Chapter 2
Fundamentals of Optical Communication

The ultimate goal of the optical signal transmission is to achieve the predetermined bit-error ratio (BER) between any two nodes in an optical network. The optical transmission system has to be properly designed in order to provide the reliable operation during its lifetime, which includes the management of key engineering parameters.

In this chapter, we describe the key optical components used in contemporary optical communication systems; basic direct detection modulation schemes; and basic coherent detection schemes. This chapter is based on [1–24].

The chapter is organized as follows. After the brief introduction, in Sect. 2.2 the key optical components are identified and described at a level sufficient to understand other chapters in the book without having any background in optical communications. In Sect. 2.3, different modulation formats with direct detection are described. Section 2.4 is devoted to different modulation schemes with coherent detection.

2.1 Introduction

The optical transmission system design [1–5] involves accounting for different effects that may degrade the signal during modulation, propagation, and detection processes. The transmission quality is assessed by the received signal-to-noise ratio (SNR), which is the ratio between signal power and noise power at the decision point. The SNR is related to the receiver sensitivity, the minimum received optical power needed to keep SNR at the specified level.

In digital optical communications, bit-error rate (BER), defined as the ratio of bits in error to total number of transmitted bit at the decision point, is commonly used as a figure of merit. In that sense, the receiver sensitivity is defined as the minimum required received optical power to keep BER below a given value. The three types of parameters important from the system engineering point of view include (1) optical signal parameters that determine the signal level, (2) the optical noise parameters that determine the BER, and (3) the impairment parameters that determine the power margin to be allocated to compensate for their impact. The optical
signal parameters defining the signal level include optical transmitter output power, extinction ratio, optical amplification gain, and photodiode responsivity. The total noise is a stochastic process composed of both additive noise components and multiplicative (nonadditive) noise components. There exist a number of impairments that deteriorate the signal quality during transmission such as fiber attenuation, chromatic dispersion, polarization mode dispersion (PMD), polarization-dependent loss (PDL), fiber nonlinearities, insertion loss, and frequency chirp; just to mention a few. A proper design process involves different steps to provide a prespecified transmission system quality and to balance different system parameters. The systems parameter can be related to power, time, wavelength, or combination of them. Given this general description of different signal and noise parameters, we turn our attention to the key optical components.

2.2 Key Optical Components

This section describes the basic optical components used in an optical transmission system. An exemplary optical network identifying the key optical components is shown in Fig. 2.1. The end-to-end optical transmission involves both electrical and optical signal paths. To perform conversion from electrical to optical domain, the optical transmitters are used, while to perform conversion in opposite direction (optical to electrical conversion), the optical receivers are used. The optical fibers serve as foundation of an optical transmission system because the optical fiber is used as a medium to transport the optical signals from source to destination. The optical fibers attenuate the signal during transmission, and someone has to use optical amplifiers, such as erbium-doped fiber amplifiers (EDFAs), Raman amplifiers, or parametric amplifiers, to restore the signal quality [1–5]. However, the process of amplification is accompanied with noise addition. The simplest optical transmission system employs only one wavelength. The wavelength division multiplexing (WDM) can

![Fig. 2.1 An exemplary optical network identifying key optical components](image-url)
be considered as an upgrade of the single-wavelength system. WDM corresponds to the scheme in which multiple optical carriers at different wavelengths are modulated by using independent electrical bit streams, as shown in Fig. 2.1, and then transmitted over the same fiber. WDM has potential of exploiting the enormous bandwidth offered by the optical fiber. During transmission of WDM signals, occasionally one or several wavelengths are to be added or dropped, which is performed by the optical component known as optical add–drop multiplexer (OADM), as illustrated in Fig. 2.1. The optical networks require the switching of information among different fibers, which is performed in optical cross-connect (OXS). To combine several distinct wavelength channels into composite channel, the wavelength multiplexers are used. On the other hand, to split the composite WDM channel into distinct wavelength channels, the wavelength demultiplexers are used. To impose the information signal, optical modulators are used. The optical modulators are commonly used in combination with semiconductor lasers.

The typical receiver configuration with direct detection is shown in Fig. 2.2 [1–5]. The main purpose of the optical receiver, terminating the lightwave path, is to convert the signal coming from single-mode fiber from optical to electrical domain and process appropriately such obtained electrical signal to recover the data being transmitted. The incoming optical signal may be preamplified by an optical amplifier and further processed by an optical filter to reduce the level of amplified spontaneous emission (ASE) noise or by wavelength demultiplexer to select a desired wavelength channel. The optical signal is converted into electrical domain by using a photodetector, followed by an electrical postamplifier. To deal with residual intersymbol interference (ISI), an equalizer may be used. The main purpose of clock recovery circuit is to provide timing for decision circuit by extracting the clock from the received signal. The clock recovery circuit is most commonly implemented using the phase-lock loop (PLL). Finally, the purpose of decision circuit is to provide the binary sequence being transmitted by comparing the sampled signal to a predetermined threshold. Whenever the received sample is larger than the threshold, the decision circuit decides in favor of bit 1, otherwise in favor of bit 0.

![Fig. 2.2 A typical direct detection receiver architecture. O/E optical to electrical and AGC automatic gain control](image-url)
The optical signal generated by semiconductor laser has to be modulated by information signal before being transmitted over the optical fiber. This can be achieved by directly modulating the bias current of semiconductor laser, which can be done even at high speed (even up to 40 Gb/s in certain lasers). Unfortunately, this concept although conceptually simple is rarely used in practice because of the frequency chirp introduced by direct modulation, nonuniform frequency response, and large current swing needed to provide operation. For transmitters operating at 10 Gb/s and above, instead, the semiconductor laser diode (LD) is commonly biased at constant current to provide continuous wave (CW) output, and external modulators are used to impose the information signal to be transmitted. The most popular modulators are electro-optic optical modulators, such as Mach–Zehnder modulators, and electroabsorption modulators. The principle of the external modulator is illustrated in Fig. 2.3. Through the external modulation process, a certain parameter of the CW signal, used as a signal carrier, is varied in accordance with the information-bearing signal. For example, a monochromatic electromagnetic wave is commonly used as a carrier, and its electrical field $E(t)$ can be represented by

$$E(t) = pA \cos(\omega t + \varphi),$$

where $A$, $\omega$, and $\varphi$ are amplitude, frequency, and phase, respectively; while $p$ denotes the polarization orientation. Each of those parameters can be used to carry information, and the information-bearing signal can be either CW or discrete. If the information-bearing signal is CW, corresponding modulation formats are amplitude modulation (AM), frequency modulation (FM), phase modulation (PM), and polarization modulation (PolM). On the other hand, if the information-bearing signal is digital, the corresponding modulations are: amplitude-shift keying (ASK), frequency-shift keying (FSK), phase-shift keying (PSK), and polarization-shift keying (PolSK).

Optical fibers serve as foundation of an optical transmission system because they transport optical signals from source to destination. The combination of low-loss and large bandwidth allows high-speed signals to be transmitted over long distances before the regeneration is needed. A low-loss optical fiber is manufactured from several different materials; the base row material is pure silica, which is mixed with different dopants in order to adjust the refractive index of optical fiber. The optical fiber consists of two waveguide layers, the core and the cladding, protected by buffer coating. The majority of the power is concentrated in the core, although some portion can spread to the cladding. There exists a difference in refractive indices between the core and cladding, which is achieved by mixing dopants, commonly added
2.2 Key Optical Components

There exist two types of optical fibers: multimode fiber (MMF) and single-mode fiber (SMF). Multimode optical fibers transfer the light through a collection of spatial transversal modes. Each mode, defined through a specified combination of electrical and magnetic components, occupies a different cross section of the optical fiber core and takes a slightly distinguished path along the optical fiber. The difference in mode path lengths in multimode optical fibers produces a difference in arrival times at the receiving point. This phenomenon is known as multimode dispersion (or intermodal dispersion) and causes signal distortion and imposes the limitations in signal bandwidth. The second type of optical fibers, SMFs, effectively eliminates multimode dispersion by limiting the number of propagating modes to a fundamental one. SMFs, however, introduce another signal impairment known as the chromatic dispersion. Chromatic dispersion is caused by the difference in velocities among different spectral components within the same mode.

The attenuation of signal propagating through optical fiber is low compared to that through other transmission media, such as copper cables or free space. Nevertheless, we have to amplify the attenuated signal from time to time, to restore the signal level, without any conversion into electrical domain. This can be done in optical amplifiers, through the process of stimulated emission. The main ingredient of an optical amplifier is the optical gain realized through the amplifier pumping (being either electrical or optical) to achieve the so-called population inversion. The common types of optical amplifiers are semiconductor optical amplifiers (SOAs), EDFAs, and Raman amplifiers. The amplification process is commonly followed by the noise process, not related to the signal, which occurs due to spontaneous emission. The amplification process degrades the SNR, because of ASE added to the signal in every amplifier stage.

Before providing more details about basic building blocks identified in this section, let us give more global picture by describing a typical optical network shown in Fig. 2.4. We can identify three ellipses representing the core network, the edge network, and the access network.
network, and the access network [1]. The long-haul core network interconnects big cities, major communications hubs, and even different continents by means of submarine transmission systems. The core networks are often called the wide area networks (WANs) or interchange carrier networks. The edge optical networks are deployed within smaller geographical areas and are commonly recognized as metropolitan area networks (MANs) or local exchange carrier networks. The access networks represent peripheral part of optical network and provide the last-mile access or the bandwidth distribution to the individual end-users. The common access networks are local area networks (LANs) and distribution networks. The common physical network topologies are mesh network (often present in core networks), ring network (in edge networks), and star networks (commonly used in access networks).

Given this general description of key optical components in the rest of this section, we provide more details about basic building blocks: optical transmitters are described in Sect. 2.2.1, optical receivers in Sect. 2.2.2, optical amplifiers in Sect. 2.2.3, optical fibers in Sect. 2.2.4, and other optical building blocks, such as multiplexers/demultiplexers, optical filters, OADMs, optical switches, couplers, etc., are described in Sect. 2.2.5.

### 2.2.1 Optical Transmitters

The role of the optical transmitter is to generate the optical signal, impose the information-bearing signal, and launch the modulated signal into the optical fiber. The semiconductor light sources are commonly used in state-of-the-art optical communication systems. The light generation process occurs in certain semiconductor materials due to recombination of electrons and holes in p–n junctions, under direct biasing. Depending on the nature of the recombination process, we can classify different semiconductor light sources as either light-emitting diodes (LEDs) in which spontaneous recombination dominates or semiconductor lasers in which the stimulated emission is a dominating mechanism. Namely, there are three basic processes in semiconductor materials, as illustrated in Fig. 2.5, by which the light interacts with matter: absorption, spontaneous emission, and stimulated emission. In normal conditions, the number of electrons in ground state (with energy $E_1$) $N_1$ is larger than the number of electrons in excited state (with energy $E_2$) $N_2$, and in the thermal equilibrium their ratio follows the Boltzmann’s statistics [1, 3, 4]

$$\frac{N_2}{N_1} = \exp\left(-\frac{h\nu}{k_B T}\right) = \exp\left(-\frac{E_2 - E_1}{k_B T}\right), \quad (2.2)$$

![Fig. 2.5](image-url) Illustrating the interaction of the light with the matter
2.2 Key Optical Components

where $h\nu$ is a photon energy ($h$ is the Plank’s constant and $\nu$ is the optical frequency proportional to the energy difference between the energy levels $E_2 - E_1$), $k_B$ is the Boltzmann’s constant, and $T$ is the absolute temperature. In the same regime, the spontaneous emission rate $dN_{2,\text{spon}}/dt = A_{21}N_2 (A_{21}$ denotes the spontaneous emission coefficient) and the stimulated emission rate $dN_{2,\text{stim}}/dt = B_{21}\rho(\nu)N_2$ ($B_{21}$ denotes the stimulated emission coefficient and $\rho(\nu)$ denotes the spectral density of electromagnetic energy) are equalized with absorption rate $dN_{1,\text{abs}}/dt = A_{12}\rho(\nu)N_1 (A_{12}$ denotes the absorption coefficient):

$$A_{21}N_2 + B_{21}\rho(\nu)N_2 = B_{12}\rho(\nu)N_1.$$  

(2.3)

In the visible or near-infrared region ($h\nu \sim 1\text{ eV}$), the spontaneous emission always dominates over stimulated emission in thermal equilibrium at room temperature ($k_B T \approx 25\text{ meV}$) [see (2.2)]. The stimulated emission rate can exceed absorption rate only when $N_2 > N_1$; the condition is referred to as population inversion and can never be realized for systems being in thermal equilibrium. The population inversion is a prerequisite for laser operation; and in atomic system, it is achieved by using three- and four-levels pumping schemes (an external energy source raises the atomic population from ground to an excited state). There are three basic components required to sustain stimulated emission and to form useful laser output: the pump source, the active medium, and the feedback mirrors. The active medium can be solid (such as in semiconductor lasers), gaseous, or liquid in nature. The pump can be electrical (e.g., semiconductor lasers), optical, or chemical. The purpose of the pump is to achieve the population inversion. The basic structure of semiconductor laser of Fabry–Perot type is shown in Fig. 2.6a, together with the equivalent model. The injection (bias) current flows through the p–n junction and stimulates the recombination of electrons and holes, leading to the generation of photons.

For the lasing action to be sustainable, the gain and phase matching condition should be satisfied. In the active medium, both gain/absorption described by $\gamma(\nu)$ and scattering described by $\alpha_s$ are present. The intensity inside the cavity can be described by the following dependence $I(z) = I_0 \exp[(\gamma(\nu) - \alpha_s)z]$. The lasing is possible when collective gain is larger than the loss after a round trip pass through the cavity:

$$I(2L) = I_0 R_1 R_2 \exp[2L(\gamma(\nu) - \alpha_s)] = I_0.$$  

(2.4)

where $R_1$ and $R_2$ are facet reflectivities (see Fig. 2.6), $L$ is the length of active medium, and $I_0$ and $I(2L)$ correspond to initial and round-trip intensities. The gain threshold is obtained by solving (2.4) per $\gamma$:

$$\gamma_{\text{th}} = \alpha_s + \frac{1}{2L} \ln \left( \frac{1}{R_1 R_2} \right) = \alpha_{\text{intr}} + \alpha_{\text{mir}},$$  

(2.5)
Fig. 2.6  Semiconductor lasers (a) Fabry–Perot semiconductor laser, (b) DFB laser, and (c) DBR laser

where the internal losses (corresponding to $\alpha_{\text{int}}$) and mirror losses ($\frac{1}{2L}\ln (1/ R_1 R_2)$) are denoted by $\alpha_{\text{tot}}$ and $\alpha_{\text{mir}}$, respectively. After the round trip the resultant phase must be equal to the initial phase, leading to the phase-matching condition:

$$\exp[-j2\beta L] = 1 \Rightarrow 2\beta L = q2\pi, \beta = 2\pi n/\lambda, \quad (2.6)$$

where $\beta$ denotes the propagation constant, $n$ is the refractive index of active medium, and $\lambda$ is the free-space wavelength. The phase-matching condition can be satisfied for many different integers $q$, representing different longitudinal modes of frequency $\nu_q = qc/(2nL)$. The separation between neighboring longitudinal modes is known as the free-spectral range:

$$\Delta\nu = \nu_q - \nu_{q-1} = q \frac{c}{2nL} - (q - 1) \frac{c}{2nL} = \frac{c}{2nL}. \quad (2.7)$$

Because of the presence of longitudinal modes, the Fabry–Perot laser belongs to the class of multimode lasers. To improve the coherence of output light and the laser modulation speed, distributed feedback (DFB) lasers, shown in Fig. 2.6b, are used. The key idea of this laser is to effectively select one of the longitudinal modes while suppressing the remaining ones. This is achieved by introducing the Bragg grating inside of the laser cavity. The wavelength of selected longitudinal mode can be determined from the Bragg condition:

$$2\Lambda = m \frac{\lambda}{n_{\text{av}}} \Rightarrow \lambda_B = \frac{2n_{\text{av}}\Lambda}{m}, \quad (2.8)$$
where $\Lambda$ is the grating period, $n_{av}$ is the average refractive index of a waveguide mode, and $\lambda/n_{av}$ is the average wavelength of the light in the waveguide mode. If the grating element is put outside of the active region or instead of the facet mirrors, the distributed Bragg reflector (DBR) laser is obtained, which is illustrated in Fig. 2.6c. Both DFB and DBR lasers belong to the class of single-mode lasers. Different semiconductor lasers shown in Fig. 2.6 are edge-emitting lasers.

Another important semiconductor laser type is vertical cavity surface emitting laser (VCSEL), which emits the light vertical to the active layer plane [1–4]. The VCSELs are usually based on In-GaAs-P layers acting as Bragg reflectors and provide the positive feedback leading to the stimulated emission.

The spectral curve of the single-mode lasers is a result of transition between discrete energy levels and can often be represented using the Lorentzian shape [1–4]:

$$g(v) = \frac{\Delta v}{2\pi \left[ (v - v_0)^2 + \left( \frac{\Delta v}{2} \right)^2 \right]},$$

(2.9)

where $v_0$ is the central optical frequency and $\Delta v$ represents the laser linewidth [1]:

$$\Delta v = \frac{n_{sp}G \left( 1 + \alpha_{\text{chirp}}^2 \right)}{4\pi P},$$

(2.10)

where $n_{sp}$ is the spontaneous emission factor, $G$ is the net rate of stimulated emission, $P$ denotes the output power, and $\alpha_{\text{chirp}}$ is the chirp factor (representing the amplitude-phase coupling parameter).

The small-signal frequency response of the semiconductor laser is determined by [3, 4]

$$H(\omega) = \frac{\Omega_R^2 + \Gamma_R^2}{(\Omega_R + \omega - j\Gamma_R)(\Omega_R - \omega + j\Gamma_R)},$$

(2.11)

where $\Gamma_R$ is the damping factor and $\Omega_R$ is the relaxation frequency.

$$\Omega_R^2 \approx \frac{G_N P_b}{\tau_p} \quad (\Gamma_R \ll \Omega_R)$$

with $G_N$ being the net rate of stimulated emission, $P_b$ being the output power corresponding to the bias current, and $\tau_p$ being the photon lifetime related to the excited energy level. The modulation bandwidth (defined as 3-dB bandwidth) is therefore determined by the relaxation frequency

$$\omega_{3\text{dB}} = \sqrt{1 + \sqrt{2}\Omega_R},$$

and for fast semiconductor lasers it can be 30 GHz. Unfortunately, the direct modulation of semiconductor lasers leads to frequency chirp, which can be described by the instantaneous frequency shift from steady-state frequency $v_0$ as follows [1]:
where \( P(t) \) is the time variation of the output power, \( \chi \) is the constant (varying from zero to several tens) related to the material and design parameters, and \( \alpha_{\text{chirp}} \) is the chirp factor defined as the ratio between the refractive index change and gain change with respect to the number of carriers \( N: \alpha_{\text{chirp}} = (dn/dN)/(dG/dN) \). The first term on the right-hand side of (2.12) represents dynamic (transient or instantaneous) chirp and the second term the adiabatic (steady-state) frequency chirp. The random fluctuation in carrier density due to spontaneous emission also leads to the linewidth enhancement proportional to \( 1 + \alpha_{\text{chirp}}^2 \). To avoid the chirp problem, the external modulation is used; while the semiconductor lasers are biased by a dc voltage to produce a continuous wave operation.

There are two types of external modulators commonly used in practice: Mach–Zehnder modulator (MZM) and electroabsorption modulator (EAM), whose operational principle is illustrated in Fig. 2.7. The MZM is based on electro-optic effect, the effect that in certain materials (such as LiNbO\(_3\)) where the refractive index \( n \) changes with respect to the voltage \( V \) applied across electrodes [4]:

\[
\Delta n = -\frac{1}{2} \Gamma n^3 r_{33}(V/d_c) \Rightarrow \Delta \phi = \frac{2\pi}{\lambda} \Delta n L, \tag{2.13}
\]

where \( \Delta n \) denotes the refractive index change, \( \Delta \phi \) is corresponding phase change, \( r_{33} \) is the electro-optic coefficient (\( \sim 30.9 \text{ pm/V} \) in LiNbO\(_3\)), \( d_c \) is the separation of electrodes, \( L \) is the electrode length, and \( \lambda \) is the wavelength of the light. The MZM (see Fig. 2.7a) is a planar waveguide structure deposited on the substrate, with two pairs of electrodes (1) for high-speed ac voltage representing the modulation data (RF) signal and (2) for dc bias voltage. Let \( V_1(t) \) and \( V_2(t) \) denote the electrical drive signals on the upper and lower electrodes, respectively. The output electrical field \( E_{\text{out}}(t) \) of the second Y-branch can be related to the input electrical field \( E_{\text{in}} \) by

\[
E_{\text{out}}(t) = \frac{1}{2} \left[ \exp \left( j \frac{\pi}{V_1(t)} \right) + \exp \left( j \frac{\pi}{V_2(t)} \right) \right] E_{\text{in}}, \tag{2.14}
\]
2.2 Key Optical Components

where $V_{\pi}$ is differential drive voltage ($V_1 - V_2 = V_{\pi}$) resulting in differential phase shift of $\pi$ rad between two waveguides. Possible modulation formats that can be used with this MZM include: on–off keying (OOK) with zero/nonzero chirp, binary phase-shift keying (BPSK), differential phase-shift keying (DPSK), quadrature phase-shift keying (QPSK), differential QPSK (DQPSK), and return-to-zero (RZ) with duty cycle 33%, 50%, or 67%. For example, for zero-chirp OOK or BPSK the following complementary drive signals should be applied:

\[
V_1(t) = V(t) - V_{\pi}/2, \quad V_2(t) = -V(t) + V_{\pi}/2 \Rightarrow \frac{E_{\text{out}}(t)}{E_{\text{in}}} = \sin\left(\frac{\pi V(t)}{V_{\pi}}\right).
\]

The electroabsorption modulator (EAM) is a semiconductor-based planar waveguide composed of multiple p-type and n-type layers that form multiple quantum wells (MQWs). The basic design of EAM is similar to that of semiconductor lasers. The MQW is used to support the quantum-confined Stark effect (the absorption spectrum being a function of applied field) more effectively. Because of similarities of EAMs and semiconductor lasers design it is possible to fabricate them on the same substrate (see Fig. 2.7b), providing that EAM and laser are electrically isolated. Bandgap of quantum wells is larger than photon energy, so that the light is completely transmitted in the absence of bias, which corresponds to the ON state. When the reverse bias is applied the input signal is absorbed, which corresponds to the OFF state. The modulation speed of EAMs is typically comparable to the modulation speed of MZMs. However, the extinction ratio (the ratio of average powers corresponding to symbol 1 and symbol 0) is lower.

### 2.2.2 Optical Receivers

The purpose of the optical receiver is to convert the optical signal into electrical domain and to recover the transmitted data. The typical OOK receiver configuration is already given in Fig. 2.2. We can identify three different stages: front-end stage, the linear channel stage, and data recovery stage. The front-end stage is composed of a photodetector and a preamplifier. The most commonly used front-end stages are high-impedance front-end and transimpedance front-end, both shown in Fig. 2.8. High-impedance front end (Fig. 2.8a) employs a large value load resistance to reduce the level of thermal noise and has a good receiver sensitivity. However, the bandwidth of this scheme is low because the RC constant is large. To achieve both the high receiver sensitivity and large bandwidth, the transimpedance front-end scheme, shown in Fig. 2.8b, is used. Even though the load resistance is high, the negative feedback reduces the effective input resistance by a factor of $G - 1$, where $G$ is the front-end amplifier gain. The bandwidth is increased for the same factor compared to high-impedance front-end scheme.

The photodiode is an integral part of both front-end stage schemes. The key role of the photodiode is to absorb photons in incoming optical signal and convert...
Fig. 2.8 Optical receiver front-end stage schemes (a) high-impedance front-end and (b) transimpedance front-end

Fig. 2.9 The semiconductor photodiodes (a) p–i–n photodiode and (b) avalanche photodiode. (c) The equivalent p–i–n photodiode model

them back to the electrical level through the process opposite to the one taking place in semiconductor lasers. The common photodiodes are p–n photodiode, p–i–n photodiode, avalanche photodiode (APD), and metal–semiconductor–metal (MSM) photodetectors [3,4]. p–n photodiode is based on a reverse-biased p–n junction. The thickness of the depletion region is often less than the absorption depth for incident light, and the photons are absorbed outside of the depletion region, leading to the slow response speed. p–i–n photodiode consists of an intrinsic region sandwiched between p- and n-type layers, as shown in Fig. 2.9a. Under the reverse bias,
the depletion depth can be made sufficiently thick to absorb most of the incident photons. Avalanche photodiode, shown in Fig. 2.9b, is a modified p–i–n photodiode that is operated at very high reverse bias. Under high-field conditions, photogenerated carriers induce generation of secondary electron–hole pairs by the process of impact ionization, and this process leads to internal electrical gain. MSM photodetectors employ interdigitated Schottky barrier contacts on one face of the device and are compatible with planar processing and optoelectronic integration. Depending on the device design, the device is illuminated through the p- or n-type contact. In Si, Ge, or GaAs diodes, the substrate is absorbing so that the device has to be illuminated through the top contact, as shown in Fig. 2.9a. On the other hand, in InGaAs or InGaAsP, the substrate is transparent, and the device can be designed to be illuminated either through the substrate or through the top contact. In order to increase the depletion region and to minimize the diffusion current component, an intrinsic layer (i-type) is introduced to the p–i–n photodiode structure. The p–i–n photodiode is reverse biased and has very high internal impedance, meaning that it acts as a current source generating the photocurrent proportional to the incoming optical signal power. The equivalent scheme of p–i–n photodiode is shown in Fig. 2.9c. Typically the internal series resistance $R_s$ is low, while the internal shunt resistance is high, so that the junction capacitance $C_p$ dominates and can be determined by

$$C_p = \varepsilon_s \frac{A}{w} = \left[ \frac{\varepsilon_s N_A N_D}{2(N_A - N_D)(V_0 - V_A)} \right]^{1/2}, \quad (2.15)$$

where $\varepsilon_s$ is the semiconductor permittivity, $A$ is the area of the space charge region (SCR), $w$ is the width of SCR, $N_A$ and $N_D$ denote dopant (acceptor and donor) densities, $V_0$ is the built-in potential across the junction, and $V_A$ is applied negative voltage. The photocurrent $i_{ph}(t)$ is proportional to the power of incident light $P(t)$, that is $i_{ph}(t) = RP(t)$, where $R [A/W]$ is the photodiode responsivity. The photodiode responsivity is related to the quantum efficiency $\eta$, defined as the ratio of number of generated electrons and the number of incident photons, by $R = \eta q/h\nu$, where $q$ is an electron charge and $h\nu$ is a photon energy. Using this model we can determine the 3-dB bandwidth of high-impedance front-end scheme as $B_{3dB} = 1/(2\pi R L C_p)$ and the 3-dB bandwidth of transimpedance front-end scheme as $B_{3dB} = (G + 1)/(2\pi R F C_p)$, which is $G$ times higher than bandwidth of high-impedance front-end scheme.

### 2.2.3 Optical Fibers

Optical fibers serve as foundation of an optical transmission system because they transport optical signals from source to destination. The combination of low-loss and extremely large bandwidth allows high-speed signals to be transmitted over long distances before the regeneration becomes necessary. A low-loss optical fiber is manufactured from several different materials; the base row material is pure silica,
which is mixed with different dopants in order to adjust the refractive index of optical fiber. The optical fiber, shown in Fig. 2.10, consists of two waveguide layers, the core (of refractive index \(n_1\)) and the cladding (of refractive index \(n_2\), protected by the jacket (the buffer coating). The majority of the power is concentrated in the core, although some portion can spread to the cladding. There is a difference in refractive indices between the core and cladding \((n_1 > n_2)\), which is achieved by a mix of dopants commonly added to the fiber core. The refractive-index profile for step-index fiber is shown in Fig. 2.10c, while the illustration of light confinement by the total internal reflection is shown in Fig. 2.10d. The ray will be totally reflected from the core–cladding interface (a guided ray) if the following condition is satisfied:

\[
n_0 \sin \theta_i < \sqrt{n_1^2 - n_2^2},
\]

where \(\theta_i\) is the angle of incidence. \(\max(n_0 \sin \theta_i)\) defines the light gathering capacity of an optical fiber and it is called the numerical aperture (NA):

\[
NA = \sqrt{n_1^2 - n_2^2} \approx n_1 \sqrt{2\Delta}, \quad \Delta = 1,
\]

where \(\Delta\) is the normalized index difference defined as \(\Delta = (n_1 - n_2)/n_1\). Therefore, from the geometrical optics point of view, light propagates in optical fiber due to series of total internal reflections that occur at the core–cladding interface. The smallest angle of incidence \(\phi\) (see Fig. 2.10d) for which the total internal reflection occurs is called the critical angle and equals \(\sin^{-1} n_2/n_1\).
There exist two types of optical fibers: MMF (shown in Fig. 2.10a) and SMF (shown in Fig. 2.10b). Multimode optical fibers transfer the light through a collection of spatial transversal modes. Each mode, defined through a specified combination of electrical and magnetic components, occupies a different cross section of the optical fiber core and takes a slightly distinguished path along the optical fiber. The difference in mode path lengths in multimode optical fibers produces a difference in arrival times at the receiving point. This phenomenon is known as multimode dispersion (or intermodal dispersion) and causes signal distortion and imposes the limitations in signal bandwidth. The second type of optical fibers, SMFs, effectively eliminates multimode dispersion by limiting the number of propagating modes to a fundamental one. The fundamental mode occupies the central portion of the optical fiber and has an energy maximum at the axis of the optical fiber core. Its radial distribution can be approximated by Gaussian curve. The number of modes \( M \) that can effectively propagate through an optical fiber is determined by the normalized frequency (\( V \) parameter or \( V \) number): \( M \approx V^2/2 \), when \( V \) is large. The normalized frequency is defined by

\[
V = \frac{2\pi a}{\lambda} \sqrt{n_1^2 - n_2^2},
\]

where \( a \) is the fiber core radius, \( \lambda \) is the carrier wavelength, and \( n_1 \) and \( n_2 \) are refractive indices related to the fiber core and the fiber cladding, respectively.

Each mode propagating through the fiber is characterized by its own propagation constant \( \beta \). The dependence of the electric and magnetic fields on axial coordinate \( z \) is expressed through the factor \( \exp(-j\beta z) \). The propagation constant must satisfy the following condition:

\[
2\pi n_2/\lambda < \beta < 2\pi n_1/\lambda. \tag{2.17}
\]

In order to evaluate the transmission characteristics of the optical fiber, the functional dependence of the mode propagation constant on the optical signal wavelength has to be known. The normalized propagation constant \( b \) is defined for that purpose:

\[
b = \frac{\beta^2 - (2\pi n_2/\lambda)^2}{(2\pi n_1/\lambda)^2 - (2\pi n_2/\lambda)^2}. \tag{2.18}
\]

The normalized propagation constant is related to the normalized frequency \( V \) by [1, 3]

\[
b(V) \approx (1.1428 - 0.9960/V)^2, \quad 1.5 \leq V \leq 2.5. \tag{2.19}
\]

The multimode dispersion can effectively be eliminated by limiting the number of propagating modes to a fundamental one: \( V \leq V_c = 2.405 \) with \( V_c \) being the cutoff frequency. The cutoff frequency is controlled by keeping the core radius small and the normalized index difference \( \Delta = (n_1 - n_2)/n_1 \) between 0.2 and 0.3%.
2.2.4 Optical Amplifiers

The purpose of an optical amplifier is to restore the signal power level, reduced due to losses during propagation, without any optical to electrical conversion. The general form of an optical amplifier is given in Fig. 2.11a. Most optical amplifiers amplify incident light through the stimulated emission, the same mechanism that is used in lasers, but without the feedback mechanism. The main ingredient is the optical gain realized through the amplifier pumping (electrical or optical) to achieve the population inversion. The optical gain, generally speaking, is not only a function of frequency, but also a function of local beam intensity. To illustrate the basic concepts, we consider the case in which the gain medium is modeled as two-level system, as shown in Fig. 2.11b. The amplification factor $G$ is defined as the ratio of amplifier output $P_{out}$ and input $P_{in}$ powers $G = P_{out}/P_{in}$. The amplification factor can be determined by knowing the dependence of evolution of power through the gain media [3]:

$$\frac{dP}{dz} = gP, \quad g(\omega) = \frac{g_0}{1 + (\omega - \omega_0)^2 T_2^2 + P/P_S}.$$  \hspace{1cm} (2.20)

$g$ is the gain coefficient, $g_0$ is the gain peak value, $\omega_0$ is the atomic transition frequency, $T_2$ is the dipole relaxation time ($<1$ ps), $\omega$ is the optical frequency of incident signal, $P$ is the incident signal power, and $P_S$ is the saturation power.

In the unsaturated regime ($P \ll P_S$), the differential equation (2.20) can be solved by separation of variables to get the following dependence of power $P(z) = P(0) \exp(gz)$, so that the amplification factor can be obtained by

$$G(\omega) = \exp[g(\omega)L].$$ \hspace{1cm} (2.21)

and corresponding full-width half-maximum (FWHM) bandwidth is determined by

$$\Delta v_A = \Delta v_g \sqrt{\frac{\ln 2}{\ln (G_0/2)}}, \quad G_0 = \exp(g_0L).$$ \hspace{1cm} (2.22)

where $\Delta v_g$ is the FWHM gain coefficient bandwidth.

![Fig. 2.11](image)

Fig. 2.11 (a) Optical amplifier principle, (b) two-level amplifier system model
The \textit{gain saturation} comes from the power dependence of the gain coefficient (2.20). The coefficient is reduced when the incident power $P$ becomes comparable with the saturation power $P_S$. Let us assume that the incident frequency is tuned to the peak gain ($\omega = \omega_0$), then from (2.20) we obtain

$$\frac{dP}{dz} = \frac{g_0 P}{1 + P/P_S}.$$ \hspace{1cm} (2.23)

By solving the differential equation (2.23) with respect to the \textit{boundary conditions}, $P(0) = P_{\text{in}}$, and $P(L) = P_{\text{out}} = GP_{\text{in}}$, we get

$$G = G_0 \exp \left[ - \frac{G - 1}{G} \frac{P_{\text{out}}}{P_S} \right].$$ \hspace{1cm} (2.24)

From (2.24) we can determine another important optical amplifier parameter, the \textit{output saturation power} as being the optical power at which the gain $G$ is reduced to $G_0/2$ (3 dB down):

$$P_{\text{sat}} = \frac{G_0 \ln 2}{G_0 - 2} P_S \approx (\ln 2) P_S \approx 0.69 P_S (G_0 > 20 \text{ dB}).$$ \hspace{1cm} (2.25)

Three common applications of optical amplifiers are (1) power boosters (of transmitters), (2) in-line amplifiers, (3) optical preamplifiers, which is illustrated in Fig. 2.12. The booster (power) amplifiers are placed at the optical transmitter side to enhance the transmitted power level or to compensate for the losses of optical elements between the laser and optical fibers, such as optical coupler, WDM multiplexers, and external optical modulators. The in-line amplifiers are placed along the transmission link to compensate for the losses incurred during propagation of optical signal. The optical preamplifiers are used to increase the signal level before the photodetection takes place, improving therefore the receiver sensitivity.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2_12.png}
\caption{Possible application of optical amplifiers (a) booster amplifier, (b) in-line amplifiers, and (c) preamplifier}
\end{figure}
Several types of optical amplifiers have been introduced so far: SOAs, fiber Raman (and Brillouin) amplifiers, rare-earth-doped fiber amplifiers (erbium-doped EDFA operating at 1,500 nm and praseodymium-doped PDFA operating at 1,300 nm), and parametric amplifiers. Semiconductor lasers act as amplifiers before reaching the threshold. To prevent the lasing, antireflection (AR) coatings are used in SOAs, as shown in Fig. 2.13. Even with AR coating, the multiple reflections must be included when considering the Fabry–Perot (FP) cavity. The gain of FP amplifiers is given by [3]

$$G_{FP} = \frac{(1 - R_1)(1 - R_2)G(v)}{(1 - G(v) \sqrt{R_1 R_2})^2 + 4G \sqrt{R_1 R_2} \sin^2 \left[ \pi (v - v_m)/\Delta v_L \right]}.$$  \hspace{1cm} (2.26)

where $R_1$ and $R_2$ denote the facet reflectivities, $G(v)$ is the single-pass amplification factor, $v_m$ is the cavity resonance frequency, and $\Delta v_L$ is the free-spectral range.

FP amplifier bandwidth can be determined as follows [3]:

$$\Delta v_A = \frac{2 \Delta v_L}{\pi} \sin^{-1} \left[ \frac{1 - G \sqrt{R_1 R_2}}{(4G \sqrt{R_1 R_2})^{1/2}} \right].$$  \hspace{1cm} (2.27)

A fiber-based Raman amplifier employs the stimulated Raman scattering (SRS) occurring in silica fibers when an intense pump propagates through it. SRS fundamentally differs from stimulated emission: in stimulated emission an incident photon stimulates emission of another identical photon, but in SRS the incident pump photon gives up its energy to create another photon of reduced energy at a lower frequency (inelastic scattering); the remaining energy is absorbed by the medium in the form of molecular vibrations (optical phonons). Raman amplifiers must be pumped optically to provide gain, as shown in Fig. 2.14.

The Raman-gain coefficient $g_R$ is related to the optical gain $g(z)$ as $g(z) = g_R I_p(z)$, where $I_p$ is the pump intensity given by $I_p = P_p/a_p$, with $P_p$ being the pump power and $a_p$ being the pump cross-sectional area. Since the cross-sectional area is different for different types of fibers, the ratio $g_R/a_p$ is the measure of the Raman-gain efficiency. The DCF efficiency can even be eight times better.
2.2 Key Optical Components

Fig. 2.14 The Raman amplifier operation principle in forward-pumping configuration

than that of a standard SMF, as shown in [3]. The evolution of the pump \( P_p \) and signal \( P_s \) powers (in distance \( z \)) can be studied by solving the system of coupled differential equations below [1, 3]:

\[
\frac{dP_s}{dz} = -\alpha_s P_s + \frac{g_R}{a_p} P_p P_s, \quad \frac{dP_p}{dz} = -\alpha_p P_p - \frac{\omega_p}{\omega_s} \frac{g_R}{a_p} P_p P_s, \tag{2.28}
\]

where \( \alpha_s \) denotes the signal cross-sectional area and \( \omega_p \) and \( \omega_s \) denote the pump and signal frequency, respectively, while other parameters are already introduced above.

In small-signal amplification regime (when the pump depletion can be neglected), the pump power evolution is exponential, \( P_p(z) = P_p(0) \exp(-\alpha_p z) \), so that the Raman amplifier gain is found to be

\[
G_A = \frac{P_s(0) \exp(g_R P_p(0) L_{\text{eff}}/a_p - \alpha_s L)}{P_s(0) \exp(-\alpha_s L)} = \exp(g_0 L),
\]

\[
g_0 = \frac{g_R P_p(0) L_{\text{eff}}}{a_p} \approx \frac{g_R P_p(0)}{a_p \alpha_p L} (\alpha_p L \ll 1). \tag{2.29}
\]

The origin of saturation in Raman amplifiers is pump power depletion, which is quite different from that in SOAs. Saturated amplifier gain \( G_S \) can be determined (assuming \( \alpha_p = \alpha_s \)) by [3]

\[
G_S = \frac{1 + r_0}{r_0 + 1/G_A r_0}, \quad r_0 = \frac{\omega_s P_s(0)}{\omega_p P_p(0)}. \tag{2.30}
\]

The amplifier gain is reduced down by 3 dB when \( G_A r_0 \approx 1 \), the condition that is satisfied when the amplified signal power becomes comparable to the input pump power \( P_s(L) = P_p(0) \). Typically \( P_p \approx 1 \text{ W} \), and channel powers in a WDM systems are around 1 mW, meaning that Raman amplifier operates in unsaturated or linear regime.

The rare-earth doped fiber amplifiers are finding increasing importance in optical communication systems. The most important class is EDFAs due to their ability to amplify in 1.55-μm wavelength range. The active medium consists of 10–30 m length of optical fiber highly doped with a rare-earth element, such as erbium (Er), ytterbium (Yb), neodymium (Nd), or praseodymium (Pr). The host fiber material
can be pure silica, a fluoride-based glass, or a multicomponent glass. General EDFA configuration is shown in Fig. 2.15.

The pumping at a suitable wavelength provides gain through population inversion. The gain spectrum depends on the pumping scheme as well as on the presence of other dopants, such as Ge or Al within the core. The amorphous nature of silica broadens the energy levels of Er$^{3+}$ into the bands, as shown in Fig. 2.16.

The pumping is primarily done in optical domain with the primary pump wavelengths at 1.48 μm and 0.98 μm. The atoms pumped to the $4I_{11/2}$ level (with 980 nm pump) decay to the primary emission transition band. The pumping with 1.48 μm is directly excited to the upper transition levels of the emission band.

EDFAs can be designed to operate in such a way that the pump and signal travel in opposite directions; this configuration is commonly referred to as backward pumping. In bidirectional pumping, the amplifier is pumped in both directions simultaneously by using two semiconductor lasers located at both fiber ends.

### 2.2.5 Other Optical Components

Different optical components can be classified, depending on whether they can operate without an external electric power source or not, into two broad categories: passive or active. Important active components are lasers, external modulators, optical amplifiers, photodiodes, optical switches, and wavelength converters. On the
other hand, important passive components are optical couplers, isolators, multiplexers/demultiplexers, and filters. Some components, such as optical filters, can be either passive or active depending on operational principle. In this section, we will briefly explain some important optical components not being described in previous sections.

The $2 \times 2$ optical coupler is a fundamental device that can either be implemented using the fiber fusing or be based on graded-index (GRIN) rods and optical filters, as shown in Fig. 2.17. The fused optical couplers (shown in Fig. 2.17a) are obtained when the cladding of two optical fibers are removed, the cores are brought together, and then heated and stretched. The obtained waveguide structure can exchange energy in the coupling region between the branches. If both inputs are used $2 \times 2$ coupler is obtained, if only one input is used $1 \times 2$ coupler is obtained. The optical couplers are recognized either as optical tap ($1 \times 2$) couplers or directional ($2 \times 2$) couplers. The power coupler splitting ratio depending on purpose can be different with typical values being 50%/50%, 10%/90%, 5%/95%, and 1%/99%. Directional coupler parameters (defined when only input 1 is active) are splitting ratio $P_{\text{out},1}/(P_{\text{out},1} + P_{\text{out},2})$, excess loss $10 \log_{10}[P_{\text{in},1}/(P_{\text{out},1} + P_{\text{out},2})]$, insertion loss $10 \log_{10}(P_{\text{in},i}/P_{\text{out},j})$, and crosstalk $10 \log_{10}(P_{\text{cross}}/P_{\text{in},1})$. The operation principle of directional coupler can be explained using coupled mode theory [15] or simple scattering (propagation) matrix $S$ approach, assuming that a coupler is lossless and reciprocal device:

$$\begin{pmatrix}
E_{\text{out},1} \\
E_{\text{out},2}
\end{pmatrix} =
S
\begin{pmatrix}
E_{\text{in},1} \\
E_{\text{in},2}
\end{pmatrix} =
\begin{pmatrix}
s_{11} & s_{12} \\
s_{21} & s_{22}
\end{pmatrix}
\begin{pmatrix}
E_{\text{in},1} \\
E_{\text{in},2}
\end{pmatrix}
= e^{-j\beta L}
\begin{pmatrix}
\cos(kL) & j\sin(kL) \\
j\sin(kL) & \cos(kL)
\end{pmatrix}
\begin{pmatrix}
E_{\text{in},1} \\
E_{\text{in},2}
\end{pmatrix},
$$

(2.31)
where $\beta$ is propagation constant, $k$ is coupling coefficient, $L$ is the coupling region length, $E_{in,1}$ and $E_{in,2}$ are corresponding inputs electrical fields, and $E_{out,1}$ and $E_{out,2}$ are corresponding output electrical fields. Scattering matrix $S$ elements are denoted with $s_{ij}$. For example, for 3-dB coupler we have to select $kL = (2m + 1)\pi / 4$ ($m$ is a positive integer) to get

$$
\begin{pmatrix}
E_{out,1} \\
E_{out,2}
\end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix}
1 & j \\
-j & 1
\end{pmatrix}
\begin{pmatrix}
E_{in,1} \\
E_{in,2}
\end{pmatrix}.
$$

(2.32)

The combination of two GRIN rods and an optical filter can effectively be used as an optical coupler, as illustrated in Fig. 2.17b. The GRIN rods are used as collimators, to collimate the light from two input ports and deliver to the output port, while the optical filter is used to select a desired wavelength channel.

The optical couplers can be used to create more complicated optical devices such as $M \times N$ optical stars, directional optical switches, different optical filters, multiplexers, etc.

An optical filter modifies the spectrum of incoming light and can mathematically be described by corresponding transfer function $H_{of}(\omega):$

$$
E_{out}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{E}_{in}(\omega) H_{of}(\omega) e^{j\omega t} d\omega,
$$

(2.33)

where $E_{in}(t)$ and $E_{out}(t)$ denote the input and output electrical field, respectively, and we used $\sim$ to denote the FT as before. Depending on the operational principle, the optical filters can be classified into two broad categories as diffraction or interference filters. The important class of optical filters is tunable optical filters, which are able to dynamically change the operating frequency to the desired wavelength channel. The basic tunable optical filter types include tunable $2 \times 2$ directional couplers, Fabry–Perot (FP) filters, Mach–Zehnder (MZ) interferometer filters, Michelson filters, and acousto-optical filters. Two basic optical filters, FP filter and MZ interferometer filter, are shown in Fig. 2.18. An FP filter is in fact a cavity between two high-reflectivity mirrors. It can act as a tunable optical filter if the cavity length is controlled for example by using a piezoelectric transducer. Tunable FP filters can also be made by using liquid crystals, dielectric thin films,

Fig. 2.18 Basic optical filters (a) Fabry–Perot filter and (b) Mach–Zehnder interferometer
2.2 Key Optical Components

A transfer function of an FP filter whose mirrors have the same reflectivity $R$, and the cavity length $L$ can be written as [4]

$$H_{FP}(\omega) = \frac{(1 - R)e^{i\pi}}{1 - Re^{i\pi}}, \quad \tau = \frac{2L}{v_g},$$

(2.34)

where $\tau$ is the round-trip time within the cavity and $v_g$ is the group velocity. The transfer function of FP filter is periodic with period being the free spectral range (FSR) $\Delta \nu = \frac{v_g}{2L}$. Another FP filter important parameter is the finesse defined as [1, 4]

$$F = \frac{\Delta \nu}{\Delta \nu_{FP}} \equiv \frac{\pi \sqrt{R}}{1 - R},$$

(2.35)

where $\Delta \nu_{FP}$ is the FP transmission peak width.

Mach–Zehnder interferometer filters, shown in Fig. 2.18b, can also be used as tunable optical filters. The first coupler splits the signal into two equal parts, which acquire different phase shifts before they interfere at the second coupler. Several MZ interferometers can be cascaded to create an optical filter. When cross output of 3-dB coupler is used, the square magnitude of transfer function is $|H_c(\omega)| = \cos^2(\omega \tau/2)$ [3], so that the transfer function of $M$-stage MZ filter based on 3-dB couplers can be written as

$$|H_{MZ}(\omega)|^2 = \prod_{m=1}^{M} \cos^2(\omega \tau_m/2),$$

(2.36)

where $\tau_m$ is the adjustable delay of $m$th ($m = 1, 2, \ldots, M$) cascade.

Multiplexers and demultiplexers are basic devices of a WDM system. Demultiplexers contain a wavelength-selective element to separate the channels of a WDM signal. Based on underlying physical principle, different demultiplexer devices can be classified as diffraction-based demultiplexers (based on a diffraction grating) and interference-based demultiplexers (based on optical filters and directional couplers).

Diffraction-based demultiplexers are based on Bragg diffraction effect and use an angular dispersive element, such as the diffraction grating. The incoming composite light signal is reflected from the grating and dispersed spatially into different wavelength components, as shown in Fig. 2.19a. Different wavelength components are focused by lenses and sent to individual optical fibers. The same device can be used as multiplexer by switching the roles of input and output ports. This device can be implemented using either conventional or GRIN lenses. To simplify design, the concave grating can be used.

The second group of optical multiplexers is based on interference effect and employs the optical couplers and filters to combine different wavelength channels into a composite WDM signal. The multiplexers employing the interference effect include thin-film filters multiplexers, and the array waveguide grating (AWG) [1–4, 8, 9, 16], which is shown in Fig. 2.19b. The AWG is highly versatile WDM device, because it can be used as a multiplexer, a demultiplexer, a drop-and-insert element, or even as a wavelength router. It consists of $M_{\text{input}}$ and $M_{\text{output}}$ slab waveguides and two identical focusing planar star couplers connected by $N$ uncoupled waveguides with a
propagation constant $\beta$. The length of adjacent waveguides in the central region differ by a constant value $\Delta L$, with corresponding phase difference being $2\pi n_c \Delta L / \lambda$, where $n_c$ is the refractive index of arrayed waveguides. Based on the phase-matching condition (see Fig. 2.19c) and knowing that the focusing is achieved when the path length difference $\Delta L$ between adjacent array waveguides be an integer multiple of the central design wavelength $\lambda_c$, that is $n_c \Delta L = m \lambda_c$, we derive the following expression for the channel spacing [16]:

$$\Delta \nu = \frac{y n_s cd n_c}{L m \lambda^2 n_g}, \quad (2.37)$$

where $d$ is the spacing between the grating array waveguides, $y$ is the spacing between the centers of output ports, $L$ is the separation between center of arrayed waveguides and center of output waveguides, $n_c$ is the refractive index of waveguides in grating array, $n_s$ is the refractive index of star coupler, $n_g$ the group index, and $m$ is the diffraction order. The FSR can be obtained by [16]
\[ \Delta v_{\text{FSR}} = \frac{c}{n_g (\Delta L + d \sin \Theta_{\text{in},i} + d \sin \Theta_{\text{out},j})}, \]

where the diffraction angles from the \( i \)th input \( \Theta_{\text{in},i} \) and \( j \)th output \( \Theta_{\text{out},j} \) ports (measured from the center of the array) can be found as \( \Theta_{\text{in},i} = iy/L \), and \( \Theta_{\text{out},j} = jy/L \), respectively.

Given this description of basic building blocks used in the state-of-the-art optical communication systems, we turn our attention to the description of different direct detection schemes.

### 2.3 Direct Detection Modulation Schemes

Basic optical modulation formats can be categorized as follows (1) On–Off Keying (OOK), where the 1 is represented by the presence of the pulse while the 0 by the absence of a pulse; (2) Amplitude-shift keying (ASK), where the information is embedded in the amplitude of the sinusoidal pulse; (3) Phase-shift keying (PSK), where the information is embedded in the phase; (4) Frequency-shift keying (FSK), where the information is embedded in the frequency; and (5) Polarization-shift keying (PolSK), where the information is embedded in the polarization. In this section, we will discuss the modulation formats with direct detection, namely (1) non-return-to-zero (NRZ), (2) return-to-zero (RZ), (3) alternate mark inversion (AMI), (4) duobinary modulation, (5) carrier-suppressed RZ, (6) NRZ-differential phase-shift keying (NRZ-DPSK), and (7) RZ-differential phase-shift keying (RZ-DPSK).

#### 2.3.1 Non-Return-to-Zero

NRZ is considered to be the simplest modulation format. In early modulation phases, it was used due to its immunity to laser phase noise and for its low bandwidth requirements relative to other modulation formats. Recently, these issues are not the main concern with the advanced technology and high speed, and so, NRZ is merely used for comparison purposes with other formats.

The block diagram of an NRZ transmitter is shown in Fig. 2.20, and it is composed of a laser source and an external modulator. In our case, the modulator is chosen to be a Mach–Zehnder modulator (MZM). The output of the NRZ transmitter is shown in Fig. 2.21.

![Non-return-to-zero transmitter](image)

**Fig. 2.20** Non-return-to-zero transmitter. MZM
Mach–Zehnder modulator
2.3.2 Return-to-Zero

RZ is one of the basic subcategories of the OOK modulation format. Its bandwidth is higher than that of the corresponding NRZ at the same data rate due to the fact that in each “1” bit period $T$, the signal amplitude reaches the amplitude specified for the “1” bit and goes down to zero. RZ is characterized by the duty cycle, which is the ratio of the pulse width at $1/\sqrt{2}$ of the maximum amplitude to the bit period $T$.

Figure 2.22 shows the block diagram of the RZ transmitter which is basically an NRZ transmitter with an extra external modulator driven by an electrical clock that can be achieved by a sinusoidal signal at the half data rate. Figure 2.23 shows the RZ signal output of the modulator.

2.3.3 Alternate Mark Inversion

AMI utilizes amplitude to transmit information, the amplitude levels are either “on” or “off,” while the optical phase levels for the “on” state are either “0” or “$\pi$,” i.e., for a “0” bit, the output of the AMI modulator is “0”, while for a “1” bit, the output is either “1” or “$-1$.” In fact, the 1’s in AMI are represented in an alternating fashion, i.e., no two consecutive “1” bits have the same sign.

Figure 2.24 shows the block diagram of the AMI transmitter with the same binary input used for the previous example for NRZ and RZ (i.e., 01101001). As noticed, the AMI transmitter is an RZ transmitter that is driven by AMI-encoded data. The driving circuit consists of a precoder and an encoder that utilize delay
2.3 Direct Detection Modulation Schemes

![Figure 2.24 The block diagram of the AMI transmitter](image)

Table 2.1 AMI modulated signal generation and detection

<table>
<thead>
<tr>
<th>$m$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_m$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$P_m$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$B_m$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>−1</td>
<td></td>
</tr>
<tr>
<td>$</td>
<td>B_m</td>
<td>$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The precoder gives out $P_m$ by adding the input bit to the value stored in the delay line by a modulo-2 adder, while the encoder gives out $B_m$ by subtracting $P_{m-1}$ from $P_m$.

Table 2.1 shows an example of calculating $P_m$ and $B_m$ from the data input $D_m$. We assume that the initial values in the delay lines are set to “0,” so at $m = 0$, $P_0 = 0$, then we calculate $P_m$ by adding $P_{m-1}$ to $D_m$. $B_m$ on the other hand is calculated by subtracting $P_{m-1}$ from $P_m$. The original sequence is recovered by photodetector, which removes the phase information.

2.3.4 Duobinary Modulation Format

The optical duobinary signal has two intensity levels “on” and “off.” The “on” state signal can have one of the two optical phases, 0 and $\pi$. The two “on” states correspond to the logic states “1” and “−1” of the duobinary encoded signal, and the “off” state corresponds to the logic state “0” of the duobinary encoded signal. According to the duobinary encoding rule, the logic states “−1,” “0,” and “1” of the duobinary encoded signal correspond to the logic states “0,” “1,” and “0” of the original binary signal, respectively. Therefore, the original signal can be recovered by inverting the directly detected signal.

The schematic diagram of the transmitter for the optical duobinary signal format is shown in Fig. 2.25. The transmitter configuration of the optical duobinary signal format is identical to that of the RZ signal format. However, the signal that drives the data modulator is not the original NRZ data, but the duobinary-encoded data
sequence. The first step in the modulation process is to feed the original NRZ data sequence $D_m$ to the differential precoder in order to avoid error propagation at the receiver caused by the preceding received data not being recovered correctly. The precoder gives out $P_m$ by adding the input bit to the value stored in the delay line by a modulo-2 adder, while the encoder gives out $B_m$ by subtracting $P_{m-1}$ from $P_m$. Table 2.2 shows an example of generating $P_m$ and $B_m$ from the data input $D_m$, as well as recovery of the transmitted sequence on receiver side by photodetector and NOT gate.

### 2.3.5 Carrier-Suppressed Return-to-Zero

Carrier-suppressed return-to-zero (CS-RZ) modulation format was proposed by Sano and Miyamoto [25]. The major difference between a CS-RZ and a conventional RZ is that in CS-RZ optical signal there is a $\pi$ phase shift between adjacent bits, as shown in Fig. 2.26. Therefore, the average optical field in a CS-RZ signal is zero (there is no DC component). In the frequency domain, this translates into a carrier suppression in the optical spectrum. In general, the generation of a CS-RZ optical signal requires two electro-optic modulators. The first intensity modulator generates a conventional chirp-free RZ optical signal and the second electro-optic phase modulator produces a $\pi$ optical phase shift between adjacent adjacent.
bits. In this transmitter configuration, the modulation bandwidths of both of these two electro-optic modulators have to be in the same level as the data rate. The figure below shows one possible version of CS-RZ transmitter. In this configuration, the first intensity modulator encodes the NRZ data. The second Mach–Zehnder type intensity modulator is biased at the minimum power transmission point and driven by a sinusoid at the half data rate. The MZ intensity modulator biased at this condition performs frequency doubling for the modulating signal, and the output pulse train is phase alternated between adjacent bits. This configuration reduces the bandwidth requirement for the electro-optical modulators. In fact, the bandwidth required for the second modulator is only a half of the signal data rate.

CS-RZ is another modulation format considered to have better tolerance to fiber nonlinearity and residual chromatic dispersion. Phase alternating between adjacent bit slots reduces the fundamental frequency components to half of the data rate and regular RZ intensity bit pattern makes it easy to find the optimum dispersion compensation. In addition, carrier suppression reduces the efficiency of four-wave mixing (FWM) in WDM systems.
NRZ-DPSK eliminates the need for a coherent reference signal at receiver by combining two basic operations at transmitter (1) differential encoding and (2) phase-shift keying. To send symbol 0, we phase advance the current signal waveform by 180°, and to send a symbol 1, we leave the phase unchanged. The transmitter and receiver block diagrams are shown in Fig. 2.27a. The receiver "measures" the relative phase difference received during two successive bit intervals. The following differentially encoded rule applied: If the incoming binary symbol \( b_k \) is 1, leave the symbol \( d_k \) unchanged with respect to previous bit; if the incoming binary symbol \( b_k \) is 0, change the symbol \( d_k \) with respect to the previous bit (see Table 2.3 for an illustrative example).

At a DPSK optical receiver, shown in Fig. 2.27b, a one-bit-delay Mach–Zehnder interferometer (MZI), correlates each bit with previous bit and makes the phase-to-intensity conversion. When the two consecutive bits are in-phase, they are added constructively in the MZI and results in a high signal level. Otherwise, if there is a \( \pi \) phase difference between the two bits, they cancel each other in the MZI and results in a low signal level. In a practical DPSK receiver, the MZI has two balanced output ports (constructive port and destructive port).

### Table 2.3  NRZ-DPSK signal generation

| \( \{b_k\} \) | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| \( \{d_{k-1}\} \) | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| Differentially encoded sequence | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| \( \{d_k\}, d_k = b_k \oplus d_{k-1} \) | 0 | 0 | \( \pi \) | 0 | 0 | \( \pi \) | 0 | 0 |
| Transmitted phase (rad) | 0 | 0 | \( \pi \) | 0 | \( \pi \) | 0 | 0 | 0 |

#### 2.3.6 NRZ-DPSK

#### 2.3.7 RZ-DPSK

In order to improve system tolerance to nonlinear distortion and to achieve a longer transmission distance, return-to-zero DPSK (RZ-DPSK) has been proposed. In this modulation format, an optical pulse appears in each bit slot, with the binary data...
encoded as either a “0” or a “π” phase shift between adjacent bits. In general, the width of the optical pulses is narrower than the bit slot and therefore, the signal optical power returns to zero at the edge of each bit slot. In order to generate the RZDPSK optical signal, one more intensity modulator has to be used compared to the generation of NRZ-DPSK. The block diagram of a RZ-DPSK transmitter is shown in Fig. 2.28.

First, an electro-optic phase modulator generates a conventional NRZ-DPSK optical signal, and then, this NRZ-DPSK optical signal is sampled by a periodic pulse train at the clock rate through an electro-optic intensity modulator. In this modulation format, the signal optical intensity is no longer constant; this will probably introduce the sensitivity to self-phase modulation (SPM). In addition, due to the narrow pulse intensity sampling, the optical spectrum of RZ-DPSK is wider than a conventional NRZ-DPSK. Intuitively, this wide optical spectrum would make the system more susceptible to chromatic dispersion. However, in long distance optical systems, periodic dispersion compensation is often used and RZ modulation format makes it easy to find the optimum dispersion compensation because of its regular bit patterns.

2.4 Coherent Detection Modulation Schemes

In order to exploit the enormous bandwidth potential of the optical fiber, different multiplexing techniques (OTDMA, WDMA, CDMA, SCMA), modulation formats (OOK, ASK, PSK, FSK, PolSK, CPFSK, DPSK, etc.), demodulation schemes (DD or coherent), and technologies were developed; and some important aspects are discussed in this section. The coherent detection offers several important advantages compared to direct detection (1) improved receiver sensitivity, (2) better frequency selectivity, (3) possibility of using constant amplitude modulation formats (FSK, PSK), (4) tunable optical receivers similar to RF receivers are possible, and (5) with coherent detection the chromatic dispersion and PMD can easier be mitigated.

Based on whether local laser operating frequency coincides with incoming optical signal, the coherent reception can identified as (1) homodyne, in which the
frequencies are the same or (2) heterodyne, in which the frequencies are different so that all related signal processing upon photodetection is performed at suitable intermediate frequency (IF). Different coherent detection can be classified into following categories (1) synchronous (PSK, FSK, ASK), (2) asynchronous (FSK, ASK), (3) differential detection (CPFSK, DPSK), (4) phase diversity receivers, (5) polarization diversity receivers, and (6) polarization multiplexing schemes. Synchronous detection schemes can further be categorized as residual carrier or suppressed carrier [Costas loop/decision-driven loop (DDL)] based.

In Fig. 2.29, we show the basic difference between direct detection (Fig. 2.29a) and coherent detection (Fig. 2.29b) schemes. Coherent detection, in addition to the photodetector and integrator already in use for direct detection, employs a local laser whose frequency and polarization is matched to the frequency and polarization of incoming optical signal. The incoming optical signal and local laser output signal are mixed in optical domain (a mirror is used in this illustrative example, in practice an optical hybrid is used instead).

Because the photodetector generates the photocurrent in proportion to the input optical power, the direct detection integrator output signal can be written as follows:

\[
\begin{align*}
 s_0'(t) &= 0; \\
 s_1'(t) &= R P_s; \\
\end{align*}
\]

(2.39)

where \( P_s \) is the input optical signal power, \( R \) is the photodiode responsivity introduced earlier, and \( T \) is the bit duration. The subscripts 0 and 1 are used to denote the transmitted bits (zero and one). It is very common in optical communications to express the photocurrent output signal in terms of number of photons per bit \( n_p = R P_s T/q \) (\( q \) is an electron charge) as given below:

\[
\begin{align*}
 s_0(t) &= 0; \\
 s_1(t) &= \frac{n_p}{T}; \\
\end{align*}
\]

(2.40)

The corresponding coherent detector integrator output signal, for homodyne synchronous detection, can be written as

\[
\begin{align*}
 s_{0,1}(t) &= \frac{1}{2T} \left( \mp \sqrt{2n_p} + \sqrt{2n_{LO}} \right)^2; \\
 0 &\leq t \leq T.
\end{align*}
\]

(2.41)
where \( n_{\text{LO}} \) denotes the average number of local laser photons per bit, sign “-” corresponds to transmitted zero, and sign “+” to transmitted one bits. The optimum coherent detection receiver (minimizing the bit error probability \( P_b \), shown in Fig. 2.30, is the matched filter (or correlation) receiver with impulse response given by \( h(t) = s_1(T-t) - s_0(T-t) \). [26]

General expression for matched filter output signal, applicable to different modulation formats (such as ASK, FSK, PSK) and shot-noise-dominated scenario, can be written as follows:

\[
s_{0,1}(t) = 2R \sqrt{P_s P_{\text{LO}}} \cos(\omega_{0,1}(t) + \theta_{0,1}(t)) + n_{\text{sh}}(t); \quad 0 \leq t \leq T, \tag{2.42}
\]

where \( P_{\text{LO}} \) denotes the local laser output signal, while \( \omega_{0,1} \) and \( \theta_{0,1} \) represent frequency and phase corresponding to bit 0 and 1. \( n_{\text{sh}}(t) \) denotes the shot noise process, which commonly can be modeled as zero-mean Gaussian with power spectral density \( N_0' = 2RqP_{\text{LO}} \). Because the bit energy can be determined by \( E' = E_{0,1}' = \int_0^T s_{0,1}^2(t) dt \approx 2R^2 P_s P_{\text{LO}} T \), the corresponding SNR is related to number of photons per bit as given below:

\[
\frac{E'}{N_0'} = \frac{2R^2 P_s P_{\text{LO}} T}{2RqP_{\text{LO}}} = \frac{RP_s T}{q} = n_p. \tag{2.43}
\]

Therefore, the general expression (2.42) can also be written in terms of \( n_p \):

\[
s_{0,1}(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_{0,1}(t) + \theta_{0,1}(t)); \quad 0 \leq t \leq T. \tag{2.44}
\]

Correlation coefficient \( \rho \) and Euclidean distance \( d \) between transmitted signals \( (s_0 \text{ and } s_1) \) are defined as

\[
\rho = \frac{1}{\sqrt{E_0 E_1}} \int_0^T s_0(t)s_1(t)dt; \quad d^2 = E_0 + E_1 - 2\rho\sqrt{E_0 E_1}, \tag{2.45}
\]

where \( E_i \) is the energy of \( i \) th \((i = 0, 1)\) bit. The probability of error is related to the Euclidean distance as follows:

\[
P_b = \frac{1}{2}\text{erfc}\left(\frac{d}{2\sqrt{N_0}}\right). \tag{2.46}
\]
For ASK systems, the transmitted signals can be written in terms of $n_p$ by

$$s_1(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_1 t); \quad s_0(t) = 0; \quad 0 \leq t \leq T \quad (2.47)$$

By using (2.45), the Euclidean distance squared is found to be $d^2 = n_p$ so that the corresponding expression for bit error probability is obtained from (2.46) as follows:

$$P_b = \frac{1}{2} \text{erfc} \left( \frac{\sqrt{n_p}}{2} \right). \quad (2.48)$$

The bit error probability of $10^{-9}$ is achieved for 72 photons per bit.

For continuous phase FSK (CPFSK) systems, the transmitted symbols can be represented by

$$s_1(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_1 t); \quad 0 \leq t \leq T,$$

$$s_0(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_0 t); \quad 0 \leq t \leq T. \quad (2.49)$$

The corresponding correlation coefficient can be found by

$$\rho = \frac{2}{T} \int_0^T \cos \omega_0 t \cos \omega_1 t dt \approx \frac{\sin 2\pi m}{2\pi m}, \quad m = \frac{\left| \omega_1 - \omega_0 \right|}{2\pi/T}, \quad (2.50)$$

where $m$ is the modulation index. The corresponding bit error probability is obtained by substituting (2.50) in (2.46) by

$$P_b = \frac{1}{2} \text{erfc} \left( \frac{n_p}{2} \left( 1 - \frac{\sin 2\pi m}{2\pi m} \right) \right). \quad (2.51)$$

For $m = 0.5p$ ($p = 1, 2, \ldots$), the number of required photons per bit to achieve $P_b$ of $10^{-9}$ is 36, while the minimum $n_p = 29.6$ is obtained for $m = 0.715$.

For direct modulation PSK (DM-PSK) systems, the transmitted symbols can be represented by

$$s_1(t) = \sqrt{\frac{2n_p}{T}} \cos(\theta_1(t)); \quad 0 \leq t \leq T; \quad \theta_1(t) = \begin{cases} \omega_1 t + \frac{\pi m}{T} & 0 \leq t \leq T/(2m) \\ \omega_1 t + \pi/2 & T/(2m) \leq t \leq T \end{cases},$$

$$s_0(t) = \sqrt{\frac{2n_p}{T}} \cos(\theta_0(t)); \quad 0 \leq t \leq T; \quad \theta_0(t) = \begin{cases} \omega_1 t - \frac{\pi m}{T} & 0 \leq t \leq T/(2m) \\ \omega_1 t - \pi/2 & T/(2m) \leq t \leq T \end{cases}. \quad (2.52)$$
where $\omega_F = |\omega_s - \omega_{LO}|$ is the intermediate frequency. The correlation coefficient is obtained from (2.45) as $\rho \approx 1/(2m) - 1$, and corresponding probability of error is given by

$$P_b = \frac{1}{2} \text{erfc} \left( \sqrt{\frac{n_p}{4m} \left( 1 - \frac{1}{4m} \right)} \right).$$

For $m = 1/2$, the required number of photons per bit to achieve $P_b$ of $10^{-9}$ is 36 (that is the same as in FSK systems), while when $m \to \infty$ the required $n_p$ is 18.

For FSK systems, with two oscillators the transmitted signals are represented by

$$s_1(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_1 t + \theta); \quad 0 \leq t \leq T,$$

$$s_0(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_0 t + \theta); \quad 0 \leq t \leq T,$$

$$\theta = \begin{cases} [\pi - \pi m] \pi; & 2p \leq m \leq 2p + 1 \\ [-\pi m] \pi; & 2p + 1 \leq m \leq 2(p + 1) \end{cases} \quad p = 0, 1, 2, \ldots$$

(We use $[x]_\pi$ to denote mod $\pi$ operation.) The correlation coefficient is obtained by substituting (2.54) in (2.45) $\rho \approx (\sin(2\pi m + \theta) - \sin \theta)/(2\pi m)$, while probability of error expression is obtained by

$$P_e = \frac{1}{2} \text{erfc} \left( \sqrt{\frac{n_p}{2} \left( 1 + \frac{|\sin \pi m|}{\pi m} \right)} \right).$$

The bit error probability $P_b$ of $10^{-9}$ is achieved for $n_p = 36/(1 + |\sin \pi m|/\pi m)$. For example, $n_p = 18$ when $m \to 0$, while $n_p = 36$ when $m = 1, 2, \ldots$. In Fig. 2.31, we provide the comparison of different FSK modulation formats in terms of required $n_p$ to achieve $P_b$ of $10^{-9}$ vs. modulation index $m$.

For PSK systems, the transmitted signal can be written by

$$s_1(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_1 t); \quad 0 \leq t \leq T,$$

$$s_0(t) = \sqrt{\frac{2n_p}{T}} \cos(\omega_1 t + \pi); \quad 0 \leq t \leq T.$$}

The corresponding probability of error is given by

$$P_e = \frac{1}{2} \text{erfc} \left( \sqrt{\frac{n_p}{2}} \right).$$

The bit error probability $P_b$ of $10^{-9}$ is achieved for $n_p = 18$ in heterodyne case and $n_p = 9$ in homodyne case. In Fig. 2.32, we provide the bit error probabilities
Fig. 2.31 Comparison of FSK systems

![Comparison of FSK systems](image1)

![Comparison of FSK systems](image2)

Fig. 2.32 BER performance comparison

For different modulation schemes discussed above. Similarly in conventional digital communications systems, the synchronous coherent detection PSK scheme performs the best. The comparison of different modulation formats in terms of receiver sensitivity, defined as required number of photons per bit $n_p$ to achieve BER of $10^{-9}$, is given in Table 2.4, which is adopted from [27]. For more details on various modulation schemes for coherent detection, an interested reader is referred to an excellent book due to Jacobsen [27].
### Table 2.4 Receiver sensitivities in terms of required $\eta_p$ to achieve BER of $10^{-9}$

<table>
<thead>
<tr>
<th>System</th>
<th>Coherent heterodyne</th>
<th>Coherent homodyne</th>
<th>IM/DD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantum limit</td>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>Super quantum limit</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>ASK</td>
<td>Matched filter</td>
<td>72</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Asynchronous</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>CPFSK $m = 0.5, 1.5, \ldots$</td>
<td></td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Matched filter $m = 0.715$</td>
<td></td>
<td>29.6</td>
<td></td>
</tr>
<tr>
<td>FSK $m = 0$</td>
<td></td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>Two oscillators $m = 1.2, \ldots$</td>
<td></td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Matched filter $m = 1.43$</td>
<td></td>
<td>29.6</td>
<td></td>
</tr>
<tr>
<td>CPFSK $m = 0.5$</td>
<td></td>
<td>61.9</td>
<td>30.9</td>
</tr>
<tr>
<td>Delay-detection $m = 0.8$</td>
<td></td>
<td>36.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$m = 1$</td>
<td>43.6</td>
<td></td>
</tr>
<tr>
<td>FSK Asynchronous $m = 1.5$</td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>DM-PSK $m = 0.5$</td>
<td></td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>Matched filter $m = 1$</td>
<td></td>
<td>24</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>$m = 2$</td>
<td>20.6</td>
<td>10.3</td>
</tr>
<tr>
<td>DM-DPSK $m = 0.5$</td>
<td></td>
<td>61.9</td>
<td>30.9</td>
</tr>
<tr>
<td>Delay-detection $m = 1$</td>
<td></td>
<td>30.6</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>$m = 2$</td>
<td>24.4</td>
<td>12.2</td>
</tr>
<tr>
<td>PolSK Matched filter</td>
<td></td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>PolSK (and FSK) $m = 0.5, 1.5, \ldots$</td>
<td></td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Matched filter $m = 0.715$</td>
<td></td>
<td>59.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$m = 1.43$</td>
<td>59.2</td>
<td></td>
</tr>
<tr>
<td>PolSK asynchronous</td>
<td></td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>PSK</td>
<td></td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>Matched filter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPSK</td>
<td></td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>DPSK IF = $1/(2T)$</td>
<td></td>
<td>24.7</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.4.1 Optical Hybrids and Balanced Receivers

So far we used an optical mirror to perform optical mixing before photodetection takes place. In practice, this operation is performed by four-port device known as optical hybrid, which is shown in Fig. 2.33.

Electrical fields at output ports $E_{1o}$ and $E_{2o}$ are related to the electrical fields at input ports $E_{1i}$ and $E_{2i}$ as follows:

$$E_{1o} = (E_{1i} + E_{2i}) \sqrt{1 - k},$$

$$E_{2o} = (E_{1i} + E_{2i} \exp(-j\phi)) \sqrt{k},$$

(2.58)
where $k$ is the power splitting ratio and $\phi$ is the phase shift introduced by the phase trimmer (see Fig. 2.33). Equation (2.58) can also be written in terms of scattering (S-) matrix as follows:

$$
E_o = \begin{bmatrix} E_{o1} \\ E_{o2} \end{bmatrix} = SE_i, \quad S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} \sqrt{1-k} & \sqrt{1-k} \\ \frac{k}{\sqrt{k}} & e^{-j\phi} \frac{k}{\sqrt{k}} \end{bmatrix}.
$$

In expressions (2.58) and (2.59), we assumed that hybrid is lossless device, which leads to: $s_{11} = |s_{11}|$, $s_{12} = |s_{12}|$, $s_{21} = |s_{21}|$, and $s_{22} = |s_{22}| \exp(j\theta_{22})$. Popular hybrids are $\pi$ hybrid, which S-matrix can be written as (by setting $\phi = \pi$ in (2.59))

$$
S = \begin{bmatrix} \sqrt{1-k} & \sqrt{1-k} \\ \frac{k}{\sqrt{k}} & -j\frac{k}{\sqrt{k}} \end{bmatrix}.
$$

and $\pi/2$ hybrid, which S-matrix can be written as

$$
S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22}e^{-j\pi/2} \end{bmatrix}.
$$

Well-known $\pi$ hybrid is 3-dB coupler ($k = 1/2$). If $\pi/2$ hybrid is symmetric $|s_{ij}| = 1/L (\forall i, j)$, the phase difference between the input electrical fields $E_{1i} = |E_{1i}|$, $E_{2i} = |E_{2i}|e^{j\theta_i}$ can be chosen in such a way that total output ports power

$$
E_o^\dagger E_o = \frac{2}{L} \left[ P_{1i} + P_{2i} + \sqrt{P_{1i}P_{2i}}(\cos \theta_i + \sin \theta_i) \right]
$$

is maximized. (We use $\dagger$ to denote Hermitian transposition–simultaneous transposition and complex conjugation.) For equal input powers, the maximum of (2.61) is obtained when $\theta_i = \pi/4$, leading to $L \geq 2 + \sqrt{2}$. The corresponding loss is $10 \log_{10}(L/2) = 3.22 \text{ dB}$. For Costas loop- and DDL-based homodyne systems, there exists optimum $k$ in corresponding S-matrix:

$$
S = \frac{1}{\sqrt{L}} \begin{bmatrix} \sqrt{1-k} & \sqrt{1-k} \\ \frac{k}{\sqrt{k}} & -j\frac{k}{\sqrt{k}} \end{bmatrix}.
$$
To reduce the relative intensity noise (RIN) of transmitting laser and to eliminate the direct-detection and signal-cross-signal interferences, the most deleterious sources for multichannel applications, the balanced receiver, shown in Fig. 2.34, is commonly used. The upper and lower photodetector’s output currents can, respectively, be written as

\[ i_1(t) = R|E_1|^2 = \frac{1}{2} R \left( P_s + P_{LO} + 2\sqrt{P_s P_{LO}} \cos \theta_s \right) + n_1(t). \]
\[ i_2(t) = R|E_2|^2 = \frac{1}{2} R \left( P_s + P_{LO} - 2\sqrt{P_s P_{LO}} \cos \theta_s \right) + n_2(t), \quad (2.62) \]

where \( \theta_s \) is the phase of incoming optical signal and \( n_i(t) (i = 1, 2) \) is the \( i \)th photodetector shot noise process of PSD \( S_{n_i} = qR|E_i|^2 \). The balanced receiver output current (see Fig. 2.34) can be written as

\[ i(t) = i_1(t) - i_2(t) = 2R \sqrt{P_s P_{LO}} \cos \theta_s + n(t), \quad n(t) = n_1(t) - n_2(t), \quad (2.63) \]

where \( n(t) \) is a zero-mean Gaussian process of PSD \( S_n = S_{n1} + S_{n2} = qR(P_s + P_{LO}) \approx qRP_{LO} \). For binary PSK signaling, \( \theta_s = \pm \pi \) and expression (2.63) is consistent with (2.56).

### 2.4.2 Dominant Coherent Detector Noise Sources

In this section, we describe the basic coherent detector noise processes, and a more detailed description of different channel impairments and receiver processes is provided in Chap. 3. The dominant sources of performance degradation in coherent detection are the laser-phase noise, photodiode shot noise, polarization noise, data-to-phase lock crosstalk for receiver with residual carrier, and adjacent channel interference for multichannel applications.

The PIN photodiode shot noise is zero-mean white Gaussian with PSD \( \text{PSD}_n(f) = qR(P_s + P_{LO}) \approx qRP_{LO} \) if the following is valid

\[ \text{Bandwidth of interest} \ll \frac{1}{T_d} \ll \frac{hP_s}{\hbar v}. \]

Noise, especially spontaneous emission, causes phase fluctuations in lasers, leading to a nonzero spectral linewidth \( \Delta v \), which is illustrated in Fig. 2.35. In gas or solid-state lasers, the linewidth \( \Delta v \) typically ranges from the subhertz to the kilohertz.
range. In semiconductor lasers, the linewidth $\Delta \nu$ is often much larger, up to the megahertz range, because a large number of photons are stored in the small cavity and due to the nonnegligible value of the linewidth enhancement factor.

The laser lightwave signal can be written as

$$x(t) = \sqrt{P_S} e^{\tilde{\phi}(t)+\phi(t)+\theta},$$

where $\phi(t)$ is the laser phase noise process, which is commonly modeled as a Wiener–Lévy process [7], that is a zero-mean Gaussian process with variance $\sigma^2 = 2\pi \Delta \nu |\tau|$. The autocorrelation function of $x(t)$ is given by

$$R_x(\tau) = P_S e^{j\omega_0 \tau} e^{-\pi \Delta \nu |\tau|}.$$  \hspace{1cm} (2.65)

The corresponding PSD of $x(t)$ can be found as Fourier transform of (2.65)

$$PSD_x(\omega) = \frac{2P_S}{\pi \Delta \nu} \left[ 1 + \left( \frac{\omega - \omega_0}{\pi \Delta \nu} \right)^2 \right]^{-1}$$

and has therefore the Lorentzian shape.

The influence of laser phase noise for BPSK signaling is illustrated in Fig. 2.36. The effect on BER curves is twofold (1) the BER curves are shifted to the right and (2) BER floor appears. For state-of-the-art optical communication systems, the laser phase noise does impose a serious limitation. However, for large medium and large multilevel modulations, such as $M$-ary PSK and $M$-ary QAM, the laser phase noise is a very important factor of performance degradation.

The polarization noise comes from discrepancies of the state of polarization (SOP) of incoming optical signal and local laser signal. Different polarization noise avoidance techniques can be classified as follows (1) polarization control, (2) polarization maintenance fibers, (3) polarization scrambling, (4) polarization diversity, and (5) polarization multiplexing. The polarized electromagnetic field launched into the fiber can be represented as

$$E(t) = \begin{bmatrix} e_x(t) \\ e_y(t) \end{bmatrix} e^{j\omega_0 t},$$

(2.67)
where \( e_x \) and \( e_y \) represent two orthogonal SOP components and \( \omega_c \) is the carrier frequency. The received field can be represented by

\[
E_s(t) = H' \begin{bmatrix} e_x(t) \\ e_y(t) \end{bmatrix} e^{j\omega_c t},
\]

where \( H' \) is the Jones matrix of birefringence. The additional transformation is needed to match the SOPs of local laser with that of incoming optical signal:

\[
E'_s(t) = H'' H' \begin{bmatrix} e_x(t) \\ e_y(t) \end{bmatrix} e^{j\omega_c t} = H \begin{bmatrix} e_x(t) \\ e_y(t) \end{bmatrix} e^{j\omega_c t}, \quad H = H'' H'.
\] (2.69)

The SOP of local laser in Stokes coordinates can be represented by \( S_{LO} = (S_{1,LO} \ S_{2,LO} \ S_{3,LO}) \). The heterodyning is possible only if \( S_{LO} = S_R \), where \( S_R \) is the SOP of received signal. The action of birefringence corresponds to rotating the point, which represents the launched SOP, on the surface of the Poincaré sphere. This rotation can be represented as product of three matrices, each of them corresponding to the rotation of reference axes in the Stokes space around axes \( s_1 \), \( s_2 \), and \( s_3 \) for \( \alpha \), \( \beta \), and \( \gamma \), respectively [9]:

\[
H(\alpha, \beta, \gamma) = H_1(\alpha)H_2(\beta)H_3(\gamma).
\]

\[
H_1(\alpha) = \begin{bmatrix} e^{j\alpha/2} & 0 \\ 0 & e^{-j\alpha/2} \end{bmatrix}, \quad H_2(\beta) = \begin{bmatrix} \cos(\beta/2) & j\sin(\beta/2) \\ j\sin(\beta/2) & \cos(\beta/2) \end{bmatrix},
\]

\[
H_3(\gamma) = \begin{bmatrix} \cos(\gamma/2) & \sin(\gamma/2) \\ -\sin(\gamma/2) & \cos(\gamma/2) \end{bmatrix}.
\] (2.70)
If we assume that the SOP of LO is aligned with $s_1$-axis in Stokes space we can write
\[ \begin{bmatrix} e'_x(t) \\ e'_y(t) \end{bmatrix} = H_1(\phi) H_2(\theta) \begin{bmatrix} e_x \\ e_y \end{bmatrix}. \] \tag{2.71}

The ratio between the power heterodyned component and the total power is
\[ p(\theta, \phi) = \frac{P_{\text{het}}}{P_{\text{tot}}} = \frac{|e''_x|^2}{|e''_x|^2 + |e''_y|^2} = \left| e^{j\phi/2} \cos \frac{\theta}{2} \right|^2 = \frac{1}{2} (1 + \cos \theta). \] \tag{2.72}

Probability density function of $\theta$ is given below [9]:
\[ \text{PDF}(\theta) = \frac{\sin \theta}{2} e^{-\frac{A^2}{4\sigma^2}(1-\cos \theta)} \left[ 1 + \frac{A^2}{4\sigma^2} (1 + \cos \theta) \right], \quad \theta \in [0, \pi]; \]
\[ A = 2R \sqrt{P_s P_{LO}}, \quad \sigma^2 = qR_{LO}. \] \tag{2.73}

The typical optical receiver is based on transimpedance FET stage, as illustrated in Fig. 2.37. The PSD of receiver noise can be written as [28]
\[ \text{PSD}_{\text{rec}}(f) = \frac{4k_B T_a}{R_f} + 4k_B T_a g_m \Gamma \left( \frac{f}{f_{\text{f,eff}}} \right)^2, \quad f_{\text{f,eff}} = g_m/2\pi C_T, \quad C_T = C_i + C_{\text{PIN}}, \] \tag{2.74}

where $g_m$ is transconductance, $\Gamma$ is the FET channel-noise factor, and $C_T$ is total capacitance (FET input capacitance $C_i$ plus PIN photodetector capacitance $C_{\text{PIN}}$). ($R_L$ is the load resistor, $R_f$ is feedback resistor, $T_a$ is absolute temperature, and $k_B$ is the Boltzmann constant.)

This amplifier stage is popular because it has large gain–bandwidth product ($G_B$) defined below as [28]:
\[ G_B = \left| -R_f \frac{g_m R_L}{1 + g_m R_L} \right| \frac{1 + g_m R_L}{2\pi C_T (R_f + R_i)} = f_{\text{f,eff}}(R_f || R_L). \] \tag{2.75}

The intensity noise comes from the variation of optical power of transmitting laser $P = \langle P \rangle + \Delta P$ ($\langle \cdot \rangle$ denotes the statistical averaging). The RIN is defined as

\[ \text{RIN} = \frac{\langle \Delta P^2 \rangle}{\langle P \rangle^2} = \frac{1}{2} \left( \frac{\Delta f}{f_0} \right)^2, \]
as \( \text{RIN} = \langle \Delta P^2 \rangle / \langle P \rangle^2 \). Because the power of transmitting laser fluctuates, the photocurrent fluctuates as well, \( I = \langle I \rangle + \Delta I \). The corresponding shot noise PSD can be determined as \( 2q \langle I \rangle = 2qR(P) \). The intensity noise PSD is simply \( \langle \Delta I^2 \rangle = R^2 \text{RIN}(P)^2 \). The SNR in the presence of shot, receiver, and RIN can be determined as signal power over total noise power:

\[
\text{SNR} = \frac{2R^2 P_s P_{LO}}{2qRP_{LO}B + \text{RIN} \cdot R^2 P_{LO}^2 B + \text{PSD}_{\text{rec}} B^2}, \tag{2.76}
\]

where \( B \) is the receiver bandwidth, and \( \text{PSD}_{\text{rec}} \) is introduced by (2.74). The SNR for balanced reception can be written as

\[
\text{SNR} = \frac{2R^2 P_s P_{LO}}{2qRP_{LO}B + \frac{\text{RIN}}{\text{CMRR}} R^2 P_{LO}^2 B + 2\text{PSD}_{\text{rec}} B^2}, \tag{2.77}
\]

where CMRR is the common mode rejection ratio of FET. Therefore, the RIN of balanced receiver is significantly reduced by balanced detection.

### 2.4.3 Homodyne Coherent Detection

Homodyne coherent detection receiver can be classified as either residual carrier receivers or suppressed carrier receivers [29]. In systems with residual carrier, shown in Fig. 2.38, the phase deviation between the mark- and space-state bits is less than \( \pi / 2 \) rad, so that the part of transmitted signal power is used for the nonmodulated carrier transmission and as a consequence some power penalty occurs.

Costas loop and DDL [29] based receivers, shown in Fig. 2.39, are two alternatives to the receivers with residual carrier. Both these alternatives employ a fully suppressed carrier transmission, in which the entire transmitted power is used for data transmission. However, at the receiver side a part of the power is used for the carrier extraction, so some power penalty is incurred with this approach too.

![Fig. 2.38 Balanced-loop-based receiver. PD Photodetector](image)
2.4.4 Phase Diversity Receivers

The general architecture of a multiport homodyne receiver is shown in Fig. 2.40a [9]. The incoming optical signal and local laser output signal can be written as

\[ S(t) = aE_s e^{\omega_0 t + \phi_s(t)}, \quad L(t) = E_{LO} e^{\omega_0 t + \phi_{LO}(t)}, \]  

(2.78)

where the information is imposed either in amplitude \(a\) or phase \(\phi_s\). Both incoming optical signal \(S\) and local laser output signal \(L\) are used as inputs to \(N\) output ports of an optical hybrid, which introduces fixed phase difference \(k(2\pi/N)\) \((k = 0, 1, \ldots, N-1)\) between the ports, so that the output electrical fields can be written as

\[ E_k(t) = \frac{1}{\sqrt{N}} \left[ S(t) e^{\frac{2\pi i}{N} k} + L(t) \right]. \]  

(2.79)
The corresponding photodetector outputs are as follows:

\[ i_k(t) = R|E_k(t)| + i_{nk}(t) \]

\[ = \frac{R}{N} \left\{ P_{LO} + aP_s + 2a \sqrt{P_sP_{LO}} \cos \left[ \phi_s(t) - \phi_{LO}(t) + k \frac{2\pi}{N} \right] \right\} + i_{nk}(t). \]

(2.80)

where \( i_{nk}(t) \) is the \( k \)th photodetector shot noise. Different versions of demodulators for ASK, DPSK, and DPFSK are shown in Fig. 2.40b. For ASK, we simply have to square photodetector outputs and add them together:

\[ y = \sum_{k=1}^{N} i_k^2. \]

(2.81)

### 2.4.5 Polarization Control and Polarization Diversity

The coherent receivers require matching the SOP of the local laser with that of the received optical signal. In practice, only the SOP of local laser can be controlled, and one possible polarization control receiver configuration is shown in Fig. 2.41. Polarization controller is commonly implemented by using four squeezers as shown in [9].
Insensitivity with respect to polarization fluctuations is possible if the receiver derives two demodulated signals from two orthogonal polarizations of the received signal, which is illustrated in Fig. 2.42. This scheme is known as polarization diversity receiver. In polarization diversity receivers, however, only one polarization is used and spectral efficiency is reduced. To double the spectral efficiency of polarization diversity schemes, the polarization multiplexing is advocated in [30–33].

### 2.4.6 Polarization Multiplexing and Coded Modulation

In polarization multiplexing [30–35], both polarizations carry independent multi-level modulated streams, which is illustrated in Fig. 2.43. $M$-ary PSK, $M$-ary QAM, and $M$-ary DPSK achieve the transmission of $\log_2 M (= m)$ bits per symbol, providing bandwidth-efficient communication. In coherent detection, the data phasor $\phi_i \in \{0, 2\pi/M, \ldots, 2\pi(M - 1)/M\}$ is sent at each $l$th transmission interval. In direct detection, the modulation is differential, the data phasor $\phi_i = \phi_{i-1} + \Delta\phi_i$ is sent instead, where $\Delta\phi_i \in \{0, 2\pi/M, \ldots, 2\pi(M - 1)/M\}$ is determined by the sequence of $\log_2 M$ input bits using an appropriate mapping rule. Let us now introduce the transmitter architecture employing forward error correction (FEC) codes. More details about coded modulation can be found in Chap. 6. If the component FEC codes are of different code rates but of the same length, the corresponding scheme is commonly referred to as multilevel coding (MLC) [34]. If all component codes are of the same code rate, corresponding scheme is referred to as the bit-interleaved coded modulation (BICM) [35]. The use of MLC allows us to adapt the code rates to the constellation mapper and channel. In MLC, the bit streams originating from $m$ different information sources are encoded using different $(n, k_i)$ FEC codes of code rate $r_i = k_i/n$. $k_i$ denotes the number of information bits of the $i$th ($i = 1, 2, \ldots, m$) component FEC code and $n$ denotes the codeword length, which
is the same for all FEC codes. The mapper accepts $m$ bits, $c = (c_1, c_2, \ldots, c_m)$, at time instance $i$ from the $(m \times n)$ interleaver column-wise and determines the corresponding $M$-ary ($M = 2^m$) constellation point $s_i = (I_i, Q_i) = |s_i| \exp(i\phi_i)$ (see Fig. 2.43a). Two dual-drive MZMs are needed, one for each polarization. The outputs of the MZMs are combined using the polarization beam combiner (PBC). The same DFB laser is used as CW source, with $x$- and $y$-polarization being separated by a polarization beam splitter (PBS).

The coherent detector receiver architecture is shown in Fig. 2.43b. The balanced outputs of I- and Q-channel branches for $x$-polarization at time instance $l$ can be written as

$$
\begin{align*}
V_{I,l}^{(x)} &= R \left| S_I^{(x)} \right| \left| L^{(x)} \right| \cos \left( \varphi_I^{(x)} + \varphi_{S,PN}^{(x)} - \varphi_{L,PN}^{(x)} \right), \\
V_{Q,l}^{(x)} &= R \left| S_I^{(x)} \right| \left| L^{(x)} \right| \sin \left( \varphi_I^{(x)} + \varphi_{S,PN}^{(x)} - \varphi_{L,PN}^{(x)} \right),
\end{align*}
$$

(2.82)

where $R$ is photodiode responsivity while $\varphi_{S,PN}$ and $\varphi_{L,PN}$ represent the laser phase noise of transmitting and receiving (local) laser, respectively. $S_I^{(x)}$ and $L^{(x)}$ represent the receiver incoming signal in $x$-polarization and local laser $x$-polarization signal, respectively. Similar expressions hold for $y$-polarization. In symbol detection block, the PMD is compensated for using one of the following possible approaches (1) blind equalization [30], (2) polarization-time coding [32] similar to space–time coding proposed for use in multi-input multi-output (MIMO) wireless
communication systems, (3) using BLAST algorithm [31], (4) by polarization interference cancellation scheme [31], or (5) carefully performed channel matrix inversion [35].

For soft decoding, the a posteriori probability (APP) demapper and bit log-likelihood ratios (LLRs) calculation blocks operate in similar fashion as described in [31–35]. The APP and LLRs calculation block are optional; they are not needed if hard decision decoding is used.

2.5 Summary

This chapter is devoted to the description of basic concepts of optical transmission systems based on both intensity modulation with direct detection and coherent detection. In Sect. 2.2, basic principles of optical transmission are provided, the basic building blocks are identified, and fundamental principles of those building blocks are described. Section 2.3 is devoted to the description of basic direct detection modulations schemes such as NRZ, RZ, AMI, duobinary modulation, carrier-suppressed RZ, NRZ-DPSK, and RZ-DSPK.

In Sect. 2.4, the basic concepts of coherent detection are introduced; including description and comparison of different coherent detection schemes for shot-noise-dominated scenario, description of dominant receiver noises, description of homodyne detection principle, phase diversity and polarization diversity principles, and polarization multiplexing.

References

References

Coding for Optical Channels
Djordjevic, I.; Ryan, W.; Vasic, B.
2010, XV, 444 p., Hardcover