Optical Communication Networks

EE654

Lecture -3

Spring 2016

3. Modulation and Demodulation

OUR GOAL IN THIS CHAPTER is to understand the processes of modulation and demodulation of digital signals. We start by discussing *modulation,* which is the process of converting digital data in electronic form to an optical signal that can be transmitted over the fiber. We then study the *demodulation* process, which is the process of converting the optical signal back into electronic form and extracting the data that was transmitted.

3.1 Modulation

The most commonly used modulation scheme in optical communication is *on-off keying* (OOK), which is illustrated in Figure 4.1. In this modulation scheme, a 1 bit is

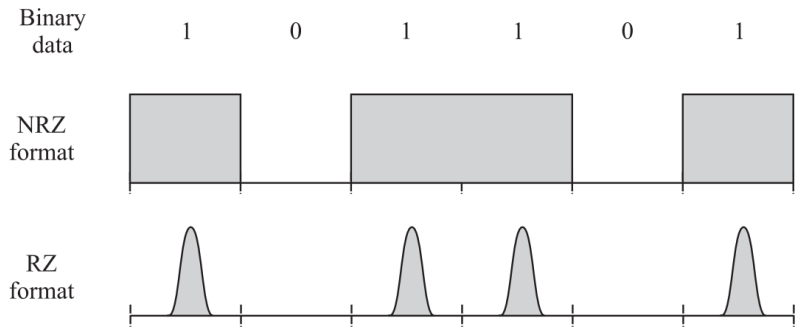

Figure 4.1 On-off keying modulation of binary digital data.

encoded by the presence of a light pulse in the bit interval or by turning a light source (laser or LED) "on." A 0 bit is encoded (ideally) by the absence of a light pulse in the bit interval or by turning a light source "off." The bit interval is the interval of time available for the transmission of a single bit. For example, at a bit rate of 1 Gb/s, the bit interval is 1 ns. As we saw in Section 3.5.4, we can either *directly* modulate the light source by turning it on or off, or use an external modulator in front of the source to perform the same function. Using an external modulator results in less chirp, and thus less of a penalty due to dispersion, and is the preferred approach for high-speed transmission over long distances.

3.1.1 signal Formats

The OOK modulation scheme can use many different signal formats. The most common signal formats are non-return-to-zero (NRZ) and return-to-zero (RZ). These formats are illustrated in Figure 4.1. In the NRZ format, the pulse for a 1 bit occupies the entire bit interval, and no pulse is used for a 0 bit. If there are two successive

1s, the pulse occupies two successive bit intervals. In the RZ format, the pulse for a 1 bit occupies only a fraction of the bit interval, and no pulse is used for a 0 bit. In electronic (digital) communication, the RZ format has meant that the pulse occupies exactly half the bit period. However, in optical communication, the term *RZ* is used in a broader sense to describe the use of pulses of duration shorter than the bit period. Thus, there are several variations of the RZ format. In some of them, the pulse occupies a substantial fraction (say, 30%) of the bit interval. The term *RZ*, without any qualification, usually refers to such systems. If, in addition, the pulses are chirped, they are also sometimes termed dispersion-managed (DM) solitons. In other RZ systems, the pulse occupies only a small fraction of the bit interval. The primary example of such a system is a (conventional) soliton system.

The major advantage of the NRZ format over the other formats is that the signal occupies a much smaller bandwidth—about half that of the RZ format. The problem with the NRZ format is that long strings of 1s or 0s will result in a total absence of any transitions, making it difficult for the receiver to acquire the bit clock, a problem we discuss in Section 4.4.8. The RZ format ameliorates this problem somewhat since long strings of 1s (but not strings of 0s) will still produce transitions. However, the RZ format requires a higher peak transmit power in order to maintain the same energy per bit, and hence the same bit error rate as the NRZ format.

A problem with all these formats is the lack of *DC balance*. An OOK modulation scheme is said to have DC balance if, for all sequences of data bits that may have to be transmitted, the average transmitted power is constant. It is important for an OOK modulation scheme to achieve DC balance because this makes it easier to set the decision threshold at the receiver (see Section 5.2).

To ensure sufficient transitions in the signal and to provide DC balance, either *line coding* or *scrambling* is used in the system. There are many different types of line codes. One form of a *binary block line code* encodes a block of k data bits into $n > k$ bits that are then modulated and sent over the fiber. At the receiver, the n bits are mapped back into the original k data bits (assuming there were no errors). Line codes can be designed so that the encoded bit sequence is DC balanced and provides sufficient transitions regardless of the input data bit sequence. An example of such a line code is the (8, 10) code that is used in the Fibre Channel standard [WF83, SV96]. This code has $k = 8$ and $n = 10$. The fiber distributed data interface (FDDI) [Ros86] uses a (4, 5) code that is significantly less complex than this (8, 10) code but does not quite achieve DC balance; the worst-case DC imbalance is 10% [Bur86].

An alternative to using line coding is to use *scrambling*. Scrambling is a one-to-one mapping of the data stream into another data stream before it is transmitted on the link. At the transmitter, a scrambler takes the incoming bits and does an EXOR operation with another carefully chosen sequence of bits. The latter sequence is chosen so as to minimize the likelihood of long sequences of 1s or 0s in the transmitted stream. The data is recovered back at the receiver by a descrambler that extracts the data from the scrambled stream. The advantage of scrambling over line coding is that it does not require any additional bandwidth. The disadvantages are that it does not guarantee DC balance, nor does it guarantee a maximum length for a sequence

of 1s or 0s. However, the probability of having long run lengths or DC imbalance is made very small by choosing the mapping so that likely input sequences with long run lengths are mapped into sequences with a small run length. However, since the mapping is one to one, it is possible to choose an input sequence that results in a bad output sequence. The mapping is chosen so that only very rare input sequences produce bad output sequences. See Problem 4.2 for an example of how scrambling is implemented and its properties.

In practice, the NRZ format is used in most high-speed communication systems, ranging from speeds of 155 Mb/s to 10 Gb/s. Scrambling is widespread and used in most communication equipment ranging from PC modems to high-speed telecommunications links. High-speed computer data links (for example, Fibre Channel, which operates at 800 Mb/s, and Gigabit Ethernet, which operates at 1 Gb/s) use line codes. See Chapter 6 for a discussion of these protocols.

The RZ format is used in certain high-bit-rate communication systems, such as chirped RZ or DM soliton systems (see Section 2.6.1). In these systems, the pulse occupies about half the bit interval, though this is usually not precise as in digital/electronic communication. The use of RZ pulses also minimizes the effects of chromatic dispersion (see Section 5.7.2). RZ modulation with pulses substantially shorter than the bit interval is used in soliton communication systems (see Section 2.6). The pulses need to be very short in such systems because they must be widely separated (by about five times their width) in order to realize the dispersion-free propagation properties of solitons.

3.1.2 Subcarrier Modulation and Multiplexing

The optical signal emitted by a laser operating in the 1310 or 1550 nm wavelength band has a center frequency around 10^{14} Hz. This frequency is the *optical carrier* frequency. In what we have studied so far, the data modulates this optical carrier. In other words, with an OOK signal, the optical carrier is simply turned on or off, depending on the bit to be transmitted.

Instead of modulating the optical carrier directly, we can have the data first modulate an electrical carrier in the microwave frequency range, typically ranging from 10 MHz to 10 GHz, as shown in Figure 4.2. The upper limit on the carrier frequency is determined by the modulation bandwidth available from the transmitter. The modulated microwave carrier then modulates the optical transmitter. If the transmitter is directly modulated, then changes in the microwave carrier amplitude get reflected as changes in the transmitted optical power envelope, as shown in Figure 4.2. The microwave carrier can itself be modulated in many different ways, including amplitude, phase, and frequency modulation, and both digital and analog modulation techniques can be employed. The figure shows an example where the microwave carrier is amplitude modulated by a binary digital data signal. The microwave carrier is called the *subcarrier,* with the optical carrier being considered the main carrier. This form of modulation is called *subcarrier modulation.*

The main motivation for using subcarrier modulation is to multiplex multiple data streams onto a single optical signal. This can be done by combining multiple microwave carriers at different frequencies and modulating the optical transmitter

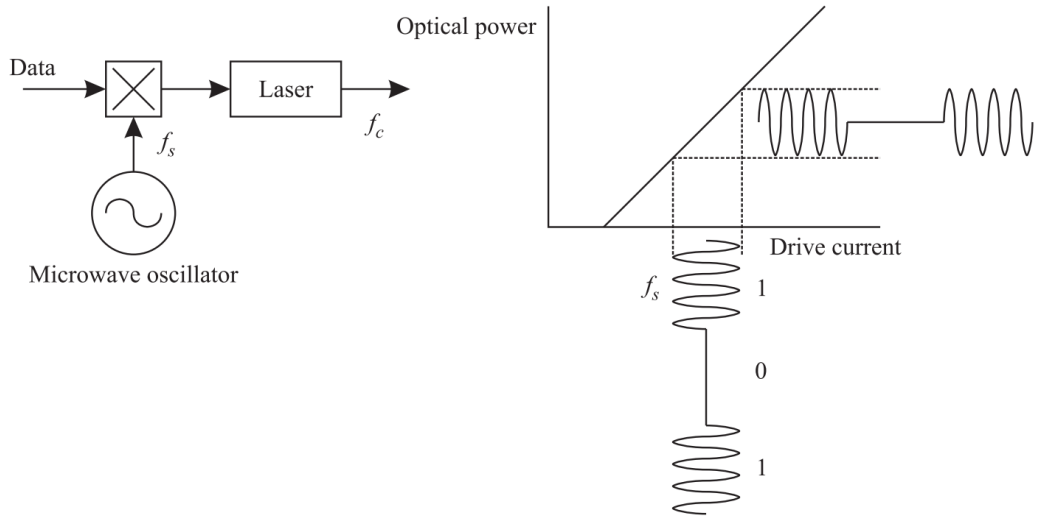

Figure 4.2 Subcarrier modulation. The data stream first modulates a microwave carrier, which, in turn, modulates the optical carrier.

with the combined signal. At the receiver, the signal is detected like any other signal, and the rest of the processing, to separate the subcarriers and extract the data from each subcarrier, is done electronically. This form of multiplexing is called *subcarrier multiplexing* (SCM).

3.1.3 Application of SCM

SCM is widely used by cable operators today for transmitting multiple analog video signals using a single optical transmitter. SCM is also being used in metropolitan-area networks to combine the signals from various users using electronic FDM followed by SCM. This reduces the cost of the network since each user does not require an optical transmitter/laser. We will study these applications further in Chapter 11.

SCM is also used to combine a control data stream along with the actual data stream. For example, most WDM systems that are deployed carry some control information about each WDM channel along with the data that is being sent. This control information has a low rate and modulates a microwave carrier that lies above the data signal bandwidth. This modulated microwave carrier is called a *pilot tone*. We will discuss the use of pilot tones in Chapter 8.

Often it is necessary to receive the pilot tones from all the WDM channels for monitoring purposes, but not the data. This can be easily done if the pilot tones use different microwave frequencies. If this is the case, and the combined WDM signal is photodetected, the detector output will contain an electronic FDM signal consisting of all the pilot tones from which the control information can be extracted. The information from all the data channels will overlap with one another and be lost.

3.2 Spectral Efficiency

We saw in Chapter 2 that the ultimate bandwidth available in silica optical fiber is about 400 nm from 1.2 μm to 1.6 μm, or about 50 THz. The natural question that arises is, therefore, what is the total capacity at which signals can be transmitted over optical fiber?

There are a few different ways to look at this question. The *spectral efficiency* of a digital signal is defined as the ratio of the bit rate to the bandwidth used by the signal. The spectral efficiency depends on the type of modulation and coding scheme used. Today's systems primarily use on-off keying of digital data and in theory can achieve a spectral efficiency of 1 b/s/Hz. In practice, the spectral efficiency of these systems is more like 0.4 b/s/Hz. Using this number, we see that the maximum capacity of optical fiber is about 20 Tb/s. The spectral efficiency can be improved by using more sophisticated modulation and coding schemes, leading to higher channel capacities than the number above. As spectral efficiency becomes increasingly important, such new schemes are being invented, typically based on proven electrical counterparts.

One such scheme that we discuss in the next section is *optical duobinary modulation*. It can increase the spectral efficiency by a factor of about 1.5, typically, achieving a spectral efficiency of 0.6 b/s/Hz.

3.2.1 Optical Duobinary Modulation

The fundamental idea of duobinary modulation (electrical or optical) is to deliberately introduce intersymbol interference (ISI) by overlapping data from adjacent bits. This is accomplished by adding a data sequence to a 1-bit delayed version of itself. For example, if the (input) data sequence is $(0, 0, 1, 0, 1, 0, 0, 1, 1, 0)$, we would instead transmit the (output) data sequence $(0, 0, 1, 0, 1, 0, 0, 1, 1, 0) + (*, 0, 0, 1, 0, 1, 0, 0, 1, 1) = (0, 0, 1, 1, 1, 1, 0, 1, 2, 1)$. Here the $*$ denotes the initial value of the input sequence, which we assume to be zero.

Note that while the input sequence is binary and consists of 0s and 1s, the output sequence is a ternary sequence consisting of 0s, 1s, and 2s. Mathematically, if we denote the input sequence by $x(nT)$ and the output sequence by $y(nT)$, duobinary modulation results if

$$y(nT) = x(nT) + x(nT - T),$$

where T is the bit period. In the example above, $x(nT) = (0, 0, 1, 0, 1, 0, 0, 1, 1, 0)$, $1 \leq n \leq 10$, and $y(nT) = (0, 0, 1, 1, 1, 1, 0, 1, 2, 1)$, $1 \leq n \leq 10$.

Since the bits overlap with each other, how do we recover the input sequence $x(nT)$ at the receiver from $y(nT)$? This can be done by constructing the signal $z(nT) = y(nT) - z(nT - T)$ at the receiver. Note that here we subtract a delayed version of $z(nT)$ from $y(nT)$, and not a delayed version of $y(nT)$ itself. This operation recovers $x(nT)$ since $z(nT) = x(nT)$, assuming we also initialize the sequence $z(0) = 0$. (For readers familiar with digital filters, $y(nT)$ is obtained from $x(nT)$ by a digital filter, and $z(nT)$ from $y(nT)$ by using the inverse of the same digital filter.) The reader should verify this by calculating $z(nT)$ for the example sequence above. To see that this holds generally, just calculate as follows:

$$\begin{aligned} z(nT) &= y(nT) - z(nT - T) \\ &= y(nT) - y(nT - T) + z(nT - 2T) \\ &= y(nT) - y(nT - T) + y(nT - 2T) - z(nT - 3T) \\ &= y(nT) - y(nT - T) + y(nT - 2T) - \ldots + (-1)^{n-1} y(T) \\ &= [x(nT) + x(nT - T)] - [x(nT - T) - x(nT - 2T)] + \ldots \\ &= x(nT) \end{aligned} \tag{4.1}$$

There is one problem with this scheme, however; a single transmission error will cause all further bits to be in error, until another transmission error occurs to correct the first one! This phenomenon is known as *error propagation*. To visualize error propagation, assume a transmission error occurs in some ternary digit in the example sequence $y(nT)$ above, and calculate the decoded sequence $z(nT)$.

The solution to the error propagation problem is to encode the actual data to be transmitted, not by the absolute value of the input sequence $x(nT)$, but by changes in the sequence $x(nT)$. Thus the sequence $x(nT) = (0, 0, 1, 0, 1, 0, 0, 1, 1, 0)$ would correspond to the data sequence $d(nT) = (0, 0, 1, 1, 1, 1, 0, 1, 0, 1)$. A 1 in the sequence $d(nT)$ is encoded by changing the sequence $x(nT)$ from a 0 to a 1, or from a 1 to a 0. To see how differential encoding solves the problem, observe that if a sequence of consecutive bits are all in error, their differences will still be correct, modulo 2.

Transmission of a ternary sequence using optical intensity modulation (the generalization of OOK for nonbinary sequences) will involve transmitting three different optical powers, say, 0, P, and $2P$. Such a modulation scheme will also considerably complicate the demodulation process. We would like to retain the advantage of binary signaling while employing duobinary signaling to reduce the transmission bandwidth.

To see how this can be done, compare $y(nT) = (0, 0, 1, 1, 1, 1, 0, 1, 2, 1)$ and $d(nT) = (0, 0, 1, 1, 1, 1, 0, 1, 0, 1)$ in our example, and observe that $y(nT) \bmod 2 = d(nT)$! This result holds in general, and thus we may think that we could simply map the 2s in $y(nT)$ to 0s and transmit the resulting binary sequence, which could then be detected using the standard scheme. However, such an approach would eliminate the bandwidth advantage of duobinary signaling, as it should, because in such a scheme the differential encoding and the duobinary encoding have done nothing but cancel each other's effects. The bandwidth advantage of duobinary signaling can only be exploited by using a ternary signaling scheme. A ternary signaling alternative to using three optical power levels is to use a combination of amplitude and phase modulation. Such a scheme is dubbed optical AM-PSK, and most studies of optical duobinary signaling today are based on AM-PSK.

Conceptually, the carrier is a continuous wave signal, a sinusoid, which we can denote by $a\cos(\omega t)$. The three levels of the ternary signal correspond to $-a\cos(\omega t) = a\cos(\omega t + \pi)$, $0 = 0\cos(\omega t)$, and $a\cos(\omega t)$, which we denote by -1, 0, and $+1$, respectively. The actual modulation is usually accomplished using an external modulator in the Mach-Zehnder arrangement (see Sections 3.3.7 and 3.5.4). These are the three signal levels corresponding to 0, 1, and 2, respectively, in $y(nT)$. This modulation scheme is clearly a combination of amplitude and phase modulation, hence the term AM-PSK. The AM-PSK signal retains the bandwidth advantage of duobinary signaling. However, for a direct detection receiver, the signals

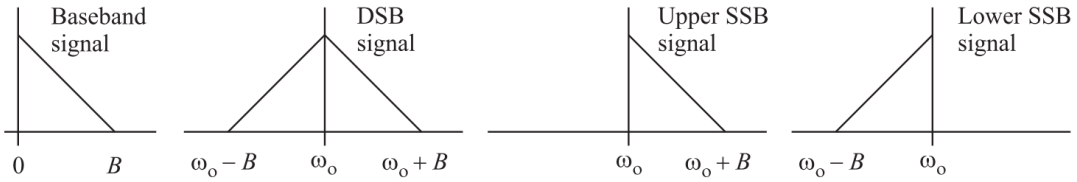

Figure 4.4 Spectrum of a baseband signal compared with the spectra of double sideband (DSB) and single sideband (SSB) modulated signals. The spectral width of the SSB signals is the same as that of the baseband signal, whereas the DSB signal has twice the spectral width of the baseband signal.

$\pm a\cos(\omega t)$ are indistinguishable so that the use of such a receiver merely identifies $2 = 0$ in $y(nT)$ naturally performing the mod 2 operation required to recover $d(nT)$ from $y(nT)$.

3.2.2 Capacity Limits of Optical Fiber

An upper limit on the spectral efficiency and the channel capacity is given by Shannon's theorem [Sha48]. Shannon's theorem says that the channel capacity C for a binary linear channel with additive noise is given by

$$C = B \log_2 \left(1 + \frac{S}{N}\right).$$

Here B is the available bandwidth and S/N is the signal-to-noise ratio. A typical value of S/N is 100. Using this number yields a channel capacity of 350 Tb/s or an equivalent spectral efficiency of 7 b/s/Hz. Clearly, such efficiencies can only be achieved through the use of multilevel modulation schemes.

In practice, today's long-haul systems operate at high power levels to overcome fiber losses and noise introduced by optical amplifiers. At these power levels, nonlinear effects come into play. These nonlinear effects can be thought of as adding additional noise, which increases as the transmitted power is increased. Therefore they in

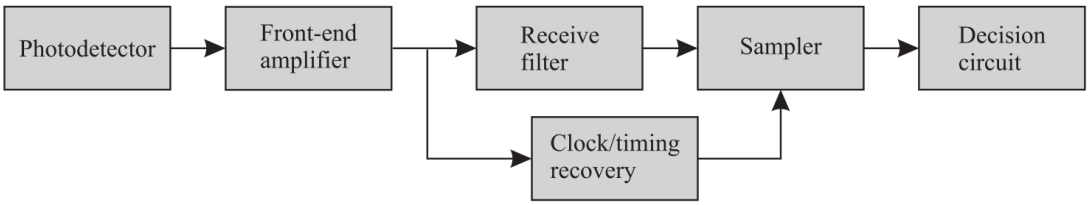

Figure 4.5 Block diagram showing the various functions involved in a receiver.

turn impose additional limits on channel capacity. Recent work to quantify the spectral efficiency, taking into account mostly cross-phase modulation [Sta99, MS00], shows that the achievable efficiencies are of the order of 3–5 b/s/Hz. Other nonlinearities such as four-wave mixing and Raman scattering may place further limitations. At the same time, we are seeing techniques to reduce the effects of these nonlinearities.

Another way to increase the channel capacity is by reducing the noise level in the system. The noise figure in today's amplifiers is limited primarily by random spontaneous emission, and these are already close to theoretically achievable limits. Advances in quantum mechanics [Gla00] may ultimately succeed in reducing these noise limits.

3.3 Demodulation

The modulated signals are transmitted over the optical fiber where they undergo attenuation and dispersion, have noise added to them from optical amplifiers, and sustain a variety of other impairments that we will discuss in Chapter 5. At the receiver, the transmitted data must be recovered with an acceptable *bit error rate* (BER). The required BER for high-speed optical communication systems today is in the range of 10^{-9} to 10^{-15}, with a typical value of 10^{-12}. A BER of 10^{-12} corresponds to one allowed bit error for every terabit of data transmitted, on average.

Recovering the transmitted data involves a number of steps, which we will discuss in this section. Our focus will be on the demodulation of OOK signals. Figure 4.5 shows the block diagram of a receiver. The optical signal is first converted to an electrical current by a *photodetector.* This electrical current is quite weak and thus we use a *front-end amplifier* to amplify it. The photodetector and front-end amplifier were discussed in Sections 3.6.1 and 3.6.2, respectively.

The amplified electrical current is then filtered to minimize the noise outside the bandwidth occupied by the signal. This filter is also designed to suitably shape the pulses so that the bit error rate is minimized. This filter may also incorporate

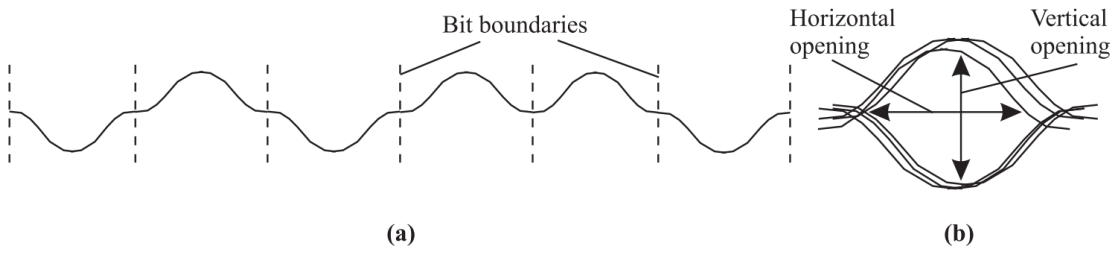

Figure 4.6 Eye diagram. (a) A typical received waveform along with the bit boundaries. (b) The received waveform of (a), wrapped around itself, on the bit boundaries to generate an eye diagram. For clarity, the waveform has been magnified by a factor of 2 relative to (a).

additional functionality, such as minimizing the intersymbol interference due to pulse spreading. If the filter performs this function, it is termed an *equalizer.* The name denotes that the filter equalizes, or cancels, the distortion suffered by the signal. Equalization is discussed in Section 4.4.9.

The signal must then be sampled at the midpoints of the bit intervals to decide whether the transmitted bit in each bit interval was a 1 or a 0. This requires that the bit boundaries be recovered at the receiver. A waveform that is periodic with period equal to the bit interval is called a *clock*. This function is termed *clock recovery,* or *timing recovery,* and is discussed in Section 4.4.8.

A widely used experimental technique to determine the goodness of the received signal is the *eye diagram*. Consider the received waveform shown in Figure 4.6(a). This is a typical shape of the received signal for NRZ modulation, after it has been filtered by the receive filter and is about to be sampled (see Figure 4.5). The bit boundaries are also shown on the figure. If the waveform is cut along at the bit boundaries and the resulting pieces are superimposed on each other, we get the resulting diagram shown in Figure 4.6(b). Such a diagram is called an *eye diagram* because of its resemblance to the shape of the human eye. An eye diagram can be easily generated experimentally using an oscilloscope to display the received signal while it is being triggered by the (recovered) clock. The vertical opening of the eye indicates the margin for bit errors due to noise. The horizontal opening of the eye indicates the margin for timing errors due to an imperfectly recovered clock.

In Section 1.5, we saw that there could be different types of repeaters, specifically 2R (regeneration with reshaping) and 3R (regeneration with reshaping and retiming). The difference between these lies primarily in the type of receiver used. A 2R receiver does not have the timing recovery circuit shown in Figure 4.5, whereas a 3R does. Also a 3R receiver may use a multirate timing recovery circuit, which is capable of recovering the clock at a variety of data rates.

3.3.1 An ideal Receiver

In principle, the demodulation process can be quite simple. Ideally, it can be viewed as "photon counting," which is the viewpoint we will take in this section. In practice, there are various impairments that are not accounted for by this model, and we discuss them in the next section.

The receiver looks for the presence or absence of light during a bit interval. If no light is seen, it infers that a 0 bit was transmitted, and if any light is seen, it infers that a 1 bit was transmitted. This is called *direct detection*. Unfortunately, even in the absence of other forms of noise, this will not lead to an ideal error-free system because of the random nature of photon arrivals at the receiver. A light signal arriving with power P can be thought of as a stream of photons arriving at average rate P/hf_c. Here, h is Planck's constant (6.63×10^{-34} J/Hz), f_c is the carrier frequency, and hf_c is the energy of a single photon. This stream can be thought of as a Poisson random process.

Note that our simple receiver does not make any errors when a 0 bit is transmitted. However, when a 1 bit is transmitted, the receiver may decide that a 0 bit was transmitted if no photons were received during that bit interval. If B denotes the bit rate, then the probability that n photons are received during a bit interval $1/B$ is given by

$$e^{-(P/hf_cB)}\frac{\left(\frac{P}{hf_cB}\right)^n}{n!}.$$

Thus the probability of not receiving any photons is $e^{-(P/hf_cB)}$. Assuming equally likely 1s and 0s, the bit error rate of this ideal receiver would be given as

$$\text{BER} = \frac{1}{2}e^{-\frac{P}{hf_cB}}.$$

Let $M = P/hf_cB$. The parameter M represents the average number of photons received during a 1 bit. Then the bit error rate can be expressed as

$$\text{BER} = \frac{1}{2}e^{-M}.$$

This expression represents the error rate of an ideal receiver and is called the *quantum limit*. To get a bit error rate of 10^{-12}, note that we would need an average of $M = 27$ photons per 1 bit.

In practice, most receivers are not ideal, and their performance is not as good as that of the ideal receiver because they must contend with various other forms of noise, as we shall soon see.

3.3.2 APractical Direct Detection Receiver

As we have seen in Section 3.6 (see Figure 3.61), the optical signal at the receiver is first photodetected to convert it into an electrical current. The main complication in recovering the transmitted bit is that in addition to the photocurrent due to the signal there are usually three other additional noise currents. The first is the *thermal noise* current due to the random motion of electrons that is always present at any finite temperature. The second is the *shot noise* current due to the random distribution of the electrons generated by the photodetection process even when the input light intensity is constant. The shot noise current, unlike the thermal noise current, is not added to the generated photocurrent but is merely a convenient representation of the variability in the generated photocurrent as a separate component. The third source of noise is the spontaneous emission due to optical amplifiers that may be used between the source and the photodetector. The amplifier noise currents are treated in Section 4.4.5 and Appendix I. In this section, we will consider only the thermal noise and shot noise currents.

The thermal noise current in a resistor R at temperature T can be modeled as a Gaussian random process with zero mean and autocorrelation function $(4k_BT/R)\delta(\tau)$. Here k_B is Boltzmann's constant and has the value 1.38×10^{-23} J/°K, and $\delta(\tau)$ is the Dirac delta function, defined as $\delta(\tau) = 0, \tau \neq 0$ and $\int_{-\infty}^{\infty} \delta(\tau)d\tau = 1$. Thus the noise is white, and in a bandwidth or frequency range B_e, the thermal noise current has the variance

$$\sigma^2_{\text{thermal}} = (4k_BT/R)B_e.$$

This value can be expressed as $I_t^2 B_e$, where I_t is the parameter used to specify the current standard deviation in units of pA/$\sqrt{\text{Hz}}$. Typical values are of the order of 1 pA/$\sqrt{\text{Hz}}$.

The electrical bandwidth of the receiver, B_e, is chosen based on the bit rate of the signal. In practice, B_e varies from $1/2T$ to $1/T$, where T is the bit period. We will also be using the parameter B_o to denote the optical bandwidth seen by the receiver. The optical bandwidth of the receiver itself is very large, but the value of B_o is usually determined by filters placed in the optical path between the transmitter and receiver. By convention, we will measure B_e in baseband units and B_o in passband units. Therefore, the minimum possible value of $B_o = 2B_e$, to prevent signal distortion.

As we saw in the previous section, the photon arrivals are accurately modeled by a Poisson random process. The photocurrent can thus be modeled as a stream

of electronic charge impulses, each generated whenever a photon arrives at the photodetector. For signal powers that are usually encountered in optical communication systems, the photocurrent can be modeled as

$$I = \bar{I} + i_s,$$

where $\bar{I}$ is a constant current, and i_s is a Gaussian random process with mean zero and autocorrelation $\sigma^2_{\text{shot}}\delta(\tau)$. For *pin* diodes, $\sigma^2_{\text{shot}} = 2e\bar{I}$. This is derived in Appendix I. The constant current $\bar{I} = \mathcal{R}P$, where $\mathcal{R}$ is the responsivity of the photodetector, which was discussed in Section 3.6. Here, we are assuming that the dark current, which is the photocurrent that is present in the absence of an input optical signal, is negligible. Thus the shot noise current is also white and in a bandwidth B_e has the variance

$$\sigma^2_{\text{shot}} = 2e\bar{I}B_e. \tag{4.2}$$

If we denote the load resistor of the photodetector by R_L, the total current in this resistor can be written as

$$I = \bar{I} + i_s + i_t,$$

where i_t has the variance $\sigma^2_{\text{thermal}} = (4k_BT/R_L)B_e$. The shot noise and thermal noise currents are assumed to be independent so that, if B_e is the bandwidth of the receiver, this current can be modeled as a Gaussian random process with mean $\bar{I}$ and variance

$$\sigma^2 = \sigma^2_{\text{shot}} + \sigma^2_{\text{thermal}}.$$

Note that both the shot noise and thermal noise variances are proportional to the bandwidth B_e of the receiver. Thus there is a trade-off between the bandwidth of a receiver and its noise performance. A receiver is usually designed so as to have just sufficient bandwidth to accommodate the desired bit rate so that its noise performance is optimized. In most practical direct detection receivers, the variance of the thermal noise component is much larger than the variance of the shot noise and determines the performance of the receiver.

3.3.3 Front-End Amplifier Noise

We saw in Chapter 3 (Figure 3.61) that the photodetector is followed by a front-end amplifier. Components within the front-end amplifier, such as the transistor, also contribute to the thermal noise. This noise contribution is usually stated by giving the *noise figure* of the front-end amplifier. The noise figure F_n is the ratio of the input signal-to-noise ratio (SNR_i) to the output signal-to-noise ratio (SNR_o). Equivalently,

the noise figure F_n of a front-end amplifier specifies the factor by which the thermal noise present at the input of the amplifier is enhanced at its output. Thus the thermal noise contribution of the receiver has variance

$$\sigma^2_{\text{thermal}} = \frac{4k_BT}{R_L} F_n B_e \tag{4.3}$$

when the front-end amplifier noise contribution is included. Typical values of F_n are 3–5 dB.

3.3.4 APD Noise

As we remarked in Section 3.6.1, the avalanche gain process in avalanche photodiodes has the effect of increasing the noise current at its output. This increased noise contribution arises from the random nature of the avalanche multiplicative gain, $G_m(t)$. This noise contribution is modeled as an increase in the shot noise component at the output of the photodetector. If we denote the responsivity of the APD by $\mathcal{R}_{\text{APD}}$, and the average avalanche multiplication gain by G_m, the average photocurrent is given by $\bar{I} = \mathcal{R}_{\text{APD}}P = G_m\mathcal{R}P$, and the shot noise current at the APD output has variance

$$\sigma^2_{\text{shot}} = 2eG_m^2 F_A(G_m)\mathcal{R}PB_e. \tag{4.4}$$

The quantity $F_A(G_m)$ is called the *excess noise factor* of the APD and is an increasing function of the gain G_m. It is given by

$$F_A(G_m) = k_A G_m + (1 - k_A)(2 - 1/G_m).$$

The quantity k_A is called the ionization coefficient ratio and is a property of the semiconductor material used to make up the APD. It takes values in the range (0–1). The excess noise factor is an increasing function of k_A, and thus it is desirable to keep k_A small. The value of k_A for silicon (which is used at 0.8 μm wavelength) is $\ll 1$, and for InGaAs (which is used at 1.3 and 1.55 μm wavelength bands) it is 0.7.

Note that $F_A(1) = 1$, and thus (4.4) also yields the shot noise variance for a *pin* receiver if we set $G_m = 1$.

3.3.5 Optical Preamplifiers

As we have seen in the previous sections, the performance of simple direct detection receivers is limited primarily by thermal noise generated inside the receiver. The performance can be improved significantly by using an optical (pre)amplifier after the receiver, as shown in Figure 4.7. The amplifier provides added gain to the input

signal. Unfortunately, as we saw in Section 3.4.2, the spontaneous emission present in the amplifier appears as noise at its output. The amplified spontaneous (ASE) noise power at the output of the amplifier for each polarization mode is given by

$$P_N = n_{\text{sp}} h f_c (G - 1) B_o, \tag{4.5}$$

where n_{sp} is a constant called the spontaneous emission factor, G is the amplifier gain, and B_o is the optical bandwidth. Two fundamental polarization modes are present in a single-mode fiber, as we saw in Chapter 2. Hence the total noise power at the output of the amplifier is $2P_N$.

The value of n_{sp} depends on the level of population inversion within the amplifier. With complete inversion $n_{\text{sp}} = 1$, but it is typically higher, around 2–5 for most amplifiers.

For convenience in the discussions to follow, we define

$$P_n = n_{\text{sp}} h f_c.$$

To understand the impact of amplifier noise on the detection of the received signal, consider the optical preamplifier system shown in Figure 4.7, used in front of a standard *pin* direct detection receiver. The photodetector produces a current that is proportional to the incident power. The signal current is given by

$$I = \mathcal{R}GP, \tag{4.6}$$

where P is the received optical power.

The photodetector produces a current that is proportional to the optical power. The optical power is proportional to the square of the electric field. Thus the noise field beats against the signal and against itself, giving rise to noise components referred to as the *signal-spontaneous* beat noise and *spontaneous-spontaneous* beat noise, respectively. In addition, shot noise and thermal noise components are also present.

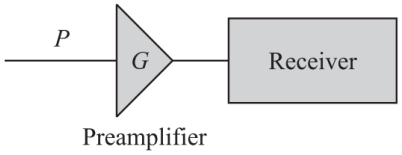

Figure 4.7 A receiver with an optical preamplifier.

3.3.6 Bit Error Rates

Earlier, we calculated the bit error rate of an ideal direct detection receiver. Next, we will calculate the bit error rate of the practical receivers already considered, which must deal with a variety of different noise impairments.

The receiver makes decisions as to which bit (0 or 1) was transmitted in each bit interval by sampling the photocurrent. Because of the presence of noise currents, the receiver could make a wrong decision resulting in an erroneous bit. In order to compute this bit error rate, we must understand the process by which the receiver makes a decision regarding the transmitted bit.

First, consider a *pin* receiver without an optical preamplifier. For a transmitted 1 bit, let the received optical power $P = P_1$, and let the mean photocurrent $\bar{I} = I_1$. Then $I_1 = \mathcal{R}P_1$, and the variance of the photocurrent is

$$\sigma_1^2 = 2eI_1B_e + 4k_BTB_e/R_L.$$

If P_0 and I_0 are the corresponding quantities for a 0 bit, $I_0 = \mathcal{R}P_0$, and the variance of the photocurrent is

$$\sigma_0^2 = 2eI_0B_e + 4k_BTB_e/R_L.$$

For ideal OOK, P_0 and I_0 are zero, but we will see later (Section 5.3) that this is not always the case in practice.

Note that it is particularly important to have a variable threshold setting in receivers if they must operate in systems with signal-dependent noise, such as optical amplifier noise. Many high-speed receivers do incorporate such a feature. However, many of the simpler receivers do not have a variable threshold adjustment and set their threshold corresponding to the average received current level, namely, $(I_1 + I_0)/2$. This threshold setting yields a higher bit error rate given by

$$\text{BER} = \frac{1}{2}\left[Q\left(\frac{(I_1 - I_0)}{2\sigma_1}\right) + Q\left(\frac{(I_1 - I_0)}{2\sigma_0}\right)\right].$$

3.3.7 Coherent Detection

We saw earlier that simple direct detection receivers are limited by thermal noise and do not achieve the shot noise limited sensitivities of ideal receivers. We saw that the sensitivity could be improved significantly by using an optical preamplifier. Another way to improve the receiver sensitivity is to use a technique called *coherent detection*.

The key idea behind coherent detection is to provide gain to the signal by mixing it with another local light signal from a so-called local-oscillator laser. At the same time, the dominant noise in the receiver becomes the shot noise due to the local oscillator, allowing the receiver to achieve the shot noise limited sensitivity. (In fact, a radio receiver works very much in this fashion except that it operates at radio, rather than light, frequencies.)

A simple coherent receiver is shown in Figure 4.10. The incoming light signal is mixed with a local-oscillator signal via a 3 dB coupler and sent to the photodetector. (We will ignore the 3 dB splitting loss induced by the coupler since it can be eliminated by a slightly different receiver design—see Problem 4.15.) Assume that the phase and polarization of the two waves are perfectly matched. The power seen by the photodetector is then

$$\begin{aligned} P_r(t) &= \left[\sqrt{2aP}\cos(2\pi f_c t) + \sqrt{2P_{\text{LO}}}\cos(2\pi f_{\text{LO}} t)\right]^2 \\ &= aP + P_{\text{LO}} + 2\sqrt{aPP_{\text{LO}}}\cos[2\pi(f_c - f_{\text{LO}})t]. \end{aligned} \tag{4.20}$$

Here, P denotes the input signal power, P_{LO} the local-oscillator power, $a = 1$ or 0 depending on whether a 1 or 0 bit is transmitted (for an OOK signal), and f_c and

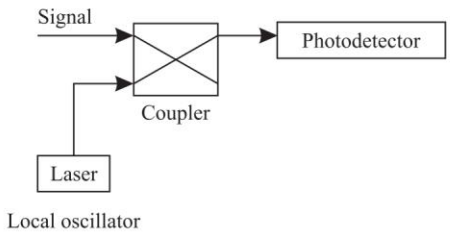

Figure 4.10 A simple coherent receiver.

f_{LO} represent the carrier frequencies of the signal and local-oscillator waves. We have neglected the $2f_c$, $2f_{LO}$, and $f_c + f_{LO}$ components since they will be filtered out by the receiver. In a *homodyne* receiver, $f_c = f_{LO}$, and in a *heterodyne* receiver, $f_c - f_{LO} = f_{IF} \neq 0$. Here, f_{IF} is called the intermediate frequency (IF), typically a few gigahertz.

To illustrate why coherent detection yields improved receiver sensitivities, consider the case of a homodyne receiver. For a 1 bit, we have

$$I_1 = \mathcal{R}(P + P_{LO} + 2\sqrt{P P_{LO}}),$$

and for a 0 bit,

$$I_0 = \mathcal{R} P_{LO}.$$

3.3.8 Timing Recovery

The process of determining the bit boundaries is called *timing recovery*. The first step is to extract the clock from the received signal. Recall that the clock is a periodic waveform whose period is the bit interval (Section 4.4). This clock is sometimes sent separately by the transmitter, for example, in a different frequency band. Usually, however, the clock must be extracted from the received signal. Even if the extracted clock has a period equal to the bit interval, it may still be out of phase with the received signal; that is, the clock may be offset from the bit boundaries. Usually, both the clock frequency (periodicity) and its phase are recovered simultaneously by a single circuit, as shown in Figure 4.11.

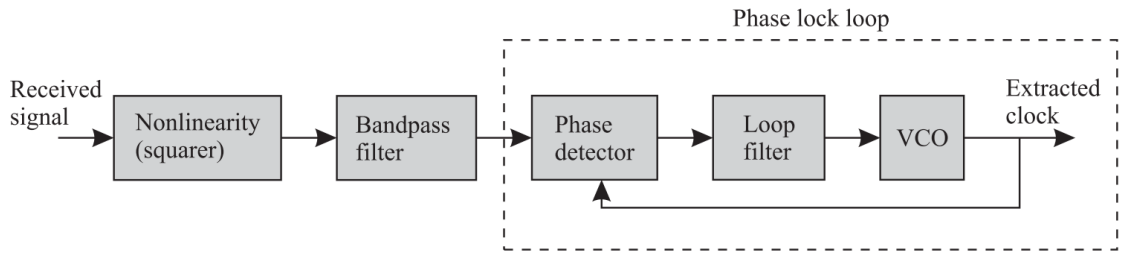

Figure 4.11 Block diagram illustrating timing, or clock, recovery at the receiver.

If we pass the received signal through a nonlinearity, typically some circuit that calculates the square of the received signal, it can be shown that the result contains a spectral component at $1/T$, where T is the bit period. Thus, we can filter the result using a bandpass filter as shown in Figure 4.11 to get a waveform that is approximately periodic with period T and that we call a *timing signal*. However, this waveform will still have considerable *jitter;* that is, successive "periods" will have slightly different durations. A "clean" clock with low jitter can be obtained by using the *phase lock loop* (PLL) circuit shown in Figure 4.11.

A PLL consists of a voltage-controlled oscillator (VCO), a phase detector, and a loop filter. A VCO is an oscillator whose output frequency can be controlled by an input voltage. A *phase detector* produces an error signal that depends on the difference in phase between its two inputs. Thus, if the timing signal and the output of the VCO are input to the phase detector, it produces an error signal that is used to adjust the output of the VCO to match the (average) frequency and phase of the timing signal. When this adjustment is complete, the output of the VCO serves as the clock that is used to sample the filtered signal in order to decide upon the values of the transmitted bits. The *loop filter* shown in Figure 4.11 is a critical element of a PLL and determines the residual jitter in the output of the VCO, as well as the ability of the PLL to track changes in the frequency and phase of the timing signal.

3.3.9 Equalization

We remarked in Section 4.4 with reference to Figure 4.5 that the receive filter that is used just prior to sampling the signal can incorporate an *equalization filter* to cancel the effects of intersymbol interference due to pulse spreading. From the viewpoint of the electrical signal that has been received, the entire optical system (including the laser, the fiber, and the photodetector) constitutes the *channel* over which the signal has been transmitted. If nonlinearities are ignored, the main distortion caused by this channel is the dispersion-induced broadening of the (electrical) pulse. Dispersion is

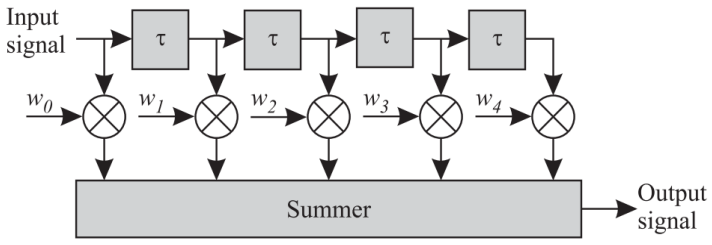

Figure 4.12 A transversal filter, a commonly used structure for equalization. The output (equalized) signal is obtained by adding together suitably delayed versions of the input signal, with appropriate weights.

a linear effect, and hence the effect of the channel on the pulse, due to dispersion, can be modeled by the response of a filter with transfer function $H_D(f)$. Hence, in principle, by using the inverse of this filter, say, $H_D^{-1}(f)$, as the equalization filter, this effect can be canceled completely at the receiver. This is what an equalization filter attempts to accomplish.

The effect of an equalization filter is very similar to the effect of dispersion compensating fiber (DCF). The only difference is that in the case of DCF, the equalization is in the optical domain, whereas equalization is done electrically when using an equalization filter. As in the case of DCF, the equalization filter depends not only on the type of fiber used but also on the fiber length.

A commonly used filter structure for equalization is shown in Figure 4.12. This filter structure is called a *transversal filter.* It is essentially a tapped delay line: the signal is delayed by various amounts and added together with individual weights. The choice of the weights, together with the delays, determines the transfer function of the equalization filter. The weights of the tapped delay line have to be adjusted to provide the best possible cancellation of the dispersion-induced pulse broadening.

Electronic equalization involves a significant amount of processing that is difficult to do at higher bit rates, such as 10 Gb/s. Thus optical techniques for dispersion compensation, such as the use of DCF for chromatic dispersion compensation, are currently much more widely used compared to electronic equalization.

3.5 Error Detection and Correction

An error-correcting code is a technique for reducing the bit error rate on a communicaton channel. It involves transmitting additional bits, called *redundancy,* along with the data bits. These additional bits carry redundant information and are used by the receiver to correct most of the errors in the data bits. This method of reducing the error rate by having the transmitter send redundant bits (using an error-correcting code) is called *forward error correction* (FEC).

An alternative is for the transmitter to use a smaller amount of redundancy, which the receiver can use to detect the presence of an error, but there is insufficient redundancy to identify/correct the errors. This approach is used in telecommunication systems based on SONET and SDH to monitor the bit error rate in the received signal. It is also widely used in data communication systems, where the receiver requests the transmitter to resend the data blocks that are detected to be in error. This technique is called *automatic repeat request* (ARQ).

A simple example of an *error-detecting code* is the *bit interleaved parity* (BIP) code. A BIP-N code adds N additional bits to the transmitted data. We can use either *even* or *odd* parity. With a BIP-N of even parity, the transmitter computes the code as follows: The first bit of the code provides even parity over the first bit of all N-bit sequences in the covered portion of the signal, the second bit provides even parity over the second bits of all N-bit sequences within the specified portion, and so on. Even parity is generated by setting the BIP-N bits so that there are an even number of 1s in each of all N-bit sequences, including the BIP-N bit. Problem 4.16 provides more details on this code.

A type of error-detecting code that is widely used in data communications is the *cyclic redundancy check* (CRC). A CRC code is based on a computation that resembles long division. The "divisor" of this computation is a bit string called a "generator polynomial." The generator polynomial actually defines the particular CRC code, and some of these polynomials are industry standards.

A CRC code forms a codeword from a data string by adding redundant bits so that the codeword is "divisible" by the generator polynomial. If a transmitted codeword is not divisible, then there was a bit error in the transmission. CRC codes can be designed to detect single bit errors, double bit errors, odd number of bit errors, and any burst of errors that has length less than the length of the generator polynomial.

FEC codes are more powerful than error-detecting codes because they can correct bit errors, which reduces the bit error rate (BER). This is especially important for optical communication systems that are expected to operate at a very low residual BER: 10^{-12} or lower. Now FEC is not necessary when there are low demands on the communication system due to relaxed channel spacing, negligible component crosstalk, negligible effect of nonlinearities, and so on. Then all that is required to achieve the specified BER is to increase the received power. However, in very high-capacity WDM systems FEC becomes necessary.

The simplest error-correcting code is a *repetition code*. In such a code, every bit is repeated some number of times, say, three times. For example, a 1 is transmitted as 111 and a 0 as 000. Thus we have one data, or information bit, plus two redundant bits of the same value. The receiver can estimate the data bit based on the value of the majority of the three received bits. For example, the received bits 101 are interpreted to mean that the data bit is a 1, and the received bits 100 are interpreted to mean the data bit is a 0.

It is easy to see how the use of such a code improves the BER, if the same energy is transmitted per bit after coding, as in the uncoded system. This amounts to transmitting three times the power in the above example, since three coded bits have to be transmitted for every data bit. In this case, the coded system has the same raw BER—the BER before error correction or decoding—as the uncoded system. However, after decoding, at least two bits in a block of three bits have to be in error for the coded system to make a wrong decision. This substantially decreases the BER of the coded system, as discussed in Problem 4.17. For example, the BER decreases from 10^{-6} for the uncoded system to 3×10^{-12} in the coded system.