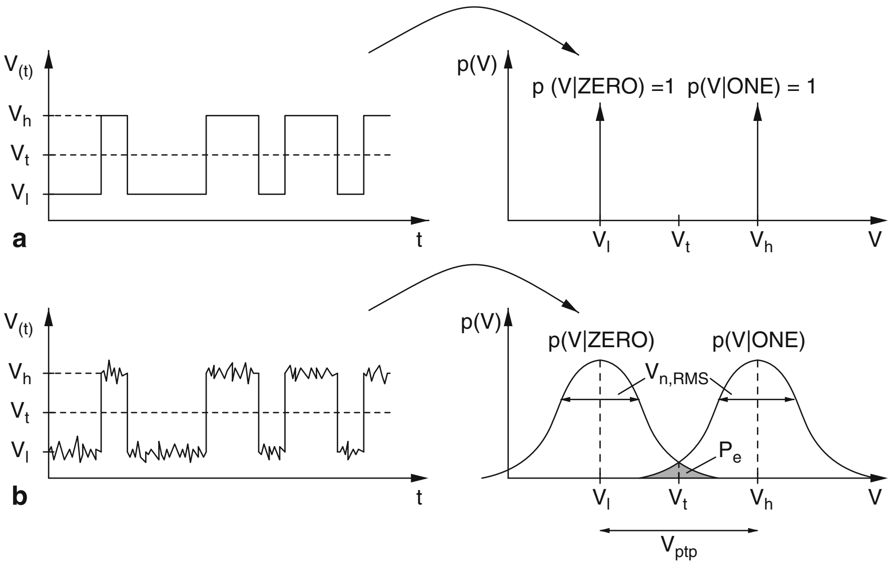


Fig. 2.18 The effect of noise on a random data pattern. **a** Data pattern and its corresponding PDF when no noise is present; the receiver always makes the correct decision if V_t is between V_h and V_l. **b** Data pattern and its corresponding PDF when noise is present; the receiver makes an errors with a probability P_e

of white noise is characterized by a *Gauss function* with an average of 0 and $V_{n,RMS}^2$ as variance:

$$p(V_n) = \frac{1}{2\pi V_{n,RMS}} e^{-V_n^2/2V_{n,RMS}^2}. \tag{2.30}$$

The simple PDF of the signal amplitude when noise is not taken into account Fig. 2.18a is now transformed into the PDF in Fig. 2.18b. When a ZERO is received, the PDF of the signal amplitude is $p(V|\text{ZERO})$ and $p(V|\text{ONE})$ when a ONE is received. These conditional distributions can be calculated respectively by means of the following formulas:

$$p(V|\text{ZERO}) = \frac{1}{2\pi V_{n,RMS}} e^{-(V_n - V_l)^2/2V_{n,RMS}^2} \tag{2.31}$$

$$p(V|\text{ONE}) = \frac{1}{2\pi V_{n,RMS}} e^{-(V_n - V_h)^2/2V_{n,RMS}^2}. \tag{2.32}$$

For the remainder of the derivation, it is assumed that the noise voltage $V_{n,RMS}$ does not depend on whether a ZERO or a ONE is received, which is a reasonable assumption in most cases[14]. In contrast to the situation without noise, there is always a possibility of making a wrong decision because the conditional PDFs only decrease asymptotically to 0. $P_{e,\text{ZERO}}$ is the probability that the receiver decides that a ONE has been transmitted while actually a ZERO is transmitted. $P_{e,\text{ONE}}$ is the probability that the receiver decides that a ZERO has been transmitted while a ONE is transmitted. Both probabilities can be expressed mathematically as follows:

$$P_{e,\text{ZERO}} = \int_{V_t}^{\infty} p(V|\text{ZERO})\, dV \tag{2.33}$$

$$P_{e,\text{ONE}} = \int_{-\infty}^{V_t} p(V|\text{ONE})\, dV. \tag{2.34}$$

For a ZERO and a ONE having an equal probability to be transmitted, the combined chance on a bit error is as follows:

$$P_e = \frac{1}{2} P_{e,\text{ZERO}} + \frac{1}{2} P_{e,\text{ONE}}. \tag{2.35}$$

This bit error probability is usually called the BER. It is also shown in Fig. 2.18b as the shaded area below the two conditional PDFs. To minimize P_e, the decision threshold must be chosen very carefully. If a ZERO and a ONE occur with equal

[14] When the receiver noise is dominated by the shot noise of the photodiode, the noise voltage for a ONE is larger than that of a ZERO because the shot noise is dependent on the current through the photodiode. This effect is neglected here.

probability and if the RMS noise voltage is equal for both logical levels, the decision threshold should be chosen exactly in the middle of V_l and V_h:

$$V_t = \frac{V_h + V_l}{2}. \tag{2.36}$$

With the help of these formulas, the BER can be written as [Cou02]:

$$P_e = Q\left(\frac{V_{ptp}}{2V_{n,RMS}}\right), \tag{2.37}$$

where V_{ptp} is the peak-to-peak signal amplitude $V_h - V_l$ (Fig. 2.18b). $Q(x)$ is called the *Q-function* and is the cumulative distribution function for the Gaussian distribution. The argument of the Q-function in (2.37) represents the *signal-to-noise-ratio* (SNR) of the signal. The graphical representation of (2.37) is given in Fig. 2.19 for a range of SNRs. As a rule of thumb, the BER decreases by two orders of magnitude for each decibel increase in SNR [Raz03]. For example, if the required BER is 10^{-12} or lower, the argument of the Q-function in (2.37) needs to be equal to or higher than 7. If the noise of the signal has an RMS value of 1 mV_{RMS}, the required peak-to-peak signal voltage needs to be larger than 14 mV (2.37)[15].

To quantify the BER as a function of the SNR (2.37), it is assumed that the bandwidth of the system is high enough in order not to introduce any signal distortion. However, as mentioned in Sect. 2.5.2, the bandwidth should be 0.6–0.7 $\cdot$ B Hz for a bit rate of B bit/s as an optimal trade-off between signal distortion and noise. As derived in Sect. 2.5.1, a finite bandwidth has two important consequences: DDJ and

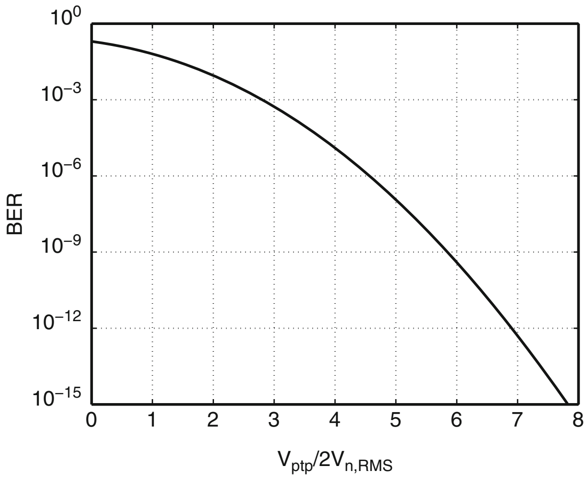


Fig. 2.19 Representation of the Q-function; with $V_{ptp}/2V_{n,RMS}$ as its argument the Q-function determines the BER of a signal corrupted by white noise

[15] The BER at the output of a receiver is determined by the signal quality at the input of this receiver as well as by the extra noise that is added by the receiver. In the remainder of this book, it is assumed that the former can always be neglected compared to the receiver noise.

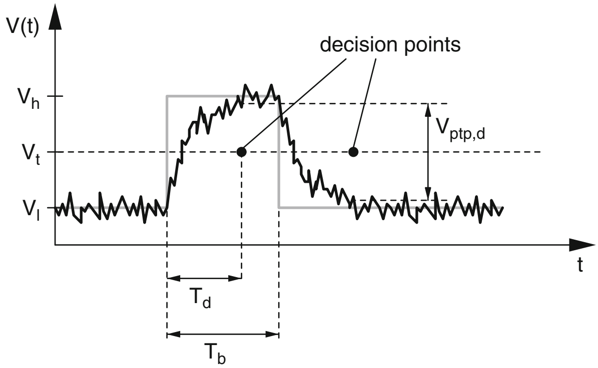


Fig. 2.20 The effect of noise and a limited bandwidth on a random data signal; the effective SNR decreases due to the finite bandwidth

a reduced signal swing. Some remarks are therefore needed to put things in a correct perspective. The effect of noise and a finite bandwidth on a random data pattern is shown in Fig. 2.20.

Without distortion, it is not important where exactly in the bit interval the decision is made as to whether a ZERO or a ONE has been transmitted. If the bandwidth is limited, this is not the case any more. DDJ and the reduced signal swing cause the optimal sampling point to be somewhere in the middle of the bit interval. This is especially important considering the fact that noise now also influences the rising and falling edges due to their finite rise and fall time. In addition to DDJ, an extra jitter component appears that is, in contrast to DDJ, random in nature because it is generated by noise. Therefore, it is called *random jitter* (RJ). A second effect of the limited bandwidth is a decreased vertical signal swing and from that a decreased SNR when compared to a signal with an infinite bandwidth. To be correct, the calculation of the BER should be done with the actual SNR. To quantify this, the vertical signal swing at the decision point, $V_{ptp,d}$, should be known. Its value is represented in Fig. 2.20. Obviously, $V_{ptp,d}$ is always smaller than V_{ptp}, the signal swing for a system with an infinite bandwidth. A more precise calculation of the BER can be obtained by altering (2.37) into the following formula:

$$P_e = Q\left(\frac{V_{ptp,d}}{2V_{n,RMS}}\right). \tag{2.38}$$

As described by (2.16), the vertical signal swing at the end of the bit interval is reduced because of a limited bandwidth. However, to calculate the BER, the signal swing at the decision point is important. Therefore, (2.16) is transformed into the following formula to calculate the effective signal swing at the decision point:

$$\frac{V_{ptp,d}}{V_{ptp}} = 1 - 2e^{\frac{-T_d}{\tau_{lp}}}. \tag{2.39}$$

A factor 2 is added because the swing is limited for a ZERO to ONE as well as for a ONE to ZERO transition. For a PRBS at a bit rate of 1 Gbit/s ($T_b = 1$ ns) and a corner frequency (bandwidth) $f_{lp} = 500$ MHz ($\tau_{lp} = 318$ ps), it follows from (2.39) that $V_{ptp,d} = 0.914 \cdot V_{ptp}$ if the decision point is at the end of the bit interval ($T_d = T_b$). To obtain a BER below 10^{-12}, it has been derived before that a peak-to-peak signal swing of 14 mV is sufficient if the RMS noise voltage equals 1 $\mathrm{mV_{RMS}}$. However, the BER increases to $3.14 \cdot 10^{-11}$ (2.38–2.39) when the finite bandwidth is taken into account. If noise generates too much RJ, it is necessary to sample before the end of the bit interval. If, for example, the decision point is moved forward so that $T_d = 0.75 \cdot T_b$, it follows from (2.39) that $V_{ptp,d} = 0.811 \cdot V_{ptp}$. The BER then equals $2.85 \cdot 10^{-9}$. Clearly, both the integrated noise as well as the bandwidth determine the BER of the receiver.

Sensitivity Another way to quantify the performance of an optical receiver can be obtained by considering its *sensitivity* P_s. This is defined as the minimal average optical input power that is needed to achieve a certain BER. To determine the sensitivity, the conversion from the light signal to the electrical output signal is described as follows:

$$V_{ptp} = P_{ptp} \cdot R_{pd} \cdot Z_r, \tag{2.40}$$

where R_{pd}[A/W] is the *responsivity* of the photodiode and $Z_r[\Omega]$ represents the transimpedance gain of the receiver. The same can be done for the RMS noise:

$$V_{n,RMS} = P_{n,RMS} \cdot R_{pd} \cdot Z_r. \tag{2.41}$$

$P_{n,RMS}$ can be considered as the equivalent integrated input noise of the receiver. The BER can now be calculated also at the input by transforming (2.37) into:

$$P_e = Q\left(\frac{P_{ptp}}{2P_{n,RMS}}\right). \tag{2.42}$$

If the sensitivity is defined at a BER of 10^{-12} for which the argument of the Q-function needs to be 7 or higher, P_s equals the peak-to-peak signal swing that is needed to obtain this: $P_s = 7 \cdot 2P_{n,RMS}$. Taking into account the finite extinction ratio of the laser diode, the total optical input power to achieve a certain BER can be calculated as follows:

$$P_{opt} = P_s + PP, \tag{2.43}$$

where PP is the power penalty due to the finite extinction ratio of the laser diode.

2.6 Conclusion

An optical communication system is used to transmit random digital data. At the transmitter, a laser diode converts the signal from the electrical domain to the optical domain. Due to the fact that the current through a laser diode should always be larger than a certain threshold current, the average transmitted power always exceeds the average signal power. The optical signal is guided to the receiver by means of an optical fiber. Mostly, a silica fiber is used because of its low loss and dispersion. However, in some specific application domains plastic fiber is applied which is characterized by a relatively high loss and dispersion but, on the contrary, also by a very low cost price.

The optical receiver, which is the topic of this thesis, is charged with the conversion of the attenuated and distorted optical signal into a truly digital bit stream that can be interpreted by the digital back-end. One of the key parameters of an optical receiver is its bandwidth which needs to be high enough to prevent the signal from being distorted. This distortion has two important consequences: a reduced signal swing and the appearance of DDJ. On the contrary, the bandwidth should also be as low as possible to guarantee low noise operation. Eventually, the most important parameter of an optical receiver is its BER. This probability on a bit error is a strong function of the ratio of the peak-to-peak signal swing and the RMS noise. Both the peak-to-peak signal swing and the integrated noise depend on the bandwidth of the receiver. An optimal trade-off between a good low-noise performance and a low distortion is obtained when, for a bit rate of B bit/s, the bandwidth equals $0.6–0.7 \cdot B$ Hz. In conclusion, the BER performance combines the effect of a finite bandwidth and the RMS noise. The sensitivity determines the required optical input power to obtain a certain BER.

http://www.springer.com/978-1-4419-9924-5

High-Speed Optical Receivers with Integrated
Photodiode in Nanoscale CMOS
Tavernier, F.; Steyaert, M.
2011, XIV, 226 p., Hardcover
ISBN: 978-1-4419-9924-5