3 chapter

Components

In this chapter, we will discuss the physical principles behind the operation of the most important components of optical communication systems. For each component, we will give a simple descriptive treatment followed by a more detailed mathematical treatment.

The components used in modern optical networks include couplers, lasers, photodetectors, optical amplifiers, optical switches, and filters and multiplexers. Couplers are simple components used to combine or split optical signals. After describing couplers, we will cover filters and multiplexers, which are used to multiplex and demultiplex signals at different wavelengths in WDM systems. We then describe various types of optical amplifiers, which are key elements used to overcome fiber and other component losses and, in many cases, can be used to amplify signals at multiple wavelengths. Understanding filters and optical amplifiers is essential to understanding the operation of lasers, which comes next. Semiconductor lasers are the main transmitters used in optical communication systems. Then we discuss photodetectors, which convert the optical signal back into the electrical domain. This is followed by optical switches, which play an important role as optical networks become more agile. Finally, we cover wavelength converters, which are used to convert signals from one wavelength to another, at the edges of the optical network, as well as inside the network.

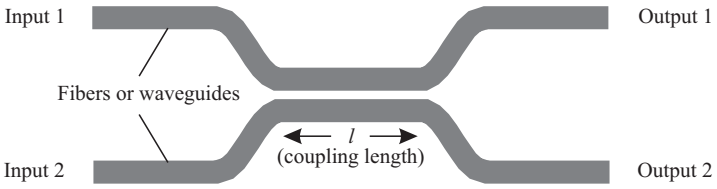

Figure 3.1 A directional coupler. The coupler is typically built by fusing two fibers together. It can also be built using waveguides in integrated optics.

3.1 Couplers

A *directional coupler* is used to combine and split signals in an optical network. A 2×2 coupler consists of two input ports and two output ports, as is shown in Figure 3.1. The most commonly used couplers are made by fusing two fibers together in the middle—these are called fused fiber couplers. Couplers can also be fabricated using waveguides in integrated optics. A 2×2 coupler, shown in Figure 3.1, takes a fraction α of the power from input 1 and places it on output 1 and the remaining fraction $1 - \alpha$ on output 2. Similarly, a fraction $1 - \alpha$ of the power from input 2 is distributed to output 1 and the remaining power to output 2. We call α the coupling ratio.

The coupler can be designed to be either wavelength selective or wavelength independent (sometimes called wavelength flat) over a usefully wide range. In a wavelength-independent device, α is independent of the wavelength; in a wavelength-selective device, α depends on the wavelength.

A coupler is a versatile device and has many applications in an optical network. The simplest application is to combine or split signals in the network. For example, a coupler can be used to distribute an input signal equally among two output ports if the coupling length, l in Figure 3.1, is adjusted such that half the power from each input appears at each output. Such a coupler is called a 3 dB coupler. An $n \times n$ star coupler is a natural generalization of the 3 dB 2×2 coupler. It is an n-input, n-output device with the property that the power from each input is divided equally among all the outputs. An $n \times n$ star coupler can be constructed by suitably interconnecting a number of 3 dB couplers, as shown in Figure 3.2. A star coupler is useful when multiple signals need to be combined and broadcast to many outputs. However, other constructions of an $n \times n$ coupler in integrated optics are also possible (see, for example, [Dra89]).

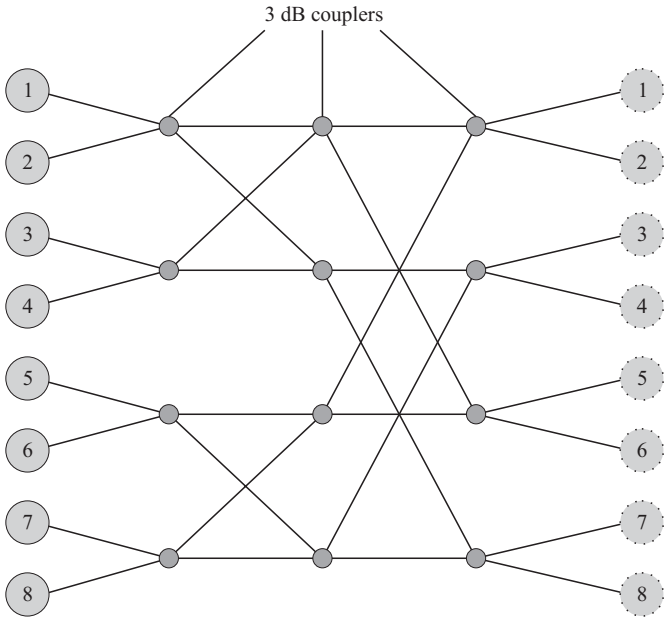

Figure 3.2 A star coupler with eight inputs and eight outputs made by combining 3 dB couplers. The power from each input is split equally among all the outputs.

Couplers are also used to tap off a small portion of the power from a light stream for monitoring purposes or other reasons. Such couplers are also called taps and are designed with values of α close to 1, typically 0.90–0.95.

Couplers are the building blocks for several other optical devices. We will explore the use of directional couplers in modulators and switches in Sections 3.5.4 and 3.7. Couplers are also the principal components used to construct *Mach-Zehnder interferometers,* which can be used as optical filters, multiplexers/demultiplexers, or as building blocks for optical modulators, switches, and wavelength converters. We will study these devices in Section 3.3.7.

So far, we have looked at wavelength-independent couplers. A coupler can be made wavelength selective, meaning that its coupling coefficient will then depend on the wavelength of the signal. Such couplers are widely used to combine signals at 1310 nm and 1550 nm into a single fiber without loss. In this case, the 1310 nm signal on input 1 is passed through to output 1, whereas the 1550 nm signal on input 2 is passed through also to output 1. The same coupler can also be used to separate the two signals coming in on a common fiber. Wavelength-dependent couplers are

also used to combine 980 nm or 1480 nm pump signals along with a 1550 nm signal into an erbium-doped fiber amplifier; see Figures 3.34 and 3.37.

In addition to the coupling ratio α, we need to look at a few other parameters while selecting couplers for network applications. The *excess loss* is the loss of the device above the fundamental loss introduced by the coupling ratio α. For example, a 3 dB coupler has a nominal loss of 3 dB but may introduce additional losses of, say, 0.2 dB. The other parameter is the variation of the coupling ratio α compared to its nominal value, due to tolerances in manufacturing, as well as wavelength dependence. In addition, we also need to maintain low *polarization-dependent loss* (PDL) for most applications.

3.1.1 Principle of Operation

When two waveguides are placed in proximity to each other, as shown in Figure 3.1, light "couples" from one waveguide to the other. This is because the propagation modes of the combined waveguide are quite different from the propagation modes of a single waveguide due to the presence of the other waveguide. When the two waveguides are identical, which is the only case we consider in this book, light launched into one waveguide couples to the other waveguide completely and then back to the first waveguide in a periodic manner. A quantitative analysis of this coupling phenomenon must be made using *coupled mode theory* [Yar97] and is beyond the scope of this book. The net result of this analysis is that the electric fields, E_{o1} and E_{o2}, at the outputs of a directional coupler may be expressed in terms of the electric fields at the inputs E_{i1} and E_{i2}, as follows:

$$\begin{pmatrix} E_{o1}(f) \\ E_{o2}(f) \end{pmatrix} = e^{-i\beta l} \begin{pmatrix} \cos(\kappa l) & i\sin(\kappa l) \\ i\sin(\kappa l) & \cos(\kappa l) \end{pmatrix} \begin{pmatrix} E_{i1}(f) \\ E_{i2}(f) \end{pmatrix}. \tag{3.1}$$

Here, l denotes the coupling length (see Figure 3.1), and β is the propagation constant in each of the two waveguides of the directional coupler. The quantity κ is called the *coupling coefficient* and is a function of the width of the waveguides, the refractive indices of the waveguiding region (core) and the substrate, and the proximity of the two waveguides. Equation (3.1) will prove useful in deriving the transfer functions of more complex devices built using directional couplers (see Problem 3.1).

Although the directional coupler is a two-input, two-output device, it is often used with only one active input, say, input 1. In this case, the power transfer function of the directional coupler is

$$\begin{pmatrix} T_{11}(f) \\ T_{12}(f) \end{pmatrix} = \begin{pmatrix} \cos^2(\kappa l) \\ \sin^2(\kappa l) \end{pmatrix}. \tag{3.2}$$

Here, $T_{ij}(f)$ represents the power transfer function from input i to output j and is defined by $T_{ij}(f) = |E_{oj}|^2/|E_{ii}|^2$. Equation (3.2) can be derived from (3.1) by setting $E_{i2} = 0$.

Note from (3.2) that for a 3 dB coupler the coupling length must be chosen to satisfy $\kappa l = (2k+1)\pi/4$, where k is a nonnegative integer.

3.1.2 Conservation of Energy

The general form of (3.1) can be derived merely by assuming that the directional coupler is lossless. Assume that the input and output electric fields are related by a general equation of the form

$$\begin{pmatrix} E_{o1} \\ E_{o2} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} E_{i1} \\ E_{i2} \end{pmatrix}. \tag{3.3}$$

The matrix

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$$

is the transfer function of the device relating the input and output electric fields and is called the *scattering matrix*. We use complex representations for the input and output electric fields, and thus the s_{ij} are also complex. It is understood that we must consider the real part of these complex fields in applications. This complex representation for the s_{ij} allows us to conveniently represent any induced phase shifts.

For convenience, we denote $\mathbf{E}_o = (E_{o1}, E_{o2})^T$ and $\mathbf{E}_i = (E_{i1}, E_{i2})^T$, where the superscript T denotes the transpose of the vector/matrix. In this notation, (3.3) can be written compactly as $\mathbf{E}_o = \mathbf{S}\mathbf{E}_i$.

The sum of the powers of the input fields is proportional to $\mathbf{E}_i^T\mathbf{E}_i^* = |E_{i1}|^2 + |E_{i2}|^2$. Here, * represents the complex conjugate. Similarly, the sum of the powers of the output fields is proportional to $\mathbf{E}_o^T\mathbf{E}_o^* = |E_{o1}|^2 + |E_{o2}|^2$. If the directional coupler is lossless, the power in the output fields must equal the power in the input fields so that

$$\begin{aligned} \mathbf{E}_o^T\mathbf{E}_o &= (\mathbf{S}\mathbf{E}_i)^T(\mathbf{S}\mathbf{E}_i)^* \\ &= \mathbf{E}_i^T(\mathbf{S}^T\mathbf{S}^*)\mathbf{E}_i^* \\ &= \mathbf{E}_i^T\mathbf{E}_i^*. \end{aligned}$$

Since this relationship must hold for arbitrary $\mathbf{E}_i$, we must have

$$\mathbf{S}^T\mathbf{S}^* = \mathbf{I}, \tag{3.4}$$

where **I** is the identity matrix. Note that this relation follows merely from conservation of energy and can be readily generalized to a device with an arbitrary number of inputs and outputs.

For a 2×2 directional coupler, by the symmetry of the device, we can set $s_{21} = s_{12} = a$ and $s_{22} = s_{11} = b$. Applying (3.4) to this simplified scattering matrix, we get

$$|a|^2 + |b|^2 = 1 \tag{3.5}$$

and

$$ab^* + ba^* = 0. \tag{3.6}$$

From (3.5), we can write

$$|a| = \cos(x) \text{ and } |b| = \sin(x). \tag{3.7}$$

If we write $a = \cos(x)e^{i\phi_a}$ and $b = \sin(x)e^{i\phi_b}$, (3.6) yields

$$\cos(\phi_a - \phi_b) = 0. \tag{3.8}$$

Thus ϕ_a and ϕ_b must differ by an odd multiple of $\pi/2$. The general form of (3.1) now follows from (3.7) and (3.8).

The conservation of energy has some important consequences for the kinds of optical components that we can build. First, note that for a 3 dB coupler, though the electric fields at the two outputs have the same magnitude, they have a relative phase shift of $\pi/2$. This relative phase shift, which follows from the conservation of energy as we just saw, plays a crucial role in the design of devices such as the Mach-Zehnder interferometer that we will study in Section 3.3.7.

Another consequence of the conservation of energy is that *lossless combining* is not possible. Thus we cannot design a device with three ports where the power input at two of the ports is completely delivered to the third port. This result is demonstrated in Problem 3.2.

3.2 Isolators and Circulators

Couplers and most other passive optical devices are *reciprocal* devices in that the devices work exactly the same way if their inputs and outputs are reversed. However, in many systems there is a need for a passive *nonreciprocal* device. An *isolator* is an example of such a device. Its main function is to allow transmission in one direction through it but block all transmission in the other direction. Isolators are used in systems at the output of optical amplifiers and lasers primarily to prevent reflections from entering these devices, which would otherwise degrade their performance. The

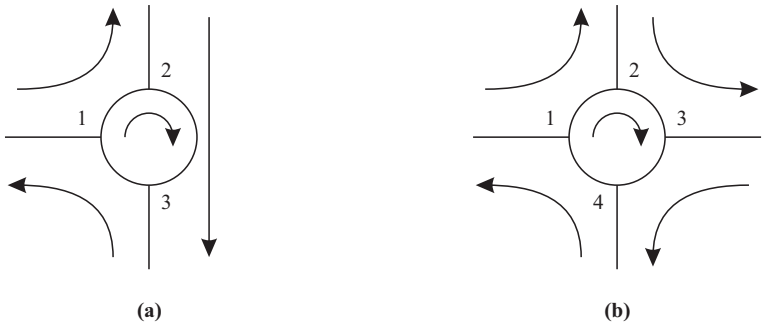

Figure 3.3 Functional representation of circulators: (a) three-port and (b) four-port. The arrows represent the direction of signal flow.

two key parameters of an isolator are its *insertion loss*, which is the loss in the forward direction and which should be as small as possible, and its *isolation*, which is the loss in the reverse direction and which should be as large as possible. The typical insertion loss is around 1 dB, and the isolation is around 40–50 dB.

A *circulator* is similar to an isolator, except that it has multiple ports, typically three or four, as shown in Figure 3.3. In a three-port circulator, an input signal on port 1 is sent out on port 2, an input signal on port 2 is sent out on port 3, and an input signal on port 3 is sent out on port 1. Circulators are useful to construct optical add/drop elements, as we will see in Section 3.3.4. Circulators operate on the same principles as isolators; therefore we only describe the details of how isolators work next.

3.2.1 Principle of Operation

In order to understand the operation of an isolator, we need to understand the notion of *polarization*. Recall from Section 2.3.3 that the *state of polarization* (SOP) of light propagating in a single-mode fiber refers to the orientation of its electric field vector on a plane that is orthogonal to its direction of propagation. At any time, the electric field vector can be expressed as a linear combination of the two orthogonal linear polarizations supported by the fiber. We will call these two polarization modes the horizontal and vertical modes.

The principle of operation of an isolator is shown in Figure 3.4. Assume that the input light signal has the vertical SOP shown in the figure. It is passed through a *polarizer*, which passes only light energy in the vertical SOP and blocks light energy in the horizontal SOP. Such polarizers can be realized using crystals, called *dichroics*,

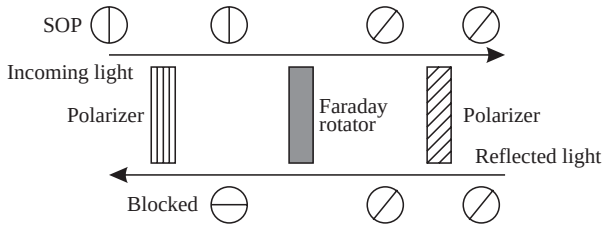

Figure 3.4 Principle of operation of an isolator that works only for a particular state of polarization of the input signal.

which have the property of selectively absorbing light with one SOP. The polarizer is followed by a *Faraday rotator*. A Faraday rotator is a nonreciprocal device, made of a crystal that rotates the SOP, say, clockwise, by 45°, regardless of the direction of propagation. The Faraday rotator is followed by another polarizer that passes only SOPs with this 45° orientation. Thus the light signal from left to right is passed through the device without any loss. On the other hand, light entering the device from the right due to a reflection, with the same 45° SOP orientation, is rotated another 45° by the Faraday rotator, and thus blocked by the first polarizer.

Note that the preceding explanation assumes a particular SOP for the input light signal. In practice we cannot control the SOP of the input, and so the isolator must work regardless of the input SOP. This requires a more complicated design, and many different designs exist. One such design for a miniature polarization-independent isolator is shown in Figure 3.5. The input signal with an arbitrary SOP is first sent through a *spatial walk-off polarizer* (SWP). The SWP splits the signal into its two orthogonally polarized components. Such an SWP can be realized using *birefringent* crystals whose refractive index is different for the two components. When light with an arbitrary SOP is incident on such a crystal, the two orthogonally polarized components are refracted at different angles. Each component goes through a Faraday rotator, which rotates the SOPs by 45°. The Faraday rotator is followed by a *half-wave plate*. The half-wave plate (a reciprocal device) rotates the SOPs by 45° in the clockwise direction for signals propagating from left to right, and by 45° in the counterclockwise direction for signals propagating from right to left. Therefore, the combination of the Faraday rotator and the half-wave plate converts the horizontal polarization into a vertical polarization and vice versa, and the two signals are combined by another SWP at the output. For reflected signals in the reverse direction, the half-wave plate and Faraday rotator cancel each other's effects, and the SOPs remain unchanged as they pass through these two devices and are thus not recombined by the SWP at the input.

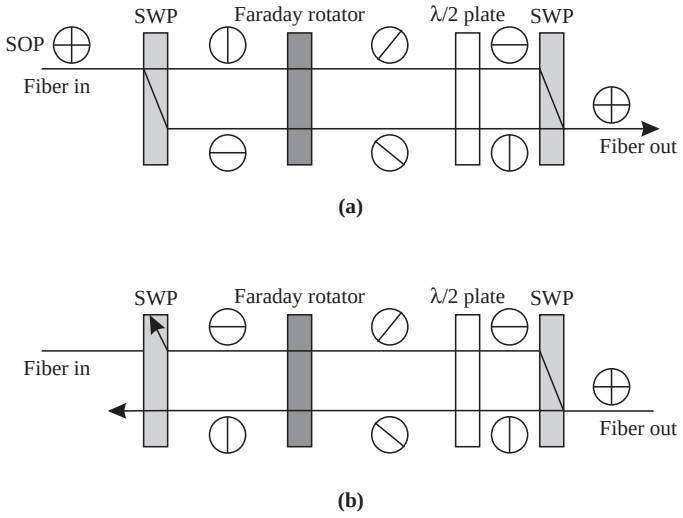

Figure 3.5 A polarization-independent isolator. The isolator is constructed along the same lines as a polarization-dependent isolator but uses spatial walk-off polarizers at the inputs and outputs. (a) Propagation from left to right. (b) Propagation from right to left.

3.3 Multiplexers and Filters

In this section, we will study the principles underlying the operation of a variety of wavelength selection technologies. Optical filters are essential components in transmission systems for at least two applications: to multiplex and demultiplex wavelengths in a WDM system—these devices are called multiplexers/demultiplexers—and to provide equalization of the gain and filtering of noise in optical amplifiers. Furthermore, understanding optical filtering is essential to understanding the operation of lasers later in this chapter.

The different applications of optical filters are shown in Figure 3.6. A simple filter is a two-port device that selects one wavelength and rejects all others. It may have an additional third port on which the rejected wavelengths can be obtained. A multiplexer combines signals at different wavelengths on its input ports onto a common output port, and a demultiplexer performs the opposite function. Multiplexers and demultiplexers are used in WDM terminals as well as in larger *wavelength crossconnects* and *wavelength add/drop multiplexers*.

Demultiplexers and multiplexers can be cascaded to realize *static* wavelength crossconnects (WXCs). In a static WXC, the crossconnect pattern is fixed at the time

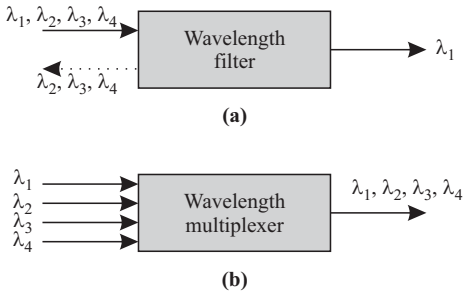

Figure 3.6 Different applications for optical filters in optical networks. (a) A simple filter, which selects one wavelength and either blocks the remaining wavelengths or makes them available on a third port. (b) A multiplexer, which combines multiple wavelengths into a single fiber. In the reverse direction, the same device acts as a demultiplexer to separate the different wavelengths.

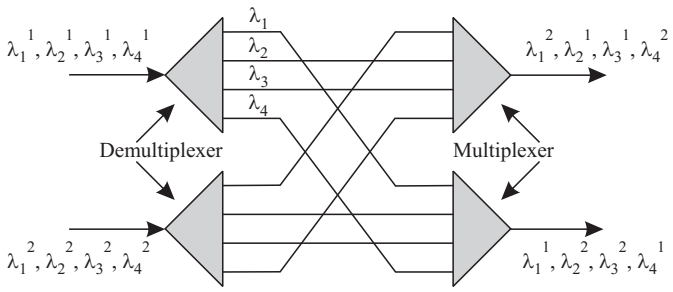

Figure 3.7 A static wavelength crossconnect. The device routes signals from an input port to an output port based on the wavelength.

the device is made and cannot be changed dynamically. Figure 3.7 shows an example of a static WXC. The device routes signals from an input port to an output port based on the wavelength. *Dynamic* WXCs can be constructed by combining optical switches with multiplexers and demultiplexers. Static WXCs are highly limited in terms of their functionality. For this reason, the devices of interest are dynamic rather than static WXCs. We will study different dynamic WXC architectures in Chapter 7.

A variety of optical filtering technologies are available. Their key characteristics for use in systems are the following:

1. Good optical filters should have low *insertion losses*. The insertion loss is the input-to-output loss of the filter.
2. The loss should be independent of the state of polarization of the input signals. The state of polarization varies randomly with time in most systems, and if the filter has a polarization-dependent loss, the output power will vary with time as well—an undesirable feature.
3. The passband of a filter should be insensitive to variations in ambient temperature. The *temperature coefficient* is measured by the amount of wavelength shift per unit degree change in temperature. The system requirement is that over the entire operating temperature range (about 100°C typically), the wavelength shift should be much less than the wavelength spacing between adjacent channels in a WDM system.
4. As more and more filters are cascaded in a WDM system, the passband becomes progressively narrower. To ensure reasonably broad passbands at the end of the cascade, the individual filters should have very flat passbands, so as to accommodate small changes in operating wavelengths of the lasers over time. This is measured by the 1 dB bandwidth, as shown in Figure 3.8.
5. At the same time, the passband skirts should be sharp to reduce the amount of energy passed through from adjacent channels. This energy is seen as *crosstalk* and degrades the system performance. The crosstalk suppression, or *isolation* of the filter, which is defined as the relative power passed through from the adjacent channels, is an important parameter as well.

In addition to all the performance parameters described, perhaps the most important consideration is cost. Technologies that require careful hand assembly tend to be more expensive. There are two ways of reducing the cost of optical filters. The first is to fabricate them using integrated-optic waveguide technology. This is analogous to semiconductor chips, although the state of integration achieved with optics is significantly less. These waveguides can be made on many substrates, including silica, silicon, InGaAs, and polymers. Waveguide devices tend to be inherently polarization dependent due to the geometry of the waveguides, and care must be taken to reduce the PDL in these devices. The second method is to realize all-fiber devices. Such devices are amenable to mass production and are inherently polarization independent. It is also easy to couple light in and out of these devices from/into other fibers. Both of these approaches are being pursued today.

All the filters and multiplexers we study use the property of *interference* among optical waves. In addition, some filters, for example, gratings, use the *diffraction* property—light from a source tends to spread in all directions depending on the

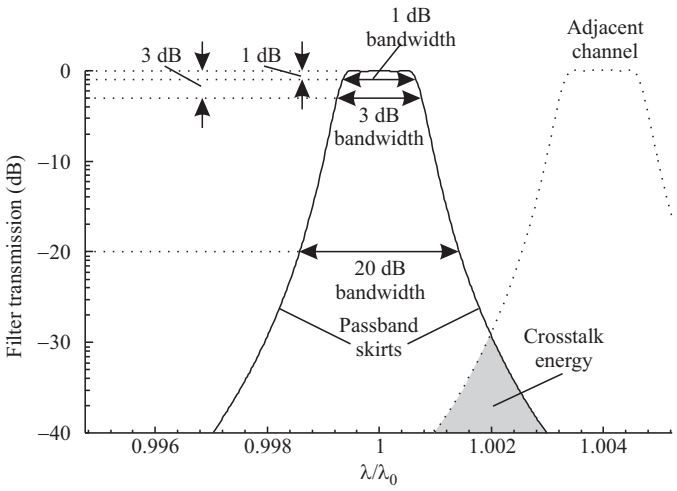

Figure 3.8 Characterization of some important spectral-shape parameters of optical filters. λ_0 is the center wavelength of the filter, and λ denotes the wavelength of the light signal.

incident wavelength. Table 3.1 compares the performance of different filtering technologies.

3.3.1 *Gratings*

The term *grating* is used to describe almost any device whose operation involves interference among multiple optical signals originating from the same source but with different relative *phase shifts*. An exception is a device where the multiple optical signals are generated by repeated traversals of a single cavity; such devices are called *etalons*. An electromagnetic wave (light) of angular frequency ω propagating, say, in the z direction has a dependence on z and t of the form $\cos(\omega t - \beta z)$. Here, β is the propagation constant and depends on the medium. The *phase* of the wave is $\omega t - \beta z$. Thus a relative phase shift between two waves from the same source can be achieved if they traverse two paths of different lengths.

Two examples of gratings are shown in Figure 3.9(a) and (b). Gratings have been widely used for centuries in optics to separate light into its constituent wavelengths. In WDM communication systems, gratings are used as demultiplexers to separate the individual wavelengths or as multiplexers to combine them. The Stimax grating of Table 3.1 is a grating of the type we describe in this section.

Table 3.1 Comparison of passive wavelength multiplexing/demultiplexing technologies. A 16-channel system with 100 GHz channel spacing is assumed. Other key considerations include center wavelength accuracy and manufacturability. All these approaches face problems in scaling with the number of wavelengths. TFMF is the dielectric thin-film multicavity filter, and AWG is the arrayed waveguide grating. For the fiber Bragg grating and the arrayed waveguide grating, the temperature coefficient can be reduced to 0.001 nm/°C by passive temperature compensation. The fiber Bragg grating is a single channel filter, and multiple filters need to be cascaded in series to demultiplex all 16 channels.

Filter Property	Fiber Bragg Grating	TFMF	AWG	Stimax Grating
1 dB BW (nm)	0.3	0.4	0.22	0.1
Isolation (dB)	25	25	25	30
Loss (dB)	0.2	7	5.5	6
PDL (dB)	0	0.2	0.5	0.1
Temp. coeff. (nm/°C)	0.01	0.0005	0.01	0.01

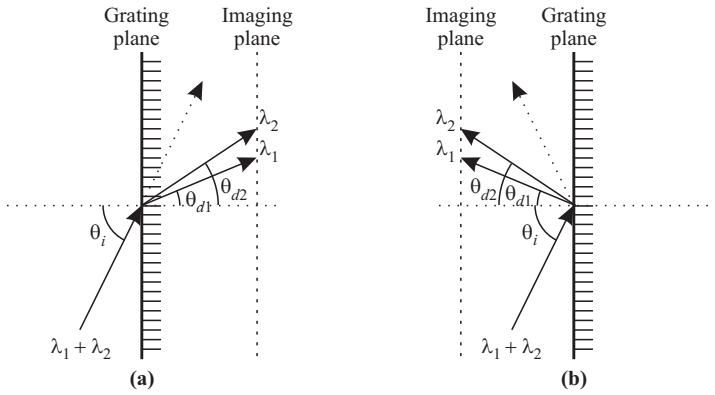

Figure 3.9 (a) A transmission grating and (b) a reflection grating. θ_i is the angle of incidence of the light signal. The angle at which the signal is diffracted depends on the wavelength ($\theta_d 1$ for wavelength λ_1 and $\theta_d 2$ for λ_2).

Consider the grating shown in Figure 3.9(a). Multiple narrow slits are spaced equally apart on a plane, called the *grating plane*. The spacing between two adjacent slits is called the *pitch* of the grating. Light incident from a source on one side of the grating is transmitted through these slits. Since each slit is narrow, by the phenomenon known as *diffraction,* the light transmitted through each slit spreads out in all directions. Thus each slit acts as a secondary source of light. Consider some other plane parallel to the grating plane at which the transmitted light from all the slits interferes. We will call this plane the *imaging plane*. Consider any point on this imaging plane. For wavelengths for which the individual interfering waves at this point are in phase, we have constructive interference and an enhancement of the light intensity at these wavelengths. For a large number of slits, which is the case usually encountered in practice, the interference is not constructive at other wavelengths, and there is little light intensity at this point from these wavelengths. Since different wavelengths interfere constructively at different points on the imaging plane, the grating effectively separates a WDM signal spatially into its constituent wavelengths. In a fiber optic system, optical fibers could be placed at different imaging points to collect light at the different wavelengths.

Note that if there were no diffraction, we would simply have light transmitted or reflected along the directed dotted lines in Figure 3.9(a) and (b). Thus the phenomenon of diffraction is key to the operation of these devices, and for this reason they are called *diffraction gratings*. Since multiple transmissions occur in the grating of Figure 3.9(a), this grating is called a *transmission grating*. If the transmission slits are replaced by narrow reflecting surfaces, with the rest of the grating surface being nonreflecting, we get the *reflection grating* of Figure 3.9(b). The principle of operation of this device is exactly analogous to that of the transmission grating. A majority of the gratings used in practice are reflection gratings since they are somewhat easier to fabricate. In addition to the plane geometry we have considered, gratings are fabricated in a concave geometry. In this case, the slits (for a transmission grating) are located on the arc of a circle. In many applications, a concave geometry leads to fewer auxiliary parts like lenses and mirrors needed to construct the overall device, say, a WDM demultiplexer, and is thus preferred.

The Stimax grating [LL84] is a reflection grating that is integrated with a concave mirror and the input and output fibers. Its characteristics are described in Table 3.1, and it has been used in commercially available WDM transmission systems. However, it is a bulk device that cannot be easily fabricated and is therefore relatively expensive. Attempts have been made to realize similar gratings in optical waveguide technology, but these devices are yet to achieve loss, PDL, and isolation comparable to the bulk version.

Principle of Operation

To understand quantitatively the principle of operation of a (transmission) grating, consider the light transmitted through adjacent slits as shown in Figure 3.10. The distance between adjacent slits—the *pitch* of the grating—is denoted by a. We assume that the light source is far enough away from the grating plane compared to a so that the light can be assumed to be incident at the same angle θ_i to the plane of the grating at each slit. We consider the light rays diffracted at an angle θ_d from the grating plane. The imaging plane, like the source, is assumed to be far away from the grating plane compared to the grating pitch. We also assume that the slits are small compared to the wavelength so that the phase change across a slit is negligible. Under these assumptions, it can be shown (Problem 3.4) that the path length difference between the rays traversing through adjacent slits is the difference in lengths between the line segments $\overline{AB}$ and $\overline{CD}$ and is given approximately by $a[\sin(\theta_i) - \sin(\theta_d)]$. Thus constructive interference at a wavelength λ occurs at the imaging plane among the rays diffracted at angle θ_d if the following *grating equation* is satisfied:

$$a[\sin(\theta_i) - \sin(\theta_d)] = m\lambda \tag{3.9}$$

for some integer m, called the *order* of the grating. The grating effects the separation of the individual wavelengths in a WDM signal since the grating equation is satisfied at different points in the imaging plane for different wavelengths. This is illustrated in Figure 3.9, where different wavelengths are shown being diffracted at the angles at which the grating equation is satisfied for that wavelength. For example, θ_{d1} is the angle at which the grating equation is satisfied for λ_1.

Note that the energy at a single wavelength is distributed over all the discrete angles that satisfy the grating equation (3.9) at this wavelength. When the grating is used as a demultiplexer in a WDM system, light is collected from only one of these angles, and the remaining energy in the other orders is lost. In fact, most of the energy will be concentrated in the zeroth-order ($m = 0$) interference maximum, which occurs at $\theta_i = \theta_d$ for all wavelengths. The light energy in this zeroth-order interference maximum is wasted since the wavelengths are not separated. Thus gratings must be designed so that the light energy is maximum at one of the other interference maxima. This is done using a technique called *blazing* [KF86, p. 386].

Figure 3.11 shows a blazed reflection grating with blaze angle α. In such a grating, the reflecting slits are inclined at an angle α to the grating plane. This has the effect of maximizing the light energy in the interference maximum whose order corresponds to the blazing angle. The grating equation for such a blazed grating can be derived as before; see Problem 3.5.

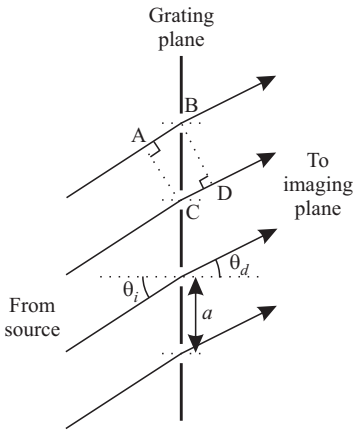

Figure 3.10 Principle of operation of a transmission grating. The reflection grating works in an analogous manner. The path length difference between rays diffracted at angle θ_d from adjacent slits is $\overline{AB} - \overline{CD} = a[\sin(\theta_i) - \sin(\theta_d)]$.

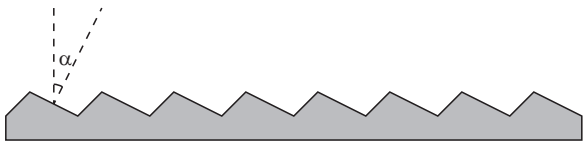

Figure 3.11 A blazed grating with blaze angle α. The energy in the interference maximum corresponding to the blaze angle is maximized.

3.3.2 *Diffraction Pattern*

So far, we have only considered the position of the diffraction maxima in the diffraction pattern. Often, we are also interested in the distribution of the intensity in the diffraction maxima. We can derive the distribution of the intensity by relaxing the assumption that the slits are much smaller than a wavelength, so that the phase change across a slit can no longer be neglected. Consider a slit of length w stretching from $y = -w/2$ to $y = w/2$. By reasoning along the same lines as we did in Figure 3.10, the light diffracted from position y at angle θ from this slit has a relative phase shift of $\phi(y) = (2\pi y \sin\theta)/\lambda$ compared to the light diffracted from $y = 0$. Thus, at the

imaging plane, the amplitude $A(\theta)$ at angle θ is given by

$$\begin{aligned}\frac{A(\theta)}{A(0)} &= \frac{1}{w}\int_{-w/2}^{w/2} \exp\left(i\phi(y)\right)\, dy \\ &= \frac{1}{w}\int_{-w/2}^{w/2} \exp\left(i2\pi(\sin\theta)y/\lambda\right)\, dy \\ &= \frac{\sin\left(\pi w \sin\theta/\lambda\right)}{\pi w \sin\theta/\lambda}. \end{aligned} \tag{3.10}$$

Observe that the amplitude distribution at the imaging plane is the Fourier transform of the rectangular slit. This result holds for a general diffracting aperture, and not just a rectangular slit. For this more general case, if the diffracting aperture or slit is described by $f(y)$, the amplitude distribution of the diffraction pattern is given by

$$A(\theta) = A(0)\int_{-\infty}^{\infty} f(y)\exp(2\pi i(\sin\theta)y/\lambda)\, dy. \tag{3.11}$$

The intensity distribution is given by $|A(\theta)|^2$. Here, we assume $f(y)$ is normalized so that $\int_{-\infty}^{\infty} f(y)\, dy = 1$. For a rectangular slit, $f(y) = 1/w$ for $|y| < w/2$ and $f(y) = 0$, otherwise, and the diffraction pattern is given by (3.10). For a pair of narrow slits spaced distance d apart,

$$f(y) = 0.5(\delta(y - d/2) + \delta(y + d/2))$$

and

$$A(\theta) = A(0)\cos\left(\pi(\sin\theta)\lambda/d\right).$$

The more general problem of N narrow slits is discussed in Problem 3.6.

3.3.3 Bragg Gratings

Bragg gratings are widely used in fiber optic communication systems. In general, any periodic perturbation in the propagating medium serves as a Bragg grating. This perturbation is usually a periodic variation of the refractive index of the medium. We will see in Section 3.5.1 that lasers use Bragg gratings to achieve single frequency operation. In this case, the Bragg gratings are "written" in waveguides. Bragg gratings written in fiber can be used to make a variety of devices such as filters, add/drop multiplexers, and dispersion compensators. We will see later that the Bragg grating principle also underlies the operation of the acousto-optic tunable filter. In this case, the Bragg grating is formed by the propagation of an acoustic wave in the medium.

Principle of Operation

Consider two waves propagating in opposite directions with propagation constants β_0 and β_1. Energy is coupled from one wave to the other if they satisfy the Bragg *phase-matching* condition

$$|\beta_0 - \beta_1| = \frac{2\pi}{\Lambda},$$

where Λ is the period of the grating. In a Bragg grating, energy from the forward propagating mode of a wave at the right wavelength is coupled into a backward propagating mode. Consider a light wave with propagation constant β_1 propagating from left to right. The energy from this wave is coupled onto a scattered wave traveling in the opposite direction at the same wavelength provided

$$|\beta_0 - (-\beta_0)| = 2\beta_0 = \frac{2\pi}{\Lambda}.$$

Letting $\beta_0 = 2\pi n_{\text{eff}}/\lambda_0$, λ_0 being the wavelength of the incident wave and n_{eff} the effective refractive index of the waveguide or fiber, the wave is reflected provided

$$\lambda_0 = 2n_{\text{eff}}\Lambda.$$

This wavelength λ_0 is called the Bragg wavelength. In practice, the reflection efficiency decreases as the wavelength of the incident wave is detuned from the Bragg wavelength; this is plotted in Figure 3.12(a). Thus if several wavelengths are transmitted into a fiber Bragg grating, the Bragg wavelength is reflected while the other wavelengths are transmitted.

The operation of the Bragg grating can be understood by reference to Figure 3.13, which shows a periodic variation in refractive index. The incident wave is reflected from each period of the grating. These reflections add in phase when the path length in wavelength λ_0 each period is equal to half the incident wavelength λ_0. This is equivalent to $n_{\text{eff}}\Lambda = \lambda_0/2$, which is the Bragg condition.

The reflection spectrum shown in Figure 3.12(a) is for a grating with a uniform refractive index pattern change across its length. In order to eliminate the undesirable side lobes, it is possible to obtain an *apodized* grating, where the refractive index change is made smaller toward the edges of the grating. (The term *apodized* means "to cut off the feet.") The reflection spectrum of an apodized grating is shown in Figure 3.12(b). Note that, for the apodized grating, the side lobes have been drastically reduced but at the expense of increasing the main lobe width.

The index distribution across the length of a Bragg grating is analogous to the grating aperture discussed in Section 3.3.2, and the reflection spectrum is obtained as the Fourier transform of the index distribution. The side lobes in the case of a uniform refractive index profile arise due to the abrupt start and end of the grating,

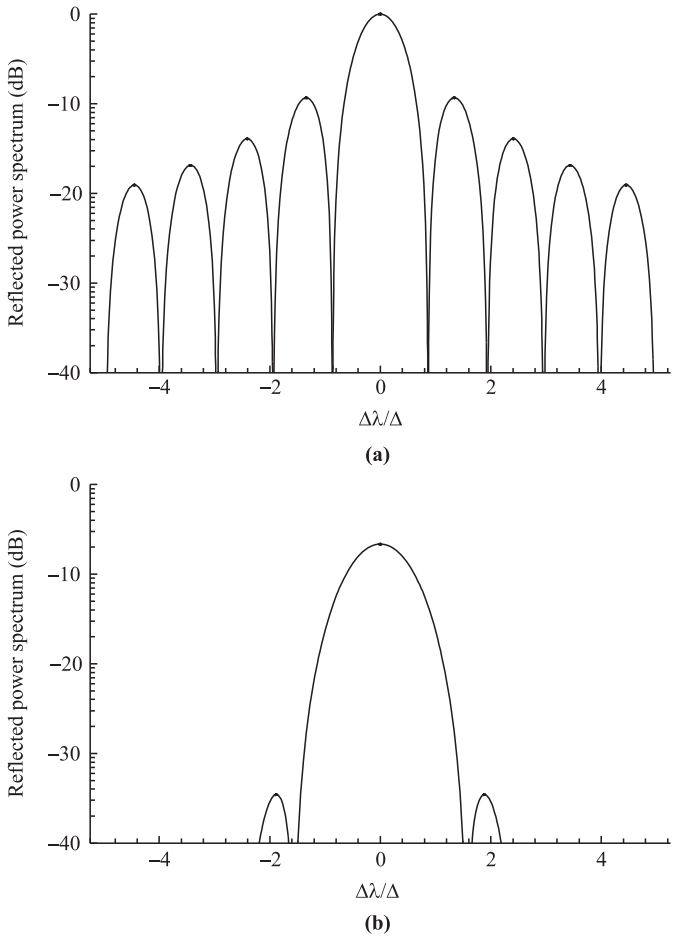

Figure 3.12 Reflection spectra of Bragg gratings with (a) uniform index profile and (b) apodized index profile. Δ is a measure of the bandwidth of the grating and is the wavelength separation between the peak wavelength and the first reflection minimum, in the uniform index profile case. Δ is inversely proportional to the length of the grating. $\Delta\lambda$ is the detuning from the phase-matching wavelength.

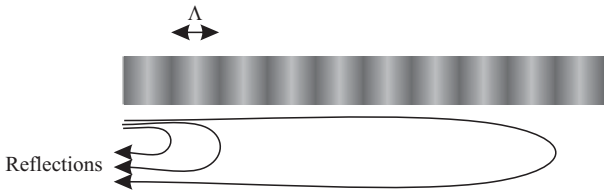

Figure 3.13 Principle of operation of a Bragg grating.

which result in a sinc(.) behavior for the side lobes. Apodization can be achieved by gradually starting and ending the grating. This technique is similar to pulse shaping used in digital communication systems to reduce the side lobes in the transmitted spectrum of the signal.

The bandwidth of the grating, which can be measured, for example, by the width of the main lobe, is inversely proportional to the length of the grating. Typically, the grating is a few millimeters long in order to achieve a bandwidth of 1 nm.

3.3.4 Fiber Gratings

Fiber gratings are attractive devices that can be used for a variety of applications, including filtering, add/drop functions, and compensating for accumulated dispersion in the system. Being all-fiber devices, their main advantages are their low loss, ease of coupling (with other fibers), polarization insensitivity, low temperature-coefficient, and simple packaging. As a result, they can be extremely low-cost devices.

Gratings are written in fibers by making use of the *photosensitivity* of certain types of optical fibers. A conventional silica fiber doped with germanium becomes extremely photosensitive. Exposing this fiber to ultraviolet (UV) light causes changes in the refractive index within the fiber core. A grating can be written in such a fiber by exposing its core to two interfering UV beams. This causes the radiation intensity to vary periodically along the length of the fiber. Where the intensity is high, the refractive index is increased; where it is low, the refractive index is unchanged. The change in refractive index needed to obtain gratings is quite small—around 10^{-4}. Other techniques, such as *phase masks,* can also be used to produce gratings. A phase mask is a diffractive optical element. When it is illuminated by a light beam, it splits the beams into different diffractive orders, which then interfere with one another to write the grating into the fiber.

Fiber gratings are classified as either *short-period* or *long-period* gratings, based on the period of the grating. Short-period gratings are also called Bragg gratings and have periods that are comparable to the wavelength, typically around 0.5 μm. We

discussed the behavior of Bragg gratings in Section 3.3.3. Long-period gratings, on the other hand, have periods that are much greater than the wavelength, ranging from a few hundred micrometers to a few millimeters.

Fiber Bragg Gratings

Fiber Bragg gratings can be fabricated with extremely low loss (0.1 dB), high wavelength accuracy ($\pm$ 0.05 nm is easily achieved), high adjacent channel crosstalk suppression (40 dB), as well as flat tops.

The temperature coefficient of a fiber Bragg grating is typically 1.25×10^{-2} nm/°C due to the variation in fiber length with temperature. However, it is possible to compensate for this change by packaging the grating with a material that has a negative thermal expansion coefficient. These passively temperature-compensated gratings have temperature coefficients of around 0.07×10^{-2} nm/°C. This implies a very small 0.07 nm center wavelength shift over an operating temperature range of 100°C, which means that they can be operated without any active temperature control.

These properties of fiber Bragg gratings make them very useful devices for system applications. Fiber Bragg gratings are finding a variety of uses in WDM systems, ranging from filters and optical add/drop elements to dispersion compensators. A simple optical drop element based on fiber Bragg gratings is shown in Figure 3.14(a). It consists of a three-port circulator with a fiber Bragg grating. The circulator transmits light coming in on port 1 out on port 2 and transmits light coming in on port 2 out on port 3. In this case, the grating reflects the desired wavelength λ_2, which is then dropped at port 3. The remaining three wavelengths are passed through. It is possible to implement an add/drop function along the same lines, by introducing a coupler to add the same wavelength that was dropped, as shown in Figure 3.14(b). Many variations of this simple add/drop element can be realized by using gratings in combination with couplers and circulators. A major concern in these designs is that the reflection of these gratings is not perfect, and as a result, some power at the selected wavelength leaks through the grating. This can cause undesirable crosstalk, and we will study this effect in Chapter 5.

Fiber Bragg gratings can also be used to compensate for dispersion accumulated along the link. We will study this application in Chapter 5 in the context of dispersion compensation.

Long-Period Fiber Gratings

Long-period fiber gratings are fabricated in the same manner as fiber Bragg gratings and are used today primarily as filters inside erbium-doped fiber amplifiers to compensate for their nonflat gain spectrum. As we will see, these devices serve as very

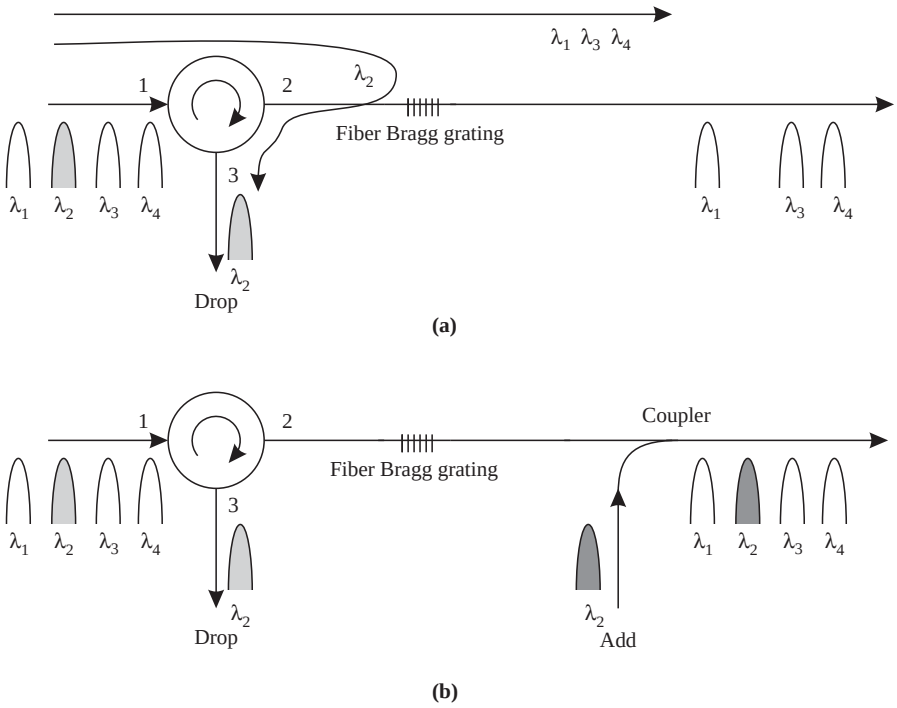

Figure 3.14 Optical add/drop elements based on fiber Bragg gratings. (a) A drop element. (b) A combined add/drop element.

efficient band rejection filters and can be tailored to provide almost exact equalization of the erbium gain spectrum. Figure 3.15 shows the transmission spectrum of such a grating. These gratings retain all the attractive properties of fiber gratings and are expected to become widely used for several filtering applications.

Principle of Operation

These gratings operate on somewhat different principles than Bragg gratings. In fiber Bragg gratings, energy from the forward propagating mode in the fiber core at the right wavelength is coupled into a backward propagating mode. In long-period gratings, energy is coupled from the forward propagating mode in the fiber core onto other forward propagating modes in the cladding. These cladding modes are extremely lossy, and their energy decays rapidly as they propagate along the fiber, due to losses at the cladding–air interface and due to microbends in the fiber. There are many cladding modes, and coupling occurs between a core mode at a given

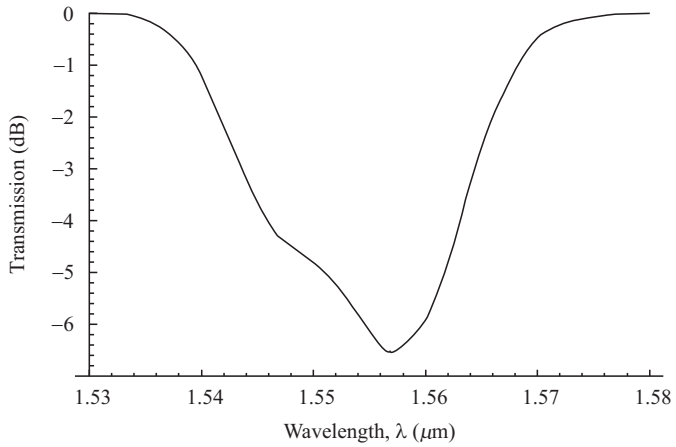

Figure 3.15 Transmission spectrum of a long-period fiber Bragg grating used as a gain equalizer for erbium-doped fiber amplifiers. (After [Ven96a].)

wavelength and a cladding mode depending on the pitch of the grating Λ, as follows: if β denotes the propagation constant of the mode in the core (assuming a single-mode fiber) and β_{cl}^{p} that of the pth-order cladding mode, then the phase-matching condition dictates that

$$\beta - \beta_{\text{cl}}^{p} = \frac{2\pi}{\Lambda}.$$

In general, the difference in propagation constants between the core mode and any one of the cladding modes is quite small, leading to a fairly large value of Λ in order for coupling to occur. This value is usually a few hundred micrometers. (Note that in Bragg gratings the difference in propagation constants between the forward and backward propagating modes is quite large, leading to a small value for Λ, typically around 0.5 μm.) If n_{eff} and n_{eff}^{p} denote the effective refractive indices of the core and pth-order cladding modes, then the wavelength at which energy is coupled from the core mode to the cladding mode can be obtained as

$$\lambda = \Lambda(n_{\text{eff}} - n_{\text{eff}}^{p}),$$

where we have used the relation $\beta = 2\pi n_{\text{eff}}/\lambda$.

Therefore, once we know the effective indices of the core and cladding modes, we can design the grating with a suitable value of Λ so as to cause coupling of energy out of a desired wavelength band. This causes the grating to act as a wavelength-dependent loss element. Methods for calculating the propagation

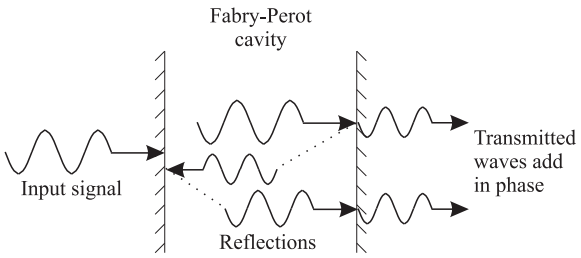

Figure 3.16 Principle of operation of a Fabry-Perot filter.

constants for the cladding modes are discussed in [Ven96b]. The amount of wavelength-dependent loss can be controlled during fabrication by controlling the UV exposure time. Complicated transmission spectra can be obtained by cascading multiple gratings with different center wavelengths and different exposures. The example shown in Figure 3.15 was obtained by cascading two such gratings [Ven96a]. These gratings are typically a few centimeters long.

3.3.5 Fabry-Perot Filters

A Fabry-Perot filter consists of the cavity formed by two highly reflective mirrors placed parallel to each other, as shown in Figure 3.16. This filter is also called a Fabry-Perot interferometer or etalon. The input light beam to the filter enters the first mirror at right angles to its surface. The output of the filter is the light beam leaving the second mirror.

This is a classical device that has been used widely in interferometric applications. Fabry-Perot filters have been used for WDM applications in several optical network testbeds. There are better filters today, such as the thin-film resonant multicavity filter that we will study in Section 3.3.6. These latter filters can be viewed as Fabry-Perot filters with wavelength-dependent mirror reflectivities. Thus the fundamental principle of operation of these filters is the same as that of the Fabry-Perot filter. The Fabry-Perot cavity is also used in lasers (see Section 3.5.1).

Compact Fabry-Perot filters are commercially available components. Their main advantage over some of the other devices is that they can be tuned to select different channels in a WDM system, as discussed later.

Principle of Operation

The principle of operation of the device is illustrated in Figure 3.16. The input signal is incident on the left surface of the cavity. After one pass through the cavity, as

shown in Figure 3.16, a part of the light leaves the cavity through the right facet and a part is reflected. A part of the reflected wave is again reflected by the left facet to the right facet. For those wavelengths for which the cavity length is an integral multiple of half the wavelength in the cavity—so that a round trip through the cavity is an integral multiple of the wavelength—all the light waves transmitted through the right facet *add in phase*. Such wavelengths are called the *resonant wavelengths* of the cavity. The determination of the resonant wavelengths of the cavity is discussed in Problem 3.7.

The power transfer function of a filter is the fraction of input light power that is transmitted by the filter as a function of optical frequency f, or wavelength. For the Fabry-Perot filter, this function is given by

$$T_{FP}(f) = \frac{\left(1 - \frac{A}{1-R}\right)^2}{\left(1 + \left(\frac{2\sqrt{R}}{1-R}\sin(2\pi f\tau)\right)^2\right)}. \tag{3.12}$$

This can also be expressed in terms of the optical free-space wavelength λ as

$$T_{FP}(\lambda) = \frac{\left(1 - \frac{A}{1-R}\right)^2}{\left(1 + \left(\frac{2\sqrt{R}}{1-R}\sin(2\pi nl/\lambda)\right)^2\right)}.$$

(By a slight abuse of notation, we use the same symbol for the power transfer function in both cases.) Here A denotes the absorption loss of each mirror, which is the fraction of incident light that is absorbed by the mirror. The quantity R denotes the *reflectivity* of each mirror (assumed to be identical), which is the fraction of incident light that is reflected by the mirror. The one-way propagation delay across the cavity is denoted by τ. The refractive index of the cavity is denoted by n and its length by l. Thus $\tau = nl/c$, where c is the velocity of light in vacuum. This transfer function can be derived by considering the sum of the waves transmitted by the filter after an odd number of passes through the cavity. This is left as an exercise (Problem 3.8).

The power transfer function of the Fabry-Perot filter is plotted in Figure 3.17 for $A = 0$ and $R = 0.75$, 0.9, and 0.99. Note that very high mirror reflectivities are required to obtain good isolation of adjacent channels.

The power transfer function $T_{FP}(f)$ is periodic in f, and the peaks, or *passbands,* of the transfer function occur at frequencies f that satisfy $f\tau = k/2$ for some positive integer k. Thus in a WDM system, even if the wavelengths are spaced sufficiently far apart compared to the width of each passband of the filter transfer function, several frequencies (or wavelengths) may be transmitted by the filter if

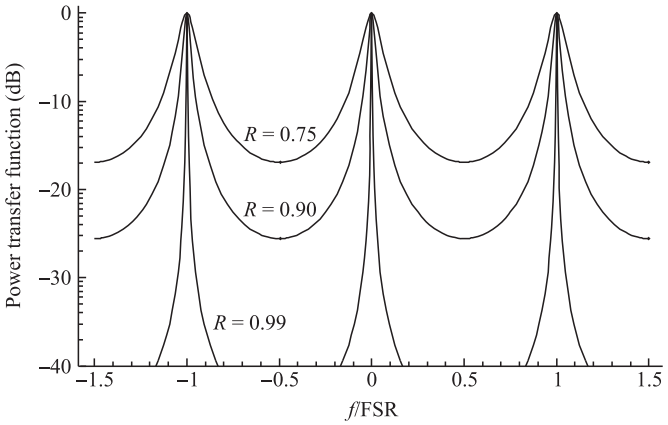

Figure 3.17 The transfer function of a Fabry-Perot filter. FSR denotes the free spectral range, f the frequency, and R the reflectivity.

they coincide with different passbands. The spectral range between two successive passbands of the filter is called the *free spectral range* (FSR). A measure of the width of each passband is its *full width* at the point where the transfer function is *half* of its *maximum* (FWHM). In WDM systems, the separation between two adjacent wavelengths must be at least a FWHM in order to minimize crosstalk. (More precisely, as the transfer function is periodic, adjacent wavelengths must be separated by a FWHM plus an integral multiple of the FSR.) Thus the ratio FSR/FWHM is an approximate (order-of-magnitude) measure of the number of wavelengths that can be accommodated by the system. This ratio is called the *finesse*, F, of the filter and is given by

$$F = \frac{\pi\sqrt{R}}{1-R}. \tag{3.13}$$

This expression can be derived from (3.12) and is left as an exercise (Problem 3.9).

If the mirrors are highly reflective, won't virtually all the input light get reflected? Also, how does light get out of the cavity if the mirrors are highly reflective? To resolve this paradox, we must look at the light energy over all the frequencies. When we do this, we will see that only a small fraction of the input light is transmitted through the cavity because of the high reflectivities of the input and output facets, but at the right frequency, all the power is transmitted. This aspect is explored further in Problem 3.10.

Tunability

A Fabry-Perot filter can be tuned to select different wavelengths in one of several ways. The simplest approach is to change the cavity length. The same effect can be achieved by varying the refractive index within the cavity. Consider a WDM system, all of whose wavelengths lie within one FSR of the Fabry-Perot filter. The frequency f_0 that is selected by the filter satisfies $f_0\tau = k/2$ for some positive integer k. Thus f_0 can be changed by changing τ, which is the one-way propagation time for the light beam across the cavity. If we denote the length of the cavity by l and its refractive index by n, $\tau = ln/c$, where c is the speed of light in vacuum. Thus τ can be changed by changing either l or n.

Mechanical tuning of the filter can be effected by moving one of the mirrors so that the cavity length changes. This permits tunability only in times of the order of a few milliseconds. For a mechanically tuned Fabry-Perot filter, a precise mechanism is needed in order to keep the mirrors parallel to each other in spite of their relative movement. The reliability of mechanical tuning mechanisms is also relatively poor.

Another approach to tuning is to use a piezoelectric material within the cavity. A piezoelectric filter undergoes compression on the application of a voltage. Thus the length of the cavity filled with such a material can be changed by the application of a voltage, thereby effecting a change in the resonant frequency of the cavity. The piezo material, however, introduces undesirable effects such as thermal instability and hysteresis, making such a filter difficult to use in practical systems.

3.3.6 Multilayer Dielectric Thin-Film Filters

A thin-film resonant cavity filter (TFF) is a Fabry-Perot interferometer, or etalon (see Section 3.3.5), where the mirrors surrounding the cavity are realized by using multiple reflective dielectric thin-film layers (see Problem 3.13). This device acts as a bandpass filter, passing through a particular wavelength and reflecting all the other wavelengths. The wavelength that is passed through is determined by the cavity length.

A thin-film resonant multicavity filter (TFMF) consists of two or more cavities separated by reflective dielectric thin-film layers, as shown in Figure 3.18. The effect of having multiple cavities on the response of the filter is illustrated in Figure 3.19. As more cavities are added, the top of the passband becomes flatter and the skirts become sharper, both very desirable filter features.

In order to obtain a multiplexer or a demultiplexer, a number of these filters can be cascaded, as shown in Figure 3.20. Each filter passes a different wavelength and reflects all the others. When used as a demultiplexer, the first filter in the cascade

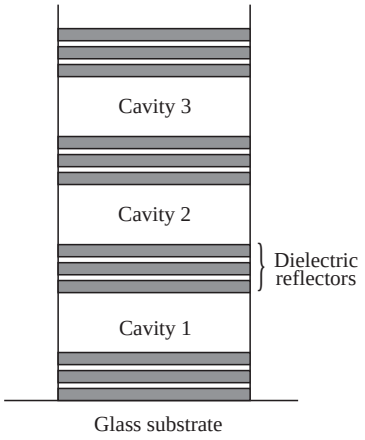

Figure 3.18 A three-cavity thin-film resonant dielectric thin-film filter. (After [SS96].)

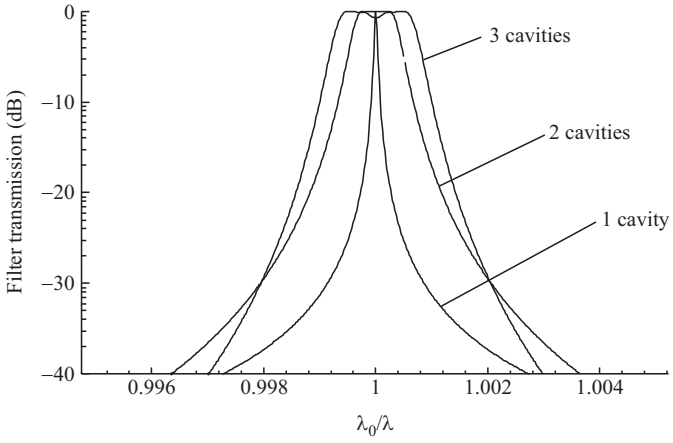

Figure 3.19 Transfer functions of single-cavity, two-cavity, and three-cavity dielectric thin-film filters. Note how the use of multiple cavities leads to a flatter passband and a sharper transition from the passband to the stop band.

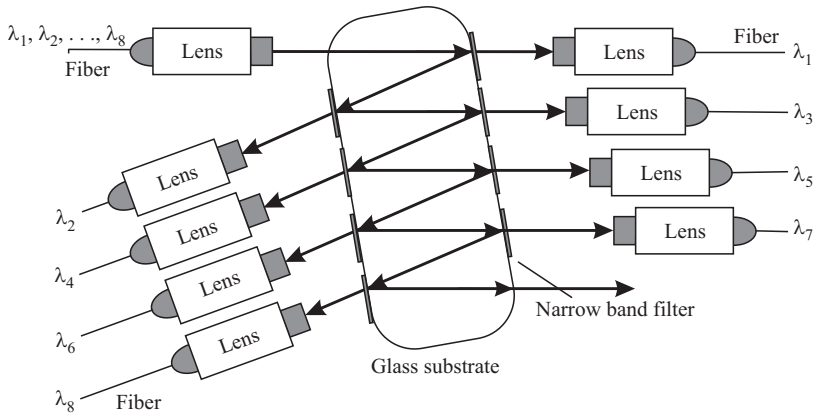

Figure 3.20 A wavelength multiplexer/demultiplexer using multilayer dielectric thin-film filters. (After [SS96].)

passes one wavelength and reflects all the others onto the second filter. The second filter passes another wavelength and reflects the remaining ones, and so on.

This device has many features that make it attractive for system applications. It is possible to have a very flat top on the passband and very sharp skirts. The device is extremely stable with regard to temperature variations, has low loss, and is insensitive to the polarization of the signal. Typical parameters for a 16-channel multiplexer are shown in Table 3.1. For these reasons, TFMFs are becoming widely used in commercial systems today. Understanding the principle of operation of these devices requires some knowledge of electromagnetic theory, and so we defer this to Appendix G.

3.3.7 *Mach-Zehnder Interferometers*

A Mach-Zehnder interferometer (MZI) is an interferometric device that makes use of two interfering paths of different lengths to resolve different wavelengths. Devices constructed on this principle have been around for some decades. Today, Mach-Zehnder interferometers are typically constructed in integrated optics and consist of two 3 dB directional couplers interconnected through two paths of differing lengths, as shown in Figure 3.21(a). The substrate is usually silicon, and the waveguide and cladding regions are silica (SiO_2).

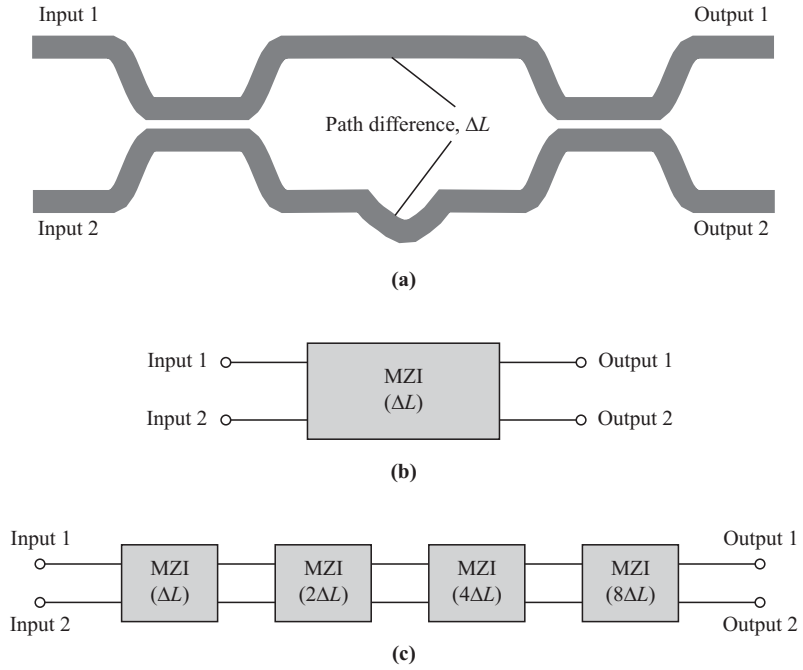

Figure 3.21 (a) An MZI constructed by interconnecting two 3 dB directional couplers. (b) A block diagram representation of the MZI in (a). ΔL denotes the path difference between the two arms. (c) A block diagram of a four-stage Mach-Zehnder interferometer, which uses different path length differences in each stage.

Mach-Zehnder interferometers are useful as both filters and (de)multiplexers. Even though there are better technologies for making narrow band filters, for example, dielectric multicavity thin-film filters, MZIs are still useful in realizing wide band filters. For example, MZIs can be used to separate the wavelengths in the 1.3 μm and 1.55 μm bands. Narrow band MZI filters are fabricated by cascading a number of stages, as we will see, and this leads to larger losses. In principle, very good crosstalk performance can be achieved using MZIs if the wavelengths are spaced such that the undesired wavelengths occur at, or close to, the nulls of the power transfer function. However, in practice, the wavelengths cannot be fixed precisely (for example, the wavelengths drift because of temperature variations or age). Moreover, the coupling ratio of the directional couplers is not 50:50 and could be wavelength dependent. As

a result, the crosstalk performance is far from the ideal situation. Also the passband of narrow band MZIs is not flat. In contrast, the dielectric multicavity thin-film filters can have flat passbands and good stop bands.

MZIs are useful as two-input, two-output multiplexers and demultiplexers. They can also be used as tunable filters, where the tuning is achieved by varying the temperature of one of the arms of the device. This causes the refractive index of that arm to change, which in turn affects the phase relationship between the two arms and causes a different wavelength to be coupled out. The tuning time required is of the order of several milliseconds. For higher channel-count multiplexers and demultiplexers, better technologies are available today. One example is the *arrayed waveguide grating* (AWG) described in the next section. Since understanding the MZI is essential to understanding the AWG, we will now describe the principle of operation of MZIs.

Principle of Operation

Consider the operation of the MZI as a demultiplexer; so only one input, say, input 1, has a signal (see Figure 3.21(a)). After the first directional coupler, the input signal power is divided equally between the two arms of the MZI, but the signal in one arm has a phase shift of $\pi/2$ with respect to the other. Specifically, the signal in the lower arm lags the one in the upper arm in phase by $\pi/2$, as discussed in Section 3.1. This is best understood from (3.1). Since there is a length difference of ΔL between the two arms, there is a further phase lag of $\beta\Delta L$ introduced in the signal in the lower arm. In the second directional coupler, the signal from the lower arm undergoes another phase delay of $\pi/2$ in going to the first output relative to the signal from the upper arm. Thus the total relative phase difference at the first or upper output between the two signals is $\pi/2 + \beta\Delta L + \pi/2$. At the output directional coupler, in going to the second output, the signal from the upper arm lags the signal from the lower arm in phase by $\pi/2$. Thus the total relative phase difference at the second or lower output between the two signals is $\pi/2 + \beta\Delta L - \pi/2 = \beta\Delta L$.

If $\beta\Delta L = k\pi$ and k is odd, the signals at the first output add in phase, whereas the signals at the second output add with opposite phases and thus cancel each other. Thus the wavelengths passed from the first input to the first output are those wavelengths for which $\beta\Delta L = k\pi$ and k is odd. The wavelengths passed from the first input to the second output are those wavelengths for which $\beta\Delta L = k\pi$ and k is even. This could have been easily deduced from the transfer function of the MZI in the following equation (3.14), but this detailed explanation will help in the understanding of the arrayed waveguide grating (Section 3.3.8).

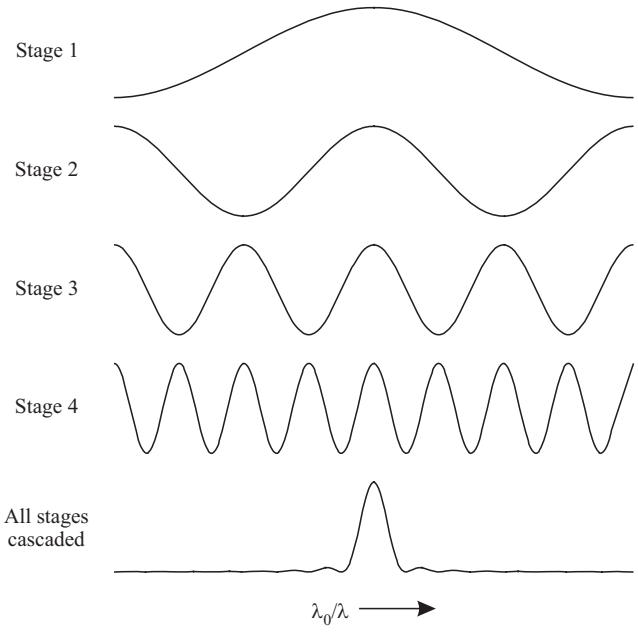

Figure 3.22 Transfer functions of each stage of a multistage MZI.

Assume that the difference between these path lengths is ΔL and that only one input, say, input 1, is active. Then it can be shown (see Problem 3.14) that the power transfer function of the Mach-Zehnder interferometer is given by

$$\begin{pmatrix} T_{11}(f) \\ T_{12}(f) \end{pmatrix} = \begin{pmatrix} \sin^2(\beta \Delta L/2) \\ \cos^2(\beta \Delta L/2) \end{pmatrix}. \tag{3.14}$$

Thus the path difference between the two arms, ΔL, is the key parameter characterizing the transfer function of the MZI. We will represent the MZI of Figure 3.21(a) using the block diagram of Figure 3.21(b).

Now consider k MZIs interconnected, as shown in Figure 3.21(c) for $k = 4$. Such a device is termed a *multistage Mach-Zehnder interferometer.* The path length difference for the kth MZI in the cascade is assumed to be $2^{k-1}\Delta L$. The transfer function of each MZI in this multistage MZI together with the power transfer function of the entire filter is shown in Figure 3.22. The power transfer function of the multistage MZI is also shown on a decibel scale in Figure 3.23.

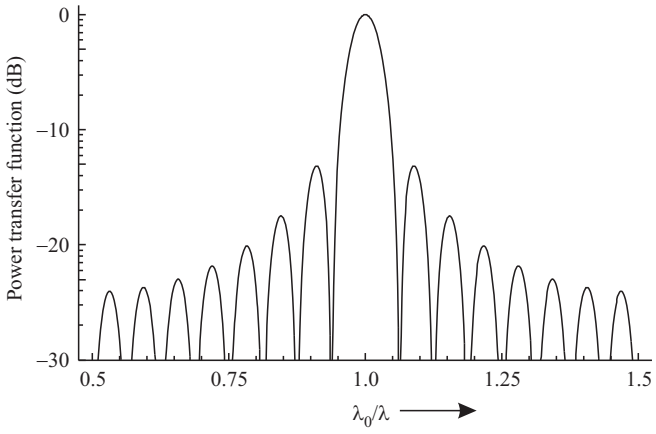

Figure 3.23 Transfer function of a multistage Mach-Zehnder interferometer.

We will now describe how an MZI can be used as a 1×2 demultiplexer. Since the device is reciprocal, it follows from the principles of electromagnetics that if the inputs and outputs are interchanged, it will act as a 2×1 multiplexer.

Consider a single MZI with a fixed value of the path difference ΔL. Let one of the inputs, say, input 1, be a wavelength division multiplexed signal with all the wavelengths chosen to coincide with the peaks or troughs of the transfer function. For concreteness, assume the propagation constant $\beta = 2\pi n_{\text{eff}}/\lambda$, where n_{eff} is the effective refractive index of the waveguide. The input wavelengths λ_i would have to be chosen such that $n_{\text{eff}}\Delta L/\lambda_i = m_i/2$ for some positive integer m_i. The wavelengths λ_i for which m is odd would then appear at the first output (since the transfer function is $\sin^2(m_i\pi/2)$), and the wavelengths for which m_i is even would appear at the second output (since the transfer function is $\cos^2(m_i\pi/2)$).

If there are only two wavelengths, one for which m_i is odd and the other for which m_i is even, we have a 1×2 demultiplexer. The construction of a $1 \times n$ demultiplexer when n is a power of two, using $n - 1$ MZIs, is left as an exercise (Problem 3.15). But there is a better method of constructing higher channel count demultiplexers, which we describe next.

3.3.8 *Arrayed Waveguide Grating*

An *arrayed waveguide grating* (AWG) is a generalization of the Mach-Zehnder interferometer. This device is illustrated in Figure 3.24. It consists of two multiport

couplers interconnected by an array of waveguides. The MZI can be viewed as a device where *two* copies of the same signal, but shifted in phase by different amounts, are added together. The AWG is a device where *several* copies of the same signal, but shifted in phase by different amounts, are added together.

The AWG has several uses. It can be used as an $n \times 1$ *wavelength multiplexer.* In this capacity, it is an n-input, 1-output device where the n inputs are signals at different wavelengths that are combined onto the single output. The inverse of this function, namely, $1 \times n$ *wavelength demultiplexing,* can also be performed using an AWG. Although these wavelength multiplexers and demultiplexers can also be built using MZIs interconnected in a suitable fashion, it is preferable to use an AWG. Relative to an MZI chain, an AWG has lower loss and flatter passband, and is easier to realize on an integrated-optic substrate. The input and output waveguides, the multiport couplers, and the arrayed waveguides are all fabricated on a single substrate. The substrate material is usually silicon, and the waveguides are silica, Ge-doped silica, or SiO_2-Ta_2O_5. Thirty-two–channel AWGs are commercially available, and smaller AWGs are being used in WDM transmission systems. Their temperature coefficient (0.01 nm/°C) is not as low as those of some other competing technologies such as fiber gratings and multilayer thin-film filters. So we will need to use active temperature control for these devices.

Another way to understand the working of the AWG as a demultiplexer is to think of the multiport couplers as lenses and the array of waveguides as a prism. The input coupler collimates the light from an input waveguide to the array of waveguides. The array of waveguides acts like a prism, providing a wavelength-dependent phase shift, and the output coupler focuses different wavelengths on different output waveguides.

The AWG can also be used as a static wavelength crossconnect. However, this wavelength crossconnect is not capable of achieving an arbitrary routing pattern. Although several interconnection patterns can be achieved by a suitable choice of

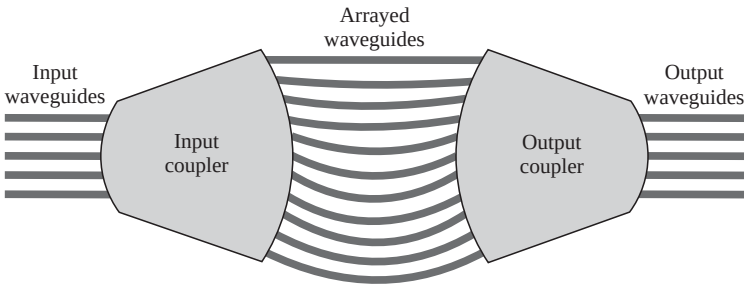

Figure 3.24 An arrayed waveguide grating.

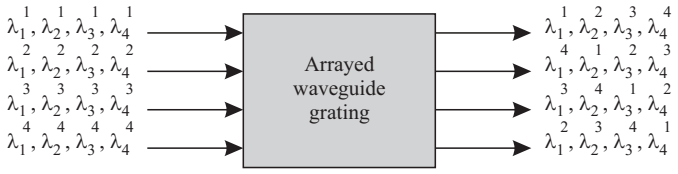

Figure 3.25 The crossconnect pattern of a static wavelength crossconnect constructed from an arrayed waveguide grating. The device routes signals from an input to an output based on their wavelength.

the wavelengths and the FSR of the device, the most useful one is illustrated in Figure 3.25. This figure shows a 4×4 static wavelength crossconnect using four wavelengths with one wavelength routed from each of the inputs to each of the outputs.

In order to achieve this interconnection pattern, the operating wavelengths and the FSR of the AWG must be chosen suitably. The FSR of the AWG is derived in Problem 3.17. Given the FSR, we leave the determination of the wavelengths to be used to achieve this interconnection pattern as another exercise (Problem 3.18).

Principle of Operation

Consider the AWG shown in Figure 3.24. Let the number of inputs and outputs of the AWG be denoted by n. Let the couplers at the input and output be $n \times m$ and $m \times n$ in size, respectively. Thus the couplers are interconnected by m waveguides. We will call these waveguides *arrayed waveguides* to distinguish them from the input and output waveguides. The lengths of these waveguides are chosen such that the difference in length between consecutive waveguides is a constant denoted by ΔL. The MZI is a special case of the AWG, where $n = m = 2$. We will now determine which wavelengths will be transmitted from a given input to a given output. The first coupler splits the signal into m parts. The relative phases of these parts are determined by the distances traveled in the coupler from the input waveguides to the arrayed waveguides. Denote the differences in the distances traveled (relative to any one of the input waveguides and any one of the arrayed waveguides) between input waveguide i and arrayed waveguide k by d_{ik}^{in}. Assume that arrayed waveguide k has a path length larger than arrayed waveguide $k-1$ by ΔL. Similarly, denote the differences in the distances traveled (relative to any one of the arrayed waveguides and any one of the output waveguides) between arrayed waveguide k and output

waveguide j by d_{kj}^{out}. Then, the relative phases of the signals from input i to output j traversing the m different paths between them are given by

$$\phi_{ijk} = \frac{2\pi}{\lambda}(n_1 d_{ik}^{\text{in}} + n_2 k\Delta L + n_1 d_{kj}^{\text{out}}), \qquad k = 1, \ldots, m. \tag{3.15}$$

Here, n_1 is the refractive index in the input and output directional couplers, and n_2 is the refractive index in the arrayed waveguides. From input i, those wavelengths λ, for which ϕ_{ijk}, $k = 1, \ldots, m$, differ by a multiple of 2π will add in phase at output j. The question is, Are there any such wavelengths?

If the input and output couplers are designed such that $d_{ik}^{\text{in}} = d_i^{\text{in}} + k\delta_i^{\text{in}}$ and $d_{kj}^{\text{out}} = d_j^{\text{out}} + k\delta_j^{\text{out}}$, then (3.15) can be written as

$$\begin{aligned}\phi_{ijk} &= \frac{2\pi}{\lambda}(n_1 d_i^{\text{in}} + n_1 d_j^{\text{out}}) \\ &\quad + \frac{2\pi k}{\lambda}(n_1\delta_i^{\text{in}} + n_2\Delta L + n_1\delta_j^{\text{out}}), \qquad k = 1, \ldots, m.\end{aligned} \tag{3.16}$$

Such a construction is possible and is called the *Rowland circle construction*. It is illustrated in Figure 3.26 and discussed further in Problem 3.16. Thus wavelengths λ that are present at input i and that satisfy $n_1\delta_i^{\text{in}} + n_2\Delta L + n_1\delta_j^{\text{out}} = p\lambda$ for some integer p add in phase at output j.

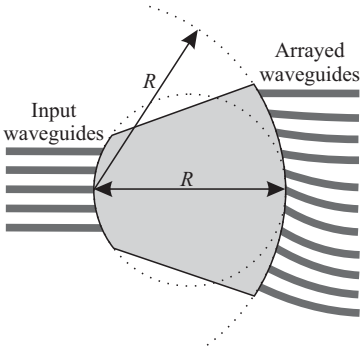

Figure 3.26 The Rowland circle construction for the couplers used in the AWG. The arrayed waveguides are located on the arc of a circle, called the *grating circle*, whose center is at the end of the central input (output) waveguide. Let the *radius* of this circle be denoted by R. The other input (output) waveguides are located on the arc of a circle whose *diameter* is equal to R; this circle is called the *Rowland circle*. The vertical spacing between the arrayed waveguides is chosen to be constant.

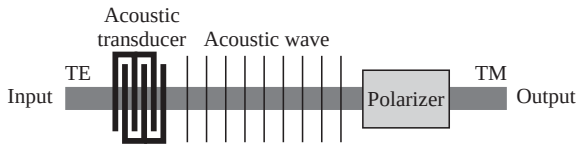

Figure 3.27 A simple AOTF. An acoustic wave introduces a grating whose pitch depends on the frequency of the acoustic wave. The grating couples energy from one polarization mode to another at a wavelength that satisfies the Bragg condition.

For use as a demultiplexer, all the wavelengths are present at the same input, say, input i. Therefore, if the wavelengths, $\lambda_1, \lambda_2, \ldots, \lambda_n$ in the WDM system satisfy $n_1\delta_i^{\text{in}} + n_2\Delta L + n_1\delta_j^{\text{out}} = p\lambda_j$ for some integer p, we infer from (3.16) that these wavelengths are demultiplexed by the AWG. Note that though δ_i^{in} and ΔL are necessary to define the precise set of wavelengths that are demultiplexed, the (minimum) spacing between them is independent of δ_i^{in} and ΔL, and determined primarily by δ_j^{out}.

Note in the preceding example that if wavelength λ'_j satisfies $n_1\delta_i^{\text{in}} + n_2\Delta L + n_1\delta_j^{\text{out}} = (p+1)\lambda'_j$, then both λ_j and λ'_j are "demultiplexed" to output j from input i. Thus like many of the other filter and multiplexer/demultiplexer structures we have studied, the AWG has a periodic response (in frequency), and all the wavelengths must lie within one FSR. The derivation of an expression for this FSR is left as an exercise (Problem 3.17).

3.3.9 *Acousto-Optic Tunable Filter*

The acousto-optic tunable filter is a versatile device. It is probably the only known *tunable* filter that is capable of selecting several wavelengths simultaneously. This capability can be used to construct a wavelength crossconnect, as we will explain later in this section.

The acousto-optic tunable filter (AOTF) is one example of several optical devices whose construction is based on the interaction of sound and light. Basically, an acoustic wave is used to create a Bragg grating in a waveguide, which is then used to perform the wavelength selection. Figure 3.27 shows a simple version of the AOTF. We will see that the operation of this AOTF is dependent on the state of polarization of the input signal. Figure 3.28 shows a more realistic polarization-independent implementation in integrated optics.

Principle of Operation

Consider the device shown in Figure 3.27. It consists of a waveguide constructed from a birefringent material and supporting only the lowest-order TE and TM modes (see Section 2.3.4). We assume that the input light energy is entirely in the TE mode. A *polarizer*, which selects only the light energy in the TM mode, is placed at the other end of the channel waveguide. If, somehow, the light energy in a narrow spectral range around the wavelength to be selected is converted to the TM mode, while the rest of the light energy remains in the TE mode, we have a wavelength-selective filter. This conversion is effected in an AOTF by launching an acoustic wave along, or opposite to, the direction of propagation of the light wave.

As a result of the propagation of the acoustic wave, the density of the medium varies in a periodic manner. The period of this density variation is equal to the wavelength of the acoustic wave. This periodic density variation acts as a Bragg grating. From the discussion of such gratings in Section 3.3.3, it follows that if the refractive indices n_{TE} and n_{TM} of the TE and TM modes satisfy the Bragg condition

$$\frac{n_{TM}}{\lambda} = \frac{n_{TE}}{\lambda} \pm \frac{1}{\Lambda}, \tag{3.17}$$

then light couples from one mode to the other. Thus light energy in a narrow spectral range around the wavelength λ that satisfies (3.17) undergoes TE to TM mode conversion. Thus the device acts as a narrow bandwidth filter when only light energy in the TE mode is input and only the light energy in the TM mode is selected at the output, as shown in Figure 3.27.

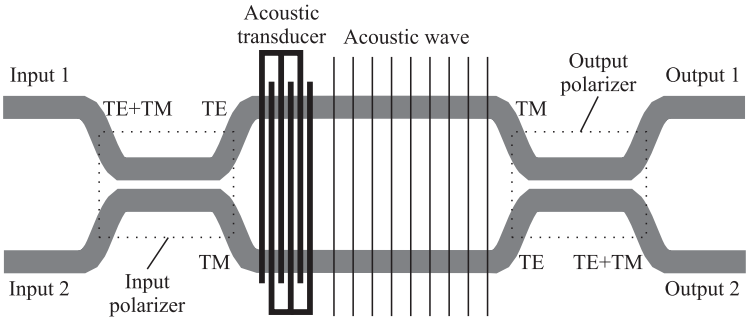

Figure 3.28 A polarization-independent integrated-optics AOTF. A polarizer splits the input signal into its constituent polarization modes, and each mode is converted in two separate arms, before being recombined at the output.

In $LiNbO_3$, the TE and TM modes have refractive indices n_{TE} and n_{TM} that differ by about 0.07. If we denote this refractive index difference by (Δn), the Bragg condition (3.17) can be written as

$$\lambda = \Lambda(\Delta n). \tag{3.18}$$

The wavelength that undergoes mode conversion and thus lies in the passband of the AOTF can be selected, or tuned, by suitably choosing the acoustic wavelength Λ. In order to select a wavelength of 1.55 μm, for $(\Delta n) = 0.07$, using (3.18), the acoustic wavelength is about 22 μm. Since the velocity of sound in $LiNbO_3$ is about 3.75 km/s, the corresponding RF frequency is about 170 MHz. Since the RF frequency is easily tuned, the wavelength selected by the filter can also be easily tuned. The typical insertion loss is about 4 dB.

The AOTF considered here is a polarization-dependent device since the input light energy is assumed to be entirely in the TE mode. A polarization-independent AOTF, shown in Figure 3.28, can be realized in exactly the same manner as a polarization-independent isolator by decomposing the input light signal into its TE and TM constituents and sending each constituent separately through the AOTF and recombining them at the output.

Transfer Function

Whereas the Bragg condition determines the wavelength that is selected, the width of the filter passband is determined by the length of the acousto-optic interaction. The longer this interaction, and hence the device, the narrower the passband. It can be shown that the wavelength dependence of the fraction of the power transmitted by the AOTF is given by

$$T(\lambda) = \frac{\sin^2\left((\pi/2)\sqrt{1 + (2\Delta\lambda/\Delta)^2}\right)}{1 + (2\Delta\lambda/\Delta)^2}.$$

This is plotted in Figure 3.29. Here $\Delta\lambda = \lambda - \lambda_0$, where λ_0 is the optical wavelength that satisfies the Bragg condition, and $\Delta = \lambda_0^2/l\Delta n$ is a measure of the filter passband width. Here, l is the length of the device (or, more correctly, the length of the acousto-optic interaction). It can be shown that the full width at half-maximum (FWHM) bandwidth of the filter is $\approx 0.8\Delta$ (Problem 3.20). This equation clearly shows that the longer the device, the narrower the passband. However, there is a trade-off here: the tuning speed is inversely proportional to l. This is because the tuning speed is essentially determined by the time it takes for a sound wave to travel the length of the filter.

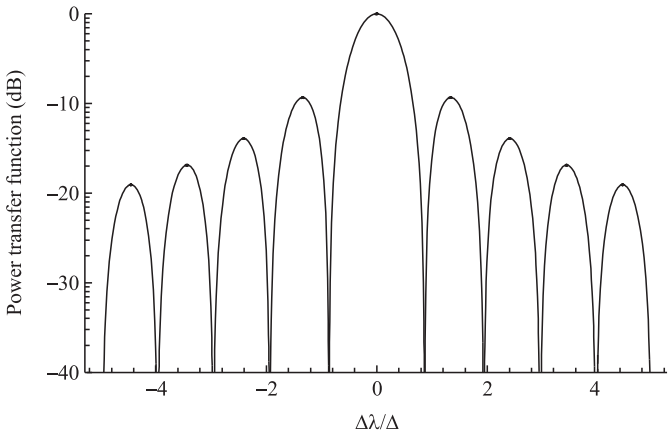

Figure 3.29 The power transfer function of the acousto-optic tunable filter.

AOTF as a Wavelength Crossconnect

The polarization-independent AOTF illustrated in Figure 3.28 can be used as a two-input, two-output dynamic wavelength crossconnect. We studied the operation of this device as a filter earlier; in this case, only one of the inputs was active. We leave it as an exercise (Problem 3.21) to show that when the second input is also active, the energy at the wavelength λ satisfying the Bragg phase-matching condition (3.18) is *exchanged* between the two ports. This is illustrated in Figure 3.30(a), where the wavelength λ_1 satisfies the Bragg condition and is exchanged between the ports.

Now the AOTF has one remarkable property that is not shared by any other tunable filter structure we know. By launching multiple acoustic waves *simultaneously*, the Bragg condition (3.18) can be satisfied for multiple optical wavelengths simultaneously. Thus multiple wavelength exchanges can be accomplished simultaneously between two ports with a single device of the form shown in Figure 3.28. This is illustrated in Figure 3.30(b), where the wavelengths λ_1 and λ_4 are exchanged between the ports. Thus this device performs the same routing function as the static crossconnect of Figure 3.7. However, the AOTF is a completely general two-input, two-output *dynamic* crossconnect since the routing pattern, or the set of wavelengths to be exchanged, can be changed easily by varying the frequencies of the acoustic waves launched in the device. In principle, larger dimensional dynamic crossconnects (with more input and output ports) can be built by suitably cascading 2×2 crossconnects. We will see in Section 3.7, however, that there are better ways of building large-scale crossconnects.

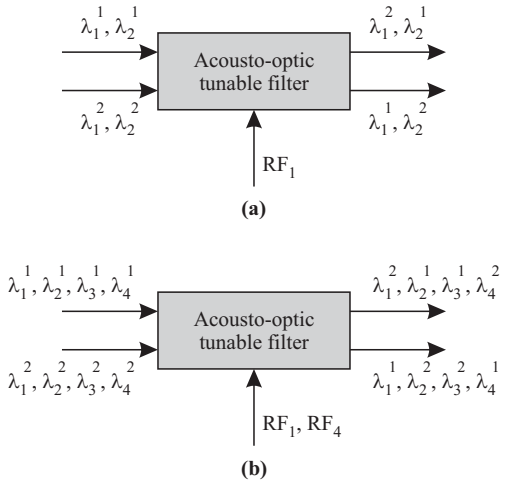

Figure 3.30 Wavelength crossconnects constructed from acousto-optic tunable filters. (a) The wavelength λ_1 is exchanged between the two ports. (b) The wavelengths λ_1 and λ_4 are simultaneously exchanged between the two ports by the simultaneous launching of two appropriate acoustic waves.

As of this writing, the AOTF has not yet lived up to its promise either as a versatile tunable filter or a wavelength crossconnect. One reason for this is the high level of crosstalk that is present in the device. As can be seen from Figure 3.29, the first side lobe in its power transfer function is not even 10 dB below the peak transmission. This problem can be alleviated to some extent by cascading two such filters. In fact, the cascade can even be built on a single substrate. But even then the first side lobe would be less than 20 dB below the peak transmission. It is harder to cascade more such devices without facing other problems such as an unacceptably high transmission loss. Another reason for the comparative failure of the AOTF today is that the passband width is fairly large (100 GHz or more) even when the acousto-optic interaction length is around 1 inch (Problem 3.22). This makes it unsuitable for use in dense WDM systems where channel spacings are now down to 50 GHz. Devices with larger interaction lengths are more difficult to fabricate. However, some recent theoretical work [Son95] indicates that some of these problems, particularly crosstalk, may be solvable. The crosstalk problems that arise in AOTFs when used as wavelength crossconnects are discussed in detail in [Jac96].

3.3.10 High Channel Count Multiplexer Architectures

With the number of wavelengths continuously increasing, designing multiplexers and demultiplexers to handle large numbers of wavelengths has become an important problem. The desired attributes of these devices are the same as what we saw at the beginning of Section 3.3. Our discussion will be based on demultiplexers, but these demultiplexers can all be used as multiplexers as well. In fact, in bidirectional applications, where some wavelengths are transmitted in one direction over a fiber and others in the opposite direction over the same fiber, the same device acts as a multiplexer for some wavelengths and a demultiplexer for others. We describe several architectural approaches to construct high channel count demultiplexers below.

Serial

In this approach, the demultiplexing is done one wavelength at a time. The demultiplexer consists of W filter stages in series, one for each of the W wavelengths. Each filter stage demultiplexes a wavelength and allows the other wavelengths to pass through. The architecture of the dielectric thin-film demultiplexer shown in Figure 3.20 is an example. One advantage of this architecture is that the filter stages can potentially be added one at a time, as more wavelengths are added. This allows a "pay as you grow" approach.

Serial approaches work for demultiplexing relatively small numbers of channels but do not scale to handle a large number of channels. This is because the insertion loss (in decibels) of the demultiplexer increases almost linearly with the number of channels to be demultiplexed. Moreover, different channels see different insertion losses based on the order in which the wavelengths are demultiplexed, which is not a desirable feature.

Single Stage

Here, all the wavelengths are demultiplexed together in a single stage. The AWG shown in Figure 3.24 is an example of such an architecture. This approach provides relatively lower losses and better loss uniformity, compared to the serial approach. However, the number of channels that can be demultiplexed is limited by the maximum number of channels that can be handled by a single device, typically around 40 channels in commercially available AWGs today.

Multistage Banding

Going to larger channel counts requires the use of multiple demultiplexing stages, due to the limitations of the serial and single-stage approaches discussed above. A popular approach used today is to divide the wavelengths into *bands*. For example,

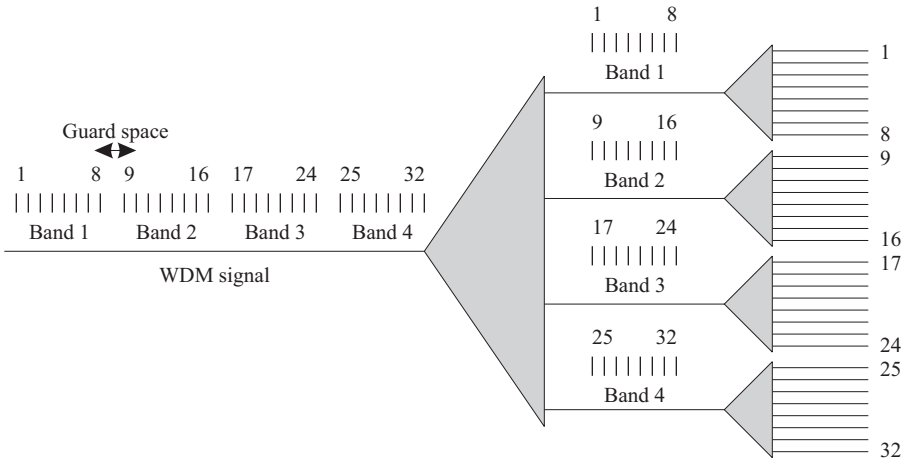

Figure 3.31 A two-stage demultiplexing approach using bands. A 32-channel demultiplexer is realized using four bands of 8 channels each.

a total of 32 wavelengths may be divided into four bands, each with 8 wavelengths. The demultiplexing is done in two stages, as shown in Figure 3.31. In the first the set of wavelengths is demultiplexed into bands. In the second stage, the bands are demultiplexed, and individual wavelengths are extracted. The scheme can be extended to more than two stages as well. It is also modular in that the demultiplexers in the second stage (or last stage in a multistage scheme) can be populated one band at a time.

One drawback of the banding approach is that we will usually need to leave a "guard" space between bands, as shown in Figure 3.31. This guard space allows the first-stage filters to be designed to provide adequate crosstalk suppression while retaining a low insertion loss.

Multistage Interleaving

Interleaving provides another approach to realizing large channel count demultiplexers. A two-stage *interleaver* is shown in Figure 3.32. In this approach the first stage separates the wavelengths into two groups. The first group consists of wavelengths 1, 3, 5, . . . and the second group consists of wavelengths 2, 4, 6, ... The second stage extracts the individual wavelengths. This approach is also modular in the sense that the last stage of demultiplexers can be populated as needed. More than two stages can be used if needed as well.

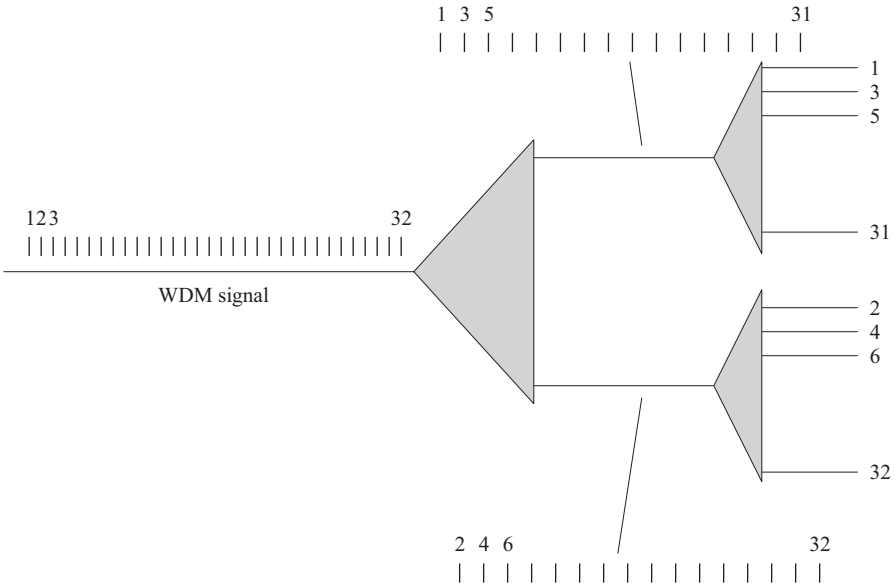

Figure 3.32 A two-stage multiplexing approach using interleaving. In this 32-channel demultiplexer, the first stage picks out every alternate wavelength, and the second stage extracts the individual wavelength.

A significant benefit of this approach is that the filters in the last stage can be much wider than the channel width. As an example, suppose we want to demultiplex a set of 32 channels spaced 50 GHz apart. After the first stage of demultiplexing, the channels are spaced 100 GHz apart, as shown in Figure 3.32. So demultiplexers with a broader passband suitable for demultiplexing 100 GHz spaced channels can be used in the second stage. In contrast, the single-stage or serial approach would require the use of demultiplexers capable of demultiplexing 50 GHz spaced channels, which are much more difficult to build. Carrying this example further, the second stage itself can in turn be made up of two stages. The first stage extracts every other 100 GHz channel, leading to a 200 GHz interchannel spacing after this stage. The final stage can then use even broader filters to extract the individual channels. Another advantage of this approach is that no guard bands are required in the channel plan.

The challenges with the interleaving approach lie in realizing the demultiplexers that perform the interleaving at all the levels except the last level. In principle,

any periodic filter can be used as an interleaver by matching its period to the desired channel spacing. For example, a fiber-based Mach-Zehnder interferometer is a common choice. These devices are now commercially available, and interleaving is becoming a popular approach toward realizing high channel count multiplexers and demultiplexers.

3.4 Optical Amplifiers

In an optical communication system, the optical signals from the transmitter are attenuated by the optical fiber as they propagate through it. Other optical components, such as multiplexers and couplers, also add loss. After some distance, the cumulative loss of signal strength causes the signal to become too weak to be detected. Before this happens, the signal strength has to be restored. Prior to the advent of optical amplifiers over the last decade, the only option was to regenerate the signal, that is, receive the signal and retransmit it. This process is accomplished by *regenerators*. A regenerator converts the optical signal to an electrical signal, cleans it up, and converts it back into an optical signal for onward transmission.

Optical amplifiers offer several advantages over regenerators. On one hand, regenerators are specific to the bit rate and modulation format used by the communication system. On the other hand, optical amplifiers are insensitive to the bit rate or signal formats. Thus a system using optical amplifiers can be more easily upgraded, for example, to a higher bit rate, without replacing the amplifiers. In contrast, in a system using regenerators, such an upgrade would require all the regenerators to be replaced. Furthermore, optical amplifiers have fairly large gain bandwidths, and as a consequence, a single amplifier can simultaneously amplify several WDM signals. In contrast, we would need a regenerator for each wavelength. Thus optical amplifiers have become essential components in high-performance optical communication systems.

Amplifiers, however, are not perfect devices. They introduce additional noise, and this noise accumulates as the signal passes through multiple amplifiers along its path due to the analog nature of the amplifier. The spectral shape of the gain, the output power, and the transient behavior of the amplifier are also important considerations for system applications. Ideally, we would like to have a sufficiently high output power to meet the needs of the network application. We would also like the gain to be flat over the operating wavelength range and to be insensitive to variations in input power of the signal. We will study the impact of optical amplifiers on the physical layer design of the system in Chapters 4 and 5. Here we explore their principle of operation.

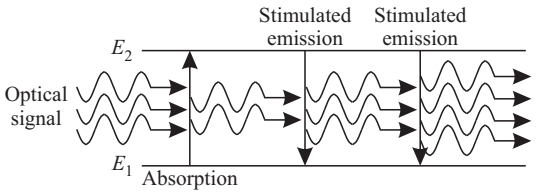

Figure 3.33 Stimulated emission and absorption in an atomic system with two energy levels.

We will consider three different types of amplifiers: *erbium-doped fiber amplifiers, Raman amplifiers,* and *semiconductor optical amplifiers.*

3.4.1 Stimulated Emission

In all the amplifiers we consider, the key physical phenomenon behind signal amplification is *stimulated emission* of radiation by atoms in the presence of an electromagnetic field. (This is not true of fiber Raman or fiber Brillouin amplifiers, which make use of fiber nonlinearities, but we do not treat these here.) This field is an optical signal in the case of optical amplifiers. Stimulated emission is the principle underlying the operation of lasers as well; we will study lasers in Section 3.5.1.

According to the principles of quantum mechanics, any physical system (for example, an atom) is found in one of a discrete number of energy levels. Accordingly, consider an atom and two of its energy levels, E_1 and E_2, with $E_2 > E_1$. An electromagnetic field whose frequency f_c satisfies $hf_c = E_2 - E_1$ induces transitions of atoms between the energy levels E_1 and E_2. Here, h is Planck's constant (6.63×10^{-34} J s). This process is depicted in Figure 3.33. Both kinds of transitions, $E_1 \rightarrow E_2$ and $E_2 \rightarrow E_1$, occur. $E_1 \rightarrow E_2$ transitions are accompanied by *absorption* of photons from the incident electromagnetic field. $E_2 \rightarrow E_1$ transitions are accompanied by the *emission* of photons of energy hf_c, the same energy as that of the incident photons. This emission process is termed *stimulated emission* to distinguish it from another kind of emission called *spontaneous emission*, which we will discuss later. Thus if stimulated emission were to dominate over absorption—that is, the incident signal causes more $E_2 \rightarrow E_1$ transitions than $E_1 \rightarrow E_2$ transitions—we would have a net increase in the number of photons of energy hf_c and an amplification of the signal. Otherwise, the signal will be attenuated.

It follows from the theory of quantum mechanics that the rate of the $E_1 \rightarrow E_2$ transitions per atom *equals* the rate of the $E_2 \rightarrow E_1$ transitions *per atom*. Let this common rate be denoted by r. If the populations (number of atoms) in the energy levels E_1 and E_2 are N_1 and N_2, respectively, we have a net increase in power (energy per unit time) of $(N_2 - N_1)rhf_c$. Clearly, for amplification to occur, this must be positive, that is, $N_2 > N_1$. This condition is known as *population inversion*. The reason for this term is that, at thermal equilibrium, lower energy levels are more highly populated, that is, $N_2 < N_1$. Therefore, at thermal equilibrium, we have only absorption of the input signal. In order for amplification to occur, we must *invert* the relationship between the populations of levels E_1 and E_2 that prevails under thermal equilibrium.

Population inversion can be achieved by supplying additional energy in a suitable form to pump the electrons to the higher energy level. This additional energy can be in optical or electrical form.

3.4.2 Spontaneous Emission

Before describing the operation of the different types of amplifiers, it is important to understand the impact of spontaneous emission. Consider again the atomic system with the two energy levels discussed earlier. Independent of any external radiation that may be present, atoms in energy level E_2 transit to the lower energy level E_1, emitting a photon of energy hf_c. The spontaneous emission rate per atom from level E_2 to level E_1 is a characteristic of the system, and its reciprocal, denoted by τ_{21}, is called the *spontaneous emission lifetime*. Thus, if there are N_2 atoms in level E_2, the rate of spontaneous emission is N_2/τ_{21}, and the spontaneous emission power is hf_cN_2/τ_{21}.

The spontaneous emission process does not contribute to the gain of the amplifier (to first order). Although the emitted photons have the same energy hf_c as the incident optical signal, they are emitted in random directions, polarizations, and phase. This is unlike the stimulated emission process, where the emitted photons not only have the same energy as the incident photons but also the same direction of propagation, phase, and polarization. This phenomenon is usually described by saying that the stimulated emission process is *coherent*, whereas the spontaneous emission process is *incoherent*.

Spontaneous emission has a deleterious effect on the system. The amplifier treats spontaneous emission radiation as another electromagnetic field at the frequency hf_c, and the spontaneous emission also gets amplified, in addition to the incident optical signal. This *amplified spontaneous emission* (ASE) appears as noise at the output of the amplifier. The implications of ASE for the design of optical communication

systems are discussed in Chapters 4 and 5. In addition, in some amplifier designs, the ASE can be large enough to *saturate* the amplifier. Saturation effects are explored in Chapter 5.

3.4.3 Erbium-Doped Fiber Amplifiers

An erbium-doped fiber amplifier (EDFA) is shown in Figure 3.34. It consists of a length of silica fiber whose core is doped with ionized atoms (ions), Er^{3+}, of the rare earth element erbium. This fiber is pumped using a pump signal from a laser, typically at a wavelength of 980 nm or 1480 nm. In order to combine the output of the pump laser with the input signal, the doped fiber is preceded by a wavelength-selective coupler.

At the output, another wavelength-selective coupler may be used if needed to separate the amplified signal from any remaining pump signal power. Usually, an isolator is used at the input and/or output of any amplifier to prevent reflections into the amplifier. We will see in Section 3.5 that reflections can convert the amplifier into a laser, making it unusable as an amplifier.

A combination of several factors has made the EDFA the amplifier of choice in today's optical communication systems: (1) the availability of compact and reliable high-power semiconductor pump lasers, (2) the fact that it is an all-fiber device, making it polarization independent and easy to couple light in and out of it, (3) the simplicity of the device, and (4) the fact that it introduces no crosstalk when amplifying WDM signals. This last aspect is discussed later in the context of semiconductor optical amplifiers.

Principle of Operation

Three of the energy levels of erbium ions in silica glass are shown in Figure 3.35 and are labeled E_1, E_2, and E_3 in order of increasing energy. Several other levels in Er^{3+} are not shown. Each energy level that appears as a discrete line in an isolated

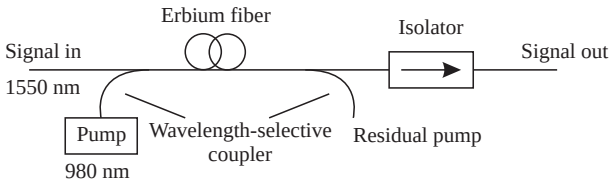

Figure 3.34 An erbium-doped fiber amplifier.

ion of erbium is split into multiple energy levels when these ions are introduced into silica glass. This process is termed *Stark splitting*. Moreover, glass is not a crystal and thus does not have a regular structure. Thus the Stark splitting levels introduced are slightly different for individual erbium ions, depending on the local surroundings seen by those ions. Macroscopically, that is, when viewed as a collection of ions, this has the effect of spreading each discrete energy level of an erbium ion into a continuous *energy band*. This spreading of energy levels is a useful characteristic for optical amplifiers since they increase the frequency or wavelength range of the signals that can be amplified. Within each energy band, the erbium ions are distributed in the various levels within that band in a nonuniform manner by a process known as *thermalization*. It is due to this thermalization process that an amplifier is capable of amplifying several wavelengths simultaneously. Note that Stark splitting denotes the phenomenon by which the energy levels of free erbium ions are split into a number of levels, or into an energy band, when the ion is introduced into silica glass. Thermalization refers to the process by which the erbium ions are distributed within the various (split) levels constituting an energy band.

Recall from our discussion of the two-energy-level atomic system that only an optical signal at the frequency f_c satisfying $hf_c = E_2 - E_1$ could be amplified in that

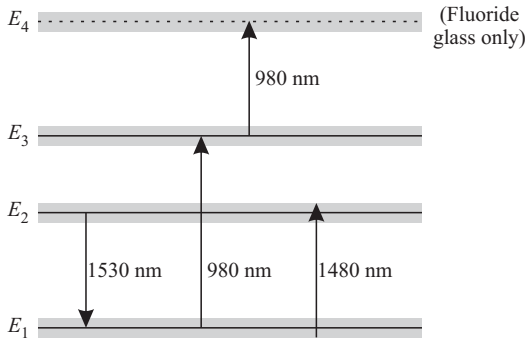

Figure 3.35 Three energy levels E_1, E_2, and E_3 of Er^{3+} ions in silica glass. The fourth energy level, E_4, is present in fluoride glass but not in silica glass. The energy levels are spread into bands by the Stark splitting process. The difference between the energy levels is labeled with the wavelength in nm of the photon corresponding to it. The upward arrows indicate wavelengths at which the amplifier can be pumped to excite the ions into the higher energy level. The 980 nm transition corresponds to the band gap between the E_1 and E_3 levels. The 1480 nm transition corresponds to the gap between the bottom of the E_1 band to the top of the E_2 band. The downward transition represents the wavelength of photons emitted due to spontaneous and stimulated emission.

case. If these levels are spread into bands, all frequencies that correspond to the energy difference between some energy in the E_2 band and some energy in the E_1 band can be amplified. In the case of erbium ions in silica glass, the set of frequencies that can be amplified by stimulated emission from the E_2 band to the E_1 band corresponds to the wavelength range 1525–1570 nm, a bandwidth of 50 nm, with a peak around 1532 nm. By a lucky coincidence, this is exactly one of the low-attenuation windows of standard optical fiber that optical communication systems use.

Denote ionic population in level E_i by N_i, $i = 1, 2, 3$. In thermal equilibrium, $N_1 > N_2 > N_3$. The population inversion condition for stimulated emission from E_2 to E_1 is $N_2 > N_1$ and can be achieved by a combination of absorption and spontaneous emission as follows. The energy difference between the E_1 and E_3 levels corresponds to a wavelength of 980 nm. So if optical power at 980 nm—called the *pump power*—is injected into the amplifier, it will cause transitions from E_1 to E_3 and vice versa. Since $N_1 > N_3$, there will be a net absorption of the 980 nm power. This process is called *pumping*.

The ions that have been raised to level E_3 by this process will quickly transit to level E_2 by the spontaneous emission process. The lifetime for this process, τ_{32}, is about 1 μs. Atoms from level E_2 will also transit to level E_1 by the spontaneous emission process, but the lifetime for this process, τ_{21}, is about 10 ms, which is much larger than the E_3 to E_2 lifetime. Moreover, if the pump power is sufficiently large, ions that transit to the E_1 level are rapidly raised again to the E_3 level only to transit to the E_2 level again. The net effect is that most of the ions are found in level E_2, and thus we have population inversion between the E_2 and E_1 levels. Therefore, if simultaneously a signal in the 1525–1570 nm band is injected into the fiber, it will be amplified by stimulated emission from the E_2 to the E_1 level.

Several levels other than E_3 are higher than E_2 and, in principle, can be used for pumping the amplifier. But the pumping process is more efficient, that is, uses less pump power for a given gain, at 980 nm than these other wavelengths. Another possible choice for the pump wavelength is 1480 nm. This choice corresponds to absorption from the bottom sublevel of the E_1 band to the top sublevel of the E_2 band itself. Pumping at 1480 nm is not as efficient as 980 nm pumping. Moreover, the degree of population inversion that can be achieved by 1480 nm pumping is lower. The higher the population inversion, the lower the noise figure of the amplifier. Thus 980 nm pumping is preferred to realize low-noise amplifiers. However, higher-power pump lasers are available at 1480 nm, compared to 980 nm, and thus 1480 nm pumps find applications in amplifiers designed to yield high output powers. Another advantage of the 1480 nm pump is that the pump power can also propagate with low loss in the silica fiber that is used to carry the signals. Therefore, the pump laser can be located remotely from the amplifier itself. This feature is used in some systems to avoid placing any active components in the middle of the link.

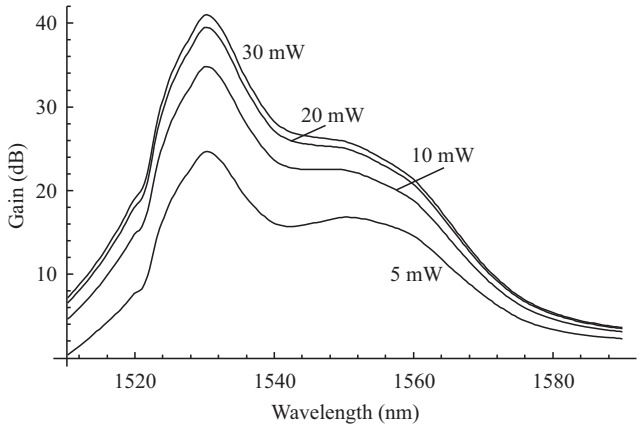

Figure 3.36 The gain of a typical EDFA as a function of the wavelength for four different values of the pump power, obtained through simulations. The length of the doped fiber is taken to be 15 m and 980 nm pumping is assumed.

Gain Flatness

Since the population levels at the various levels within a band are different, the gain of an EDFA becomes a function of the wavelength. In Figure 3.36, we plot the gain of a typical EDFA as a function of the wavelength for different values of the pump power. When such an EDFA is used in a WDM communication system, different WDM channels undergo different degrees of amplification. This is a critical issue, particularly in WDM systems with cascaded amplifiers, and is discussed in Section 5.5.2.

One way to improve the flatness of the amplifier gain profile is to use fluoride glass fiber instead of silica fiber, doped with erbium [Cle94]. Such amplifiers are called erbium-doped fluoride fiber amplifiers (EDFFAs). The fluoride glass produces a naturally flatter gain spectrum compared to silica glass. However, there are a few drawbacks to using fluoride glass. The noise performance of EDFFAs is poorer than EDFAs. One reason is that they must be pumped at 1480 nm and cannot be pumped at 980 nm. This is because fluoride glass has an additional higher energy level E_4 above the E_3 level, as shown in Figure 3.35, with the difference in energies between these two levels corresponding to 980 nm. This causes the 980 nm pump power to be absorbed for transitions from the E_3 to E_4 level, which does not produce useful gain. This phenomenon is called *excited state absorption*.

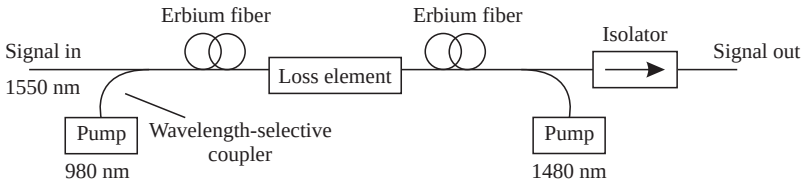

Figure 3.37 A two-stage erbium-doped fiber amplifier with a loss element inserted between the first and second stage.

In addition to this drawback, fluoride fiber itself is difficult to handle. It is brittle, difficult to splice with conventional fiber, and susceptible to moisture. Nevertheless, EDFFAs are now commercially available devices.

Another approach to flatten the EDFA gain is to use a filter inside the amplifier. The EDFA has a relatively high gain at 1532 nm, which can be reduced by using a notch filter in that wavelength region inside the amplifier. Some of the filters described in Section 3.3 can be used for this purpose. Long-period fiber gratings and dielectric thin-film filters are currently the leading candidates for this application.

Multistage Designs

In practice, most amplifiers deployed in real systems are more complicated than the simple structure shown in Figure 3.34. Figure 3.37 shows a more commonly used two-stage design. The two stages are optimized differently. The first stage is designed to provide high gain and low noise, and the second stage is designed to produce high output power. As we will see in Problem 4.5 in Chapter 4, the noise performance of the whole amplifier is determined primarily by the first stage. Thus this combination produces a high-performance amplifier with low noise and high output power. Another important consideration in the design is to provide redundancy in the event of the failure of a pump, the only active component of the amplifier. The amplifier shown in the figure uses two pumps and can be designed so that the failure of one pump has only a small impact on the system performance. Another feature of the two-stage design that we will address in Problem 4.5 is that a loss element can be placed between the two stages with negligible impact on the performance. This loss element may be a gain-flattening filter, a simple optical add/drop multiplexer, or a dispersion compensation module used to compensate for accumulated dispersion along the link.

L-Band EDFAs

So far, we have focused mostly on EDFAs operating in the C-band (1530–1565 nm). Erbium-doped fiber, however, has a relatively long tail to the gain shape extending well beyond this range to about 1605 nm. This has stimulated the development of systems in the so-called L-band from 1565 to 1625 nm. Note that current L-band EDFAs do not yet cover the top portion of this band from 1610 to 1625 nm.

L-band EDFAs operate on the same principle as C-band EDFAs. However, there are significant differences in the design of L- and C-band EDFAs. The gain spectrum of erbium is much flatter intrinsically in the L-band than in the C-band. This makes it easier to design gain-flattening filters for the L-band. However, the erbium gain coefficient in the L-band is about three times smaller than in the C-band. This necessitates the use of either much longer doped fiber lengths or fiber with higher erbium doping concentrations. In either case, the pump powers required for L-band EDFAs are much higher than their C-band counterparts. Due to the smaller absorption cross sections in the L-band, these amplifiers also have higher amplified spontaneous emission. Finally, many of the other components used inside the amplifier, such as isolators and couplers, exhibit wavelength-dependent losses and are therefore specified differently for the L-band than for the C-band. There are several other subtleties associated with L-band amplifiers; see [Flo00] for a summary.

As a result of the significant differences between C- and L-band amplifiers, these amplifiers are usually realized as separate devices rather than as a single device. In a practical system application, the C- and L-band wavelengths on a fiber are first separated by a demultiplexer, then amplified by separate amplifiers, and recombined together afterward.

3.4.4 Raman Amplifiers

In Section 2.5.3, we studied stimulated Raman scattering (SRS) as one of the nonlinear impairments that affect signals propagating through optical fiber. The same nonlinearity can be exploited to provide amplification as well. As we saw in Figure 2.17, the Raman gain spectrum is fairly broad, and the peak of the gain is centered about 13 THz below the frequency of the pump signal used. In the near-infrared region of interest to us, this corresponds to a wavelength separation of about 100 nm. Therefore, by pumping a fiber using a high-power pump laser, we can provide gain to other signals, with a peak gain obtained 13 THz below the pump frequency. For instance, using pumps around 1460–1480 nm provides Raman gain in the 1550–1600 nm window.

A few key attributes distinguish Raman amplifiers from EDFAs. Unlike EDFAs, we can use the Raman effect to provide gain at any wavelength. An EDFA provides

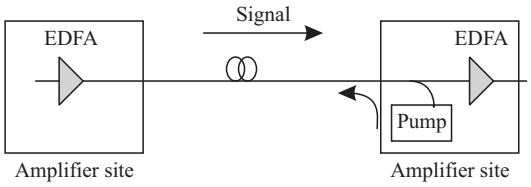

Figure 3.38 Distributed Raman amplifier using a backward propagating pump, shown operating along with discrete erbium-doped fiber amplifiers.

gain in the C- and L-bands (1528–1605 nm). Thus Raman amplification can potentially open up other bands for WDM, such as the 1310 nm window, or the so-called S-band lying just below 1528 nm. Also, we can use multiple pumps at different wavelengths and different powers simultaneously to tailor the overall Raman gain shape.

Second, Raman amplification relies on simply pumping the same silica fiber used for transmitting the data signals, so that it can be used to produce a *lumped* or *discrete* amplifier, as well as a *distributed* amplifier. In the lumped case, the Raman amplifier consists of a sufficiently long spool of fiber along with the appropriate pump lasers in a package. In the distributed case, the fiber can simply be the fiber span of interest, with the pump attached to one end of the span, as shown in Figure 3.38.

Today the most popular use of Raman amplifiers is to complement EDFAs by providing additional gain in a distributed manner in ultra-long-haul systems. The biggest challenge in realizing Raman amplifiers lies in the pump source itself. These amplifiers require high-power pump sources of the order of 1 W or more, at the right wavelength. We will study some techniques for realizing these pump sources in Section 3.5.5.

The noise sources in Raman amplifiers are somewhat different from EDFAs. The Raman gain responds instantaneously to the pump power. Therefore fluctuations in pump power will cause the gain to vary and will appear as crosstalk to the desired signals. This is not the case with EDFAs. We will see in Section 3.4.6 that the response time of the gain is much slower—on the order of milliseconds—in those devices. Therefore, for Raman amplifiers, it is important to keep the pump at a constant power. Having the pump propagate in the opposite direction to the signal helps dramatically because fluctuations in pump power are then averaged over the propagation time over the fiber. To understand this, first consider the case where the pump propagates along with the signal in the same direction. The two waves travel at approximately the same velocity. In this case, when the pump power is high at the input, the signal sees high gain, and when the power is low, the signal sees a lower

gain. Now consider the case when the signal and pump travel in opposite directions. To keep things simple, suppose that the pump power varies between two states: high and low. As the signal propagates through the fiber, whenever it overlaps with the pump signal in the high power state, it sees a high gain. When it overlaps with the pump signal in the low power state, it sees a lower gain. If the pump fluctuations are relatively fast compared to the propagation time of the signal across the fiber, the gain variations average out, and by the time the signal exits the fiber, it has seen a constant gain.

Another major concern with Raman amplifiers is crosstalk between the WDM signals due to Raman amplification. A modulated signal at a particular wavelength depletes the pump power, effectively imposing the same modulation on the pump signal. This modulation on the pump then affects the gain seen by the next wavelength, effectively appearing as crosstalk on that wavelength. Again, having the pump propagate in the opposite direction to the signal dramatically reduces this effect. For these reasons, most Raman amplifiers use a counterpropagating pump geometry.

Another source of noise is due to the back-reflections of the pump signal caused by Rayleigh scattering in the fiber. Spontaneous emission noise is relatively low in Raman amplifiers. This is usually the dominant source of noise because, by careful design, we can eliminate most of the other noise sources.

3.4.5 *Semiconductor Optical Amplifiers*

Semiconductor optical amplifiers (SOAs) actually preceded EDFAs, although we will see that they are not as good as EDFAs for use as amplifiers. However, they are finding other applications in switches and wavelength converter devices. Moreover, the understanding of SOAs is key to the understanding of semiconductor lasers, the most widely used transmitters today.

Figure 3.39 shows the block diagram of a semiconductor optical amplifier. The SOA is essentially a *pn*-junction. As we will explain shortly, the depletion layer that is formed at the junction acts as the *active region*. Light is amplified through stimulated emission when it propagates through the active region. For an amplifier, the two ends of the active region are given an antireflection (AR) coating to eliminate ripples in the amplifier gain as a function of wavelength. Alternatively, the facets may also be angled slightly to reduce the reflection. In the case of a semiconductor laser, there would be no AR coating.

SOAs differ from EDFAs in the manner in which population inversion is achieved. First, the populations are not those of ions in various energy states but of *carriers—electrons* or *holes*—in a semiconductor material. Holes can also be thought of as charge carriers similar to electrons except that they have a positive charge. A semiconductor consists of two bands of electron energy levels: a band of

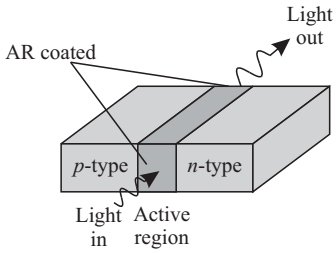

Figure 3.39 Block diagram of a semiconductor optical amplifier. Amplification occurs when light propagates through the active region. The facets are given an antireflective coating to prevent undesirable reflections, which cause ripple in the amplifier gain.

low-mobility levels called the *valence band* and a band of high-mobility levels called the *conduction band*. These bands are separated by an energy difference called the *bandgap* and denoted by E_g. No energy levels exist in the bandgap. Consider a *p*-type semiconductor material. At thermal equilibrium, there is only a very small concentration of electrons in the conduction band of the material, as shown in Figure 3.40(a). With reference to the previous discussion of EDFAs, it is convenient to think of the conduction band as the higher energy band E_2, and the valence band as the lower energy band E_1. The terms *higher* and *lower* refer to the electron energy in these bands. (Note that if we were considering an *n*-type semiconductor, we would be considering hole energies rather than electron energies, the conduction band would be the lower energy band E_1, and the valence band, the higher energy band E_2.) In the population inversion condition, the electron concentration in the conduction band is much higher, as shown in Figure 3.40(b). This increased concentration is such that, in the presence of an optical signal, there are more electrons transiting from the conduction band to the valence band by the process of stimulated emission than there are electrons transiting from the valence band to the conduction band by the process of absorption. In fact, for SOAs, this condition must be used as the defining one for population inversion, or optical gain.

Population inversion in an SOA is achieved by forward-biasing a *pn*-junction. A *pn-junction* consists of two semiconductors: a *p*-type semiconductor that is doped with suitable impurity atoms so as to have an excess concentration of holes, and an *n*-type semiconductor that has an excess concentration of electrons. When the two semiconductors are in juxtaposition, as in Figure 3.41(a), holes diffuse from the *p*-type semiconductor to the *n*-type semiconductor, and electrons diffuse from the *n*-type semiconductor to the *p*-type semiconductor. This creates a region with net negative charge in the *p*-type semiconductor and a region with net positive

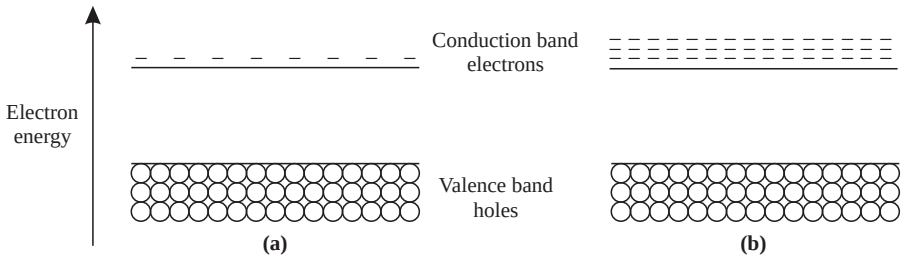

Figure 3.40 The energy bands in a *p*-type semiconductor and the electron concentration at (a) thermal equilibrium and (b) population inversion.

charge in the *n*-type semiconductor, as shown in Figure 3.41(b). These regions are devoid of free charge carriers and are together termed the *depletion region*. When no voltage (bias) is applied to the *pn*-junction, the minority carrier concentrations (electrons in the *p*-type region and holes in the *n*-type region) remain at their thermal equilibrium values. When the junction is *forward biased*—positive bias is applied to the *p*-type and negative bias to the *n*-type—as shown in Figure 3.41(c), the width of the depletion region is reduced, and there is a drift of electrons from the *n*-type region to the *p*-type region. This drift increases the electron concentration in the conduction band of the *p*-type region. Similarly, there is a drift of holes from the *p*-type to the *n*-type region that increases the hole concentration in the valence band of the *n*-type region. When the forward-bias voltage is sufficiently high, these increased minority carrier concentrations result in population inversion, and the *pn*-junction acts as an optical amplifier.

In practice, a simple *pn*-junction is not used, but a thin layer of a different semiconductor material is sandwiched between the *p*-type and *n*-type regions. Such a device is called a *heterostructure*. This semiconductor material then forms the *active region* or *layer*. The material used for the active layer has a slightly smaller bandgap and a higher refractive index than the surrounding *p*-type and *n*-type regions. The smaller bandgap helps to confine the carriers injected into the active region (electrons from the *n*-type region and holes from the *p*-type region). The larger refractive index helps to confine the light during amplification since the structure now forms a dielectric waveguide (see Section 2.3.4).

In semiconductor optical amplifiers, the population inversion condition (stimulated emission exceeds absorption) must be evaluated as a function of optical frequency or wavelength. Consider an optical frequency f_c such that $hf_c > E_g$, where E_g is the bandgap of the semiconductor material. The lowest optical frequency (or largest wavelength) that can be amplified corresponds to this bandgap. As the

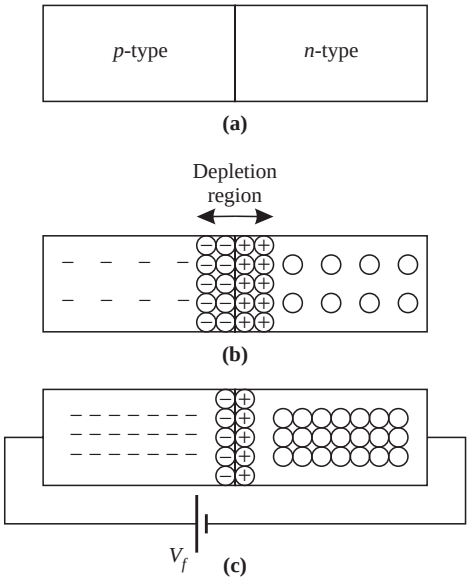

Figure 3.41 A forward-biased *pn*-junction used as an amplifier. (a) A *pn*-junction. (b) Minority carrier concentrations and depletion region with no bias voltage applied. (c) Minority carrier concentrations and depletion region with a forward-bias voltage, V_f.

forward-bias voltage is increased, the population inversion condition for this wavelength is reached first. As the forward bias voltage increases further, the electrons injected into the *p*-type region occupy progressively higher energy levels, and signals with smaller wavelengths can be amplified. In practice, bandwidths on the order of 100 nm can be achieved with SOAs. This is much larger than what is achievable with EDFAs. Signals in the 1.3 and 1.55 μm bands can even be simultaneously amplified using SOAs. Nevertheless, EDFAs are widely preferred to SOAs for several reasons. The main reason is that SOAs introduce severe crosstalk when they are used in WDM systems. This is discussed next. The gains and output powers achievable with EDFAs are higher. The coupling losses and the polarization-dependent losses are also lower with EDFAs since the amplifier is also a fiber. Due to the higher input coupling loss, SOAs have higher *noise figures* relative to EDFAs. (We will discuss noise figure in Section 4.4.5. For our purposes here, we can think of it as a measure of the noise introduced by the amplifier.) Finally, the SOA requires very high-quality antireflective coatings on its facets (reflectivity of less than 10^{-4}), which is not easy

to achieve. Higher values of reflectivity create ripples in the gain spectrum and cause gain variations due to temperature fluctuations. (Think of this device as a Fabry-Perot filter with very poor reflectivity, and the spectrum as similar to the one plotted in Figure 3.17 for the case of poor reflectivity.) Alternatively, the SOA facets can be angled to obtain the desired reflectivities, at the cost of an increased polarization dependence.

3.4.6 Crosstalk in SOAs

Consider an SOA to which is input the sum of two optical signals at different wavelengths. Assume that both wavelengths are within the bandwidth of the SOA. The presence of one signal will deplete the minority carrier concentration by the stimulated emission process so that the population inversion seen by the other signal is reduced. Thus the other signal will not be amplified to the same extent and, if the minority carrier concentrations are not very large, may even be absorbed! (Recall that if the population inversion condition is not achieved, there is net absorption of the signal.) Thus, for WDM networks, the gain seen by the signal in one channel varies with the presence or absence of signals in the other channels. This phenomenon is called *crosstalk*, and it has a detrimental effect on system performance.

This crosstalk phenomenon depends on the spontaneous emission lifetime from the high-energy to the low-energy state. If the lifetime is large enough compared to the rate of fluctuations of power in the input signals, the electrons cannot make the transition from the high-energy state to the lower-energy state in response to these fluctuations. Thus there is no crosstalk whatsoever. In the case of SOAs, this lifetime is on the order of nanoseconds. Thus the electrons can easily respond to fluctuations in power of signals modulated at gigabit/second rates, resulting in a major system impairment due to crosstalk. In contrast, the spontaneous emission lifetime in an EDFA is about 10 ms. Thus crosstalk is introduced only if the modulation rates of the input signals are less than a few kilohertz, which is not usually the case. Thus EDFAs are better suited for use in WDM systems than SOAs.

There are several ways of reducing the crosstalk introduced by SOAs. One way is to operate the amplifier in the small signal region where the gain is relatively independent of the input power of the signal. Another is to *clamp* the gain of the amplifier using a variety of techniques, so that even at high signal powers, its gain remains relatively constant, independent of the input signal. Also, if a sufficiently large number of signals at different wavelengths are present, although each signal varies in power, the total signal power into the amplifier can remain fairly constant.

The crosstalk effect is not without its uses. We will see in Section 3.8.2 that it can be used to make a *wavelength converter.*

3.5 Transmitters

We will study many different types of light sources in this section. The most important one is the laser, of which there are many different types. Lasers are used as transmitters as well as to pump both erbium-doped and Raman amplifiers.

When using a laser as a light source for WDM systems, we need to consider the following important characteristics:

1. Lasers need to produce a reasonably high output power. For WDM systems, the typical laser output powers are in the 0–10 dBm range. Related parameters are the threshold current and slope efficiency. Both of these govern the efficiency of converting electrical power into optical power. The *threshold current* is the drive current at which the laser starts to emit optical power, and the *slope efficiency* is the ratio of output optical power to drive current.
2. The laser needs to have a narrow *spectral width* at a specified operating wavelength so that the signal can pass through intermediate filters and multiple channels can be placed close together. The side-mode suppression ratio is a related parameter, which we will discuss later. In the case of a tunable laser, the operating wavelength can be varied.
3. Wavelength stability is an important criterion. When maintained at constant temperature, the wavelength drift over the life of the laser needs to be small relative to the wavelength spacing between adjacent channels.
4. For lasers that are modulated, chromatic dispersion can be an important limiting factor that affects the link length. We will see in Chapter 5 that the dispersion limit can be stated in terms of a penalty as a function of the total accumulated dispersion along the link.

Pump lasers are required to produce much higher power levels than lasers used as WDM sources. Pump lasers used in erbium-doped fiber amplifiers put out 100–200 mW of power, and pump lasers for Raman amplifiers may go up to a few watts.

3.5.1 Lasers

A laser is essentially an optical amplifier enclosed within a reflective cavity that causes it to oscillate via positive feedback. *Semiconductor lasers* use semiconductors as the gain medium, whereas *fiber lasers* typically use erbium-doped fiber as the gain medium. Semiconductor lasers are by far the most popular light sources for optical communication systems. They are compact, usually only a few hundred micrometers

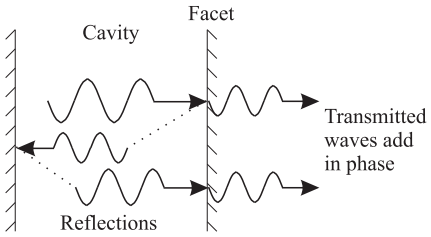

Figure 3.42 Reflection and transmission at the facets of a Fabry-Perot cavity.

in size. Since they are essentially *pn*-junctions, they can be fabricated in large volumes using highly advanced integrated semiconductor technology. The lack of any need for optical pumping, unlike fiber lasers, is another advantage. In fact, a fiber laser typically uses a semiconductor laser as a pump! Semiconductor lasers are also highly efficient in converting input electrical (pump) energy into output optical energy.

Both semiconductor and erbium fiber lasers are capable of achieving high output powers, typically between 0 and 20 dBm, although semiconductor lasers used as WDM sources typically have output powers between 0 and 10 dBm. Fiber lasers are used mostly to generate periodic trains of very short pulses (by using a technique called mode locking, discussed later in this section).

Principle of Operation

Consider any of the optical amplifiers described, and assume that a part of the optical energy is reflected at the ends of the amplifying or *gain medium*, or *cavity*, as shown in Figure 3.42. Further assume that the two ends of the cavity are plane and parallel to each other. Thus the gain medium is placed in a *Fabry-Perot cavity* (see Section 3.3.5). Such an optical amplifier is called a *Fabry-Perot amplifier*. The two end faces of the cavity (which play the role of the mirrors) are called *facets*.

The result of placing the gain medium in a Fabry-Perot cavity is that the gain is high only for the resonant wavelengths of the cavity. The argument is the same as that used in the case of the Fabry-Perot filter (Section 3.3.5). After one pass through the cavity, as shown in Figure 3.42, part of the light leaves the cavity through the right facet, and part is reflected. Part of the reflected wave is again reflected by the left facet to the right facet. For the resonant wavelengths of the cavity, all the light waves transmitted through the right facet *add in phase*. As a result of in-phase addition, the amplitude of the transmitted wave is greatly increased for these resonant wavelengths

compared to other wavelengths. Thus, when the facets are at least partially reflecting, the gain of the optical amplifier becomes a function of the wavelength.

If the combination of the amplifier gain and the facet reflectivity is sufficiently large, the amplifier will start to "oscillate," or produce light output, even in the absence of an input signal. For a given device, the point at which this happens is called its *lasing threshold*. Beyond the threshold, the device is no longer an amplifier but an oscillator or *laser*. This occurs because the stray spontaneous emission, which is always present at all wavelengths within the bandwidth of the amplifier, gets amplified even without an input signal and appears as the light output. This process is quite similar to what happens in an electronic oscillator, which can be viewed as an (electronic) amplifier with positive feedback. (In electronic oscillators, the thermal noise current due to the random motion of electrons serves the same purpose as spontaneous emission.) Since the amplification process is due to stimulated emission, the light output of a laser is *coherent*. The term *laser* is an acronym for *light amplification* by *stimulated emission* of *radiation*.

Longitudinal Modes

For laser oscillation to occur at a particular wavelength, two conditions must be satisfied. First, the wavelength must be within the bandwidth of the gain medium that is used. Thus, if a laser is made from erbium-doped fiber, the wavelength must lie in the range 1525–1560 nm. The second condition is that the length of the cavity must be an integral multiple of half the wavelength in the cavity. For a given laser, all the wavelengths that satisfy this second condition are called the *longitudinal modes* of that laser. The adjective "longitudinal" is used to distinguish these from the waveguide modes (which should strictly be called spatial modes) that we studied in Section 2.2.

The laser described earlier is called a *Fabry-Perot laser* (FP laser) and will usually oscillate simultaneously in several longitudinal modes. Such a laser is termed a *multiple-longitudinal mode* (MLM) laser. MLM lasers have large spectral widths, typically around 10 nm. A typical spectrum of the output of an MLM laser is shown in Figure 3.43(a). We saw in Section 2.4 that for high-speed optical communication systems, the spectral width of the source must be as narrow as possible to minimize the effects of chromatic dispersion. Similarly, a narrow spectral width is also needed to minimize crosstalk in WDM systems (see Section 3.3). Thus it is desirable to design a laser that oscillates in a single-longitudinal mode (SLM) only. The spectrum of the output of an SLM laser is shown in Figure 3.43(b). Single-longitudinal mode oscillation can be achieved by using a filtering mechanism in the laser that selects the desired wavelength and provides loss at the other wavelengths. An important

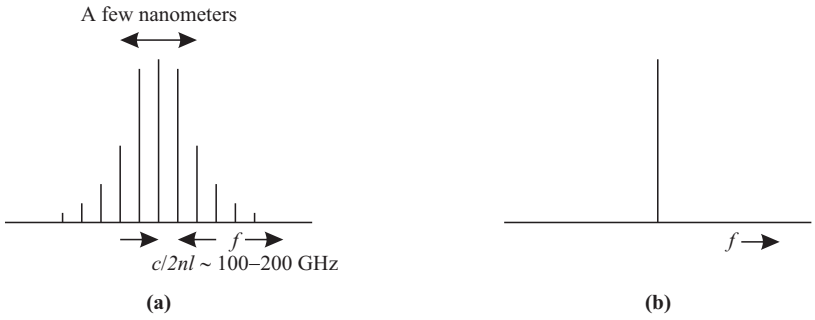

Figure 3.43 The spectrum of the output of (a) an MLM laser and (b) an SLM laser. The laser cavity length is denoted by l, and its refractive index by n. The frequency spacing between the modes of an MLM laser is then $c/2nl$.

attribute of such a laser is its *side-mode suppression ratio,* which determines the level to which the other longitudinal modes are suppressed, compared to the main mode. This ratio is typically more than 30 dB for practical SLM lasers. We will now consider some mechanisms that are commonly employed for realizing SLM lasers.

Distributed-Feedback Lasers

In the Fabry-Perot laser described earlier, the feedback of the light occurs from the reflecting facets at the ends of the cavity. Thus the feedback can be said to be *localized* at the facets. Light feedback can also be provided in a *distributed* manner by a series of closely spaced reflectors. The most common means of achieving this is to provide a periodic variation in the width of the cavity, as shown in Figure 3.44(a) and (b).

In the corrugated section of the cavity, the incident wave undergoes a series of reflections. The contributions of each of these reflected waves to the resulting transmitted wave from the cavity add in phase if the period of the corrugation is an integral multiple of half the wavelength in the cavity. The reasoning for this condition is the same as that used for the Fabry-Perot cavity. This condition is called the Bragg condition and was discussed in Section 3.3.3. The Bragg condition will be satisfied for a number of wavelengths, but the strongest transmitted wave occurs for the wavelength for which the corrugation period is *equal* to half the wavelength, rather than some other integer multiple of it. Thus this wavelength gets preferentially amplified at the expense of the other wavelengths. By suitable design of the device, this effect can be used to suppress all other longitudinal modes so that

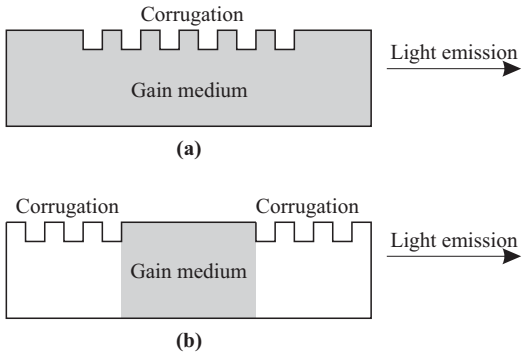

Figure 3.44 The structure of (a) a DFB laser and (b) a DBR laser. In a DFB laser, the gain and wavelength selection are obtained in the same region, whereas in a DBR laser, the wavelength selection region is outside the gain region.

the laser oscillates in a single-longitudinal mode whose wavelength is equal to twice the corrugation period. By varying the corrugation period at the time of fabrication, different operating wavelengths can be obtained.

Any laser that uses a corrugated waveguide to achieve single-longitudinal mode operation can be termed a distributed-feedback laser. However, the acronym *DFB laser* is used only when the corrugation occurs within the gain region of the cavity, as shown in Figure 3.44(a). When the corrugation is outside the gain region, as in Figure 3.44(b), the laser is called a *distributed Bragg reflector* (DBR) laser. The main advantage of DBR lasers is that the gain region is decoupled from the wavelength selection region. Thus it is possible to control both regions independently. For example, by changing the refractive index of the wavelength selection region, the laser can be tuned to a different wavelength without affecting its other operating parameters. Indeed, this is how many of the tunable lasers that we will study in Section 3.5.3 are realized.

DFB lasers are inherently more complex to fabricate than FP lasers and thus relatively more expensive. However, DFB lasers are required in almost all high-speed transmission systems today. FP lasers are used for shorter-distance data communication applications.

Reflections into a DFB laser cause its wavelength and power to fluctuate and are prevented by packaging the laser with an isolator in front of it. The laser is also usually packaged with a thermoelectric (TE) cooler and a photodetector attached to its rear facet. The TE cooler is necessary to maintain the laser at a constant operating temperature to prevent its wavelength from drifting. The temperature sensitivity of

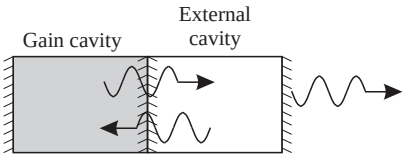

Figure 3.45 The structure of an external cavity laser.

a semiconductor DFB laser operating in the 1.55 μm wavelength region is about 0.1 nm/°C. The photodetector monitors the optical power leaking out of the rear facet, which is proportional to the optical power coming out of the laser.

The packaging of a DFB laser contributes a significant fraction of the overall cost of the device. For WDM systems, it is very useful to package multiple DFB lasers at different wavelengths inside a single package. This device can then serve as a multiwavelength light source or, alternatively, as a tunable laser (only one of the lasers in the array is turned on, depending on the desired wavelength). These lasers can all be grown on a single substrate in the form of an array. Four- and eight-wavelength laser arrays have been fabricated in research laboratories, but have not quite progressed to volume manufacturing. The primary reason for this is the relatively low yield of the array as a whole. If one of the lasers doesn't meet specifications, the entire array will have to be discarded.

External Cavity Lasers

Suppression of oscillation at more than one longitudinal mode can also be achieved by using another cavity—called an *external cavity*—following the primary cavity where gain occurs. This is illustrated in Figure 3.45. Just as the primary cavity has resonant wavelengths, so does the external cavity. This effect can be achieved, for example, by using reflecting facets for the external cavity as well. The net result of having an external cavity is that the laser is capable of oscillating only at those wavelengths that are resonant wavelengths of *both* the primary and external cavity. By suitable design of the two cavities, it can be ensured that only one wavelength in the gain bandwidth of the primary cavity satisfies this condition. Thus the laser oscillation can be confined to a single-longitudinal mode.

Instead of another Fabry-Perot cavity, as shown in Figure 3.45, we can use a diffraction grating (see Section 3.3.1) in the external cavity, as shown in Figure 3.46. Such a laser is called a *grating external cavity* laser. In this case, the facet of the gain cavity facing the grating is given an antireflection coating. The wavelengths reflected by the diffraction grating back to the gain cavity are determined by the

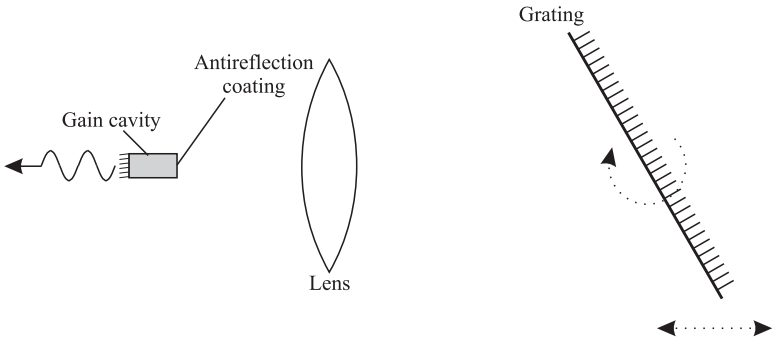

Figure 3.46 The structure of a grating external cavity laser. By rotating the grating, we can tune the wavelength of the laser.

pitch of the grating (see Section 3.3.1) and its tilt angle (see Figure 3.46) with respect to the gain cavity. An external cavity laser, in general, uses a *wavelength-selective mirror* instead of a wavelength-flat mirror. (A highly polished and/or metal-coated facet used in conventional lasers acts as a wavelength-flat mirror.) The reflectivity of a wavelength-selective mirror is a function of the wavelength. Thus only certain wavelengths experience high reflectivities and are capable of lasing. If the wavelength-selective mirror is chosen suitably, only one such wavelength will occur within the gain bandwidth, and we will have a single-mode laser.

Several of the filters discussed in Section 3.3 can be used as wavelength-selective mirrors in external cavity lasers. We have already seen the use of the diffraction grating (Section 3.3.1) and Fabry-Perot filter (Section 3.3.5) in external cavity lasers. These laser structures are used today primarily in optical test instruments and are not amenable to low-cost volume production as SLM light sources for transmission systems. One version of the external cavity laser, though, appears to be particularly promising for this purpose. This device uses a fiber Bragg grating in front of a conventional FP laser with its front facet AR coated. This device then acts as an SLM DBR laser. It can be fabricated at relatively low cost compared to DFB lasers and is inherently more temperature stable in wavelength due to the low temperature-coefficient of the fiber grating.

One disadvantage of external cavity lasers is that they cannot be modulated directly at high speeds. This is related to the fact that the cavity length is large.

Vertical Cavity Surface-Emitting Lasers

In this section, we will study another class of lasers that achieve single-longitudinal mode operation in a slightly different manner. As we saw in Figure 3.43, the frequency

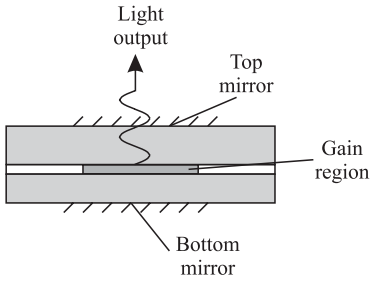

Figure 3.47 The structure of a VCSEL.

spacing between the modes of an MLM laser is $c/2nl$, where l is the length of the cavity and n is its refractive index. If we were to make the length of the cavity sufficiently small, the mode spacing increases such that only one longitudinal mode occurs within the gain bandwidth of the laser. It turns out that making a very thin active layer is much easier if the active layer is deposited on a semiconductor substrate, as illustrated in Figure 3.47. This leads to a vertical cavity with the mirrors being formed on the top and bottom surfaces of the semiconductor wafer. The laser output is also taken from one of these (usually top) surfaces. For these reasons, such lasers are called *vertical cavity surface-emitting lasers* (VCSELs). The other lasers that we have been discussing hitherto can thus be referred to as *edge-emitting lasers*.

Since the gain region has a very short length, very high mirror reflectivities are required in order for laser oscillation to occur. Such high mirror reflectivities are difficult to obtain with metallic surfaces. A stack of alternating low- and high-index dielectrics serves as a highly reflective, though wavelength-selective, mirror. The reflectivity of such a mirror is discussed in Problem 3.13. Such dielectric mirrors can be deposited at the time of fabrication of the laser.

One problem with VCSELs is the large ohmic resistance encountered by the injected current. This leads to considerable heating of the device and the need for efficient thermal cooling. Many of the dielectric materials used to make the mirrors have low thermal conductivity. So the use of such dielectric mirrors makes room temperature operation of VCSELs difficult to achieve since the heat generated by the device cannot be dissipated easily. For this reason, for several years after they were first demonstrated in 1979, VCSELs were not capable of operating at room temperature. However, significant research effort has been expended on new materials and techniques, VCSELs operating at 1.3 μm at room temperature have been demonstrated [Har00].

The advantages of VCSELs, compared to edge-emitting lasers, include simpler and more efficient fiber coupling, easier packaging and testing, and their ability

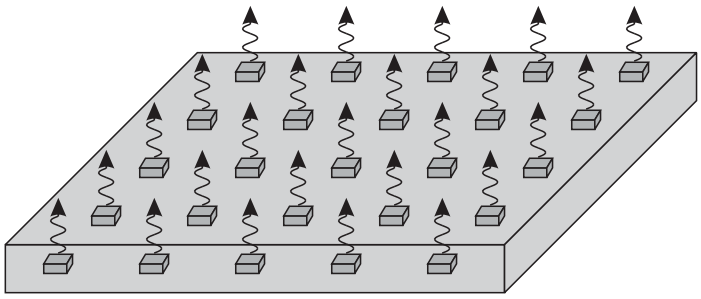

Figure 3.48 A two-dimensional array of vertical cavity surface-emitting lasers.

to be integrated into multiwavelength arrays. VCSELs operating at 0.85 μm are commercially available and used for low-cost, short-distance multimode fiber interconnections. In addition, 1.3 μm VCSELs have been commercially available.

In a WDM system, many wavelengths are transmitted simultaneously over each link. Usually, this requires a separate laser for each wavelength. The cost of the transmitters can be significantly reduced if all the lasers can be integrated on a single substrate. This is the main motivation for the development of arrayed lasers such as the DFB laser arrays that we discussed earlier. Moreover, an arrayed laser can be used as a tunable laser simply by turning on only the one required laser in the array. The use of surface-emitting lasers enables us to fabricate a two-dimensional array of lasers, as shown in Figure 3.48. Much higher array packing densities can be achieved using surface-emitting lasers than edge-emitting ones because of this added dimension. However, it is harder to couple light from the lasers in this array onto optical fiber since multiplexers that work conveniently with this two-dimensional geometry are not readily available. These arrayed lasers have the same yield problem as other arrayed laser structures; if one of the lasers does not meet specifications, the entire array will have to be discarded.

Mode-Locked Lasers

Mode-locked lasers are used to generate narrow optical pulses that are needed for the high-speed TDM systems that we will study in Chapter 12. Consider a Fabry-Perot laser that oscillates in N longitudinal modes, which are adjacent to each other. This means that if the wavelengths of the modes are $\lambda_0, \lambda_1, \ldots, \lambda_{N-1}$, the cavity length l satisfies $l = (k+i)\lambda_i/2$, $i = 0, 1, \ldots, N-1$, for some integer k. From this condition, it can be shown (see Problem 3.7) that the corresponding frequencies $f_0, f_1, \ldots, f_{N-1}$ of these modes must satisfy $f_i = f_0 + i\Delta f$, $i = 0, 1, \ldots, N-1$. The oscillation at

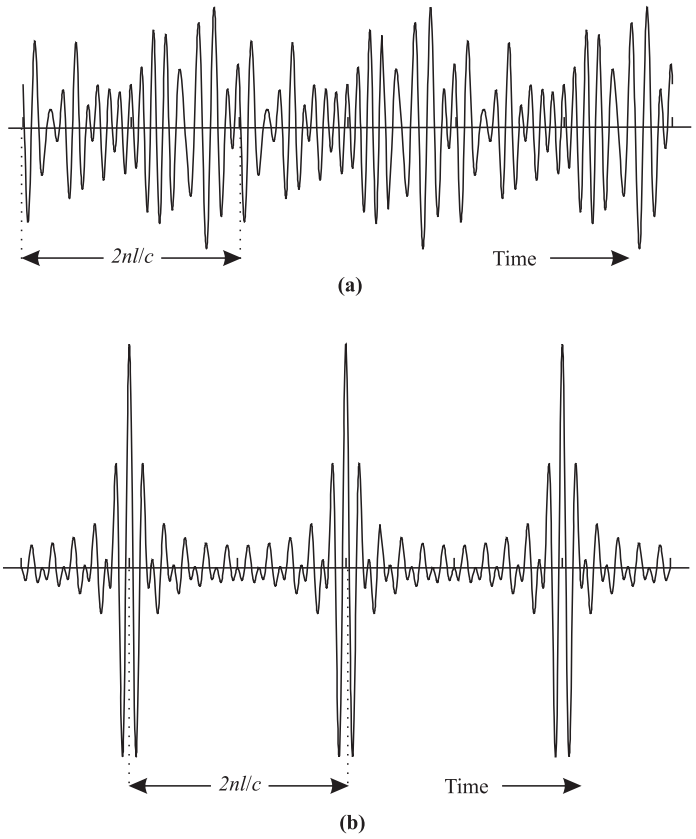

Figure 3.49 Output oscillation of a laser oscillating simultaneously in 10 longitudinal modes. (a) The phases of the modes are chosen at random. (b) All the phases are equal to each other; such a laser is said to be mode locked.

frequency f_i is of the form $a_i \cos(2\pi f_i t + \phi_i)$, where a_i is the amplitude and ϕ_i the phase of mode i. (Strictly speaking, this is the distribution in time of the electric field associated with the longitudinal mode.) Thus the total laser output oscillation takes the form

$$\sum_{i=0}^{N-1} a_i \cos(2\pi f_i t + \phi_i).$$

This expression is plotted in Figure 3.49 for $N = 10$, for different sets of values of the ϕ_i. In Figure 3.49(a), the ϕ_i are chosen at random, and in Figure 3.49(b), they

are chosen to be equal to each other. All the a_i are chosen to be equal in both cases, and the frequency f_0 has been diminished from its typical value for the purpose of illustration.

From Figure 3.49(a), we observe that the output amplitude of an MLM laser varies rapidly with time when it is not mode locked. We have also seen in Figure 3.43(a) that the frequency spacing between adjacent longitudinal modes is $c/2nl$. If $n = 3$ and $l = 200\ \mu$m, which are typical values for semiconductor lasers, this frequency spacing is 250 GHz. Thus these amplitude fluctuations occur extremely rapidly (at a time scale on the order of a few picoseconds) and pose no problems for on-off modulation even at bit rates of a few tens of gigabits per second.

We see from Figure 3.49(b) that when the ϕ_i are chosen to be equal to each other, the output oscillation of the laser takes the form of a periodic train of narrow pulses. A laser operating in this manner is called a *mode-locked laser* and is the most common means of generating narrow optical pulses.

The time interval between two pulses of a mode-locked laser is $2nl/c$, as indicated in Figure 3.49(b). For a typical semiconductor laser, as we have seen earlier, this corresponds to a few picoseconds. For modulation in the 1–10 GHz range, the interpulse interval should be in the 0.1–1 ns range. Cavity lengths, l, of the order of 1–10 cm (assuming $n = 1.5$) are required in order to realize mode-locked lasers with interpulse intervals in this range. These large cavity lengths are easily obtained using fiber lasers, which require the length anyway to obtain sufficient gain to induce lasing.

The most common means of achieving mode lock is by modulating the gain of the laser cavity. Either amplitude or frequency modulation can be used. Mode locking using amplitude modulation is illustrated in Figure 3.50. The gain of the cavity is modulated with a period equal to the interpulse interval, namely, $2nl/c$. The amplitude of this modulation is chosen such that the average gain is insufficient for any single mode to oscillate. However, if a large number of modes are in phase, there can be a sufficient buildup in the energy inside the cavity for laser oscillation to occur at the instants of high gain, as illustrated in Figure 3.50.

Gain modulation of the fiber laser can be achieved by introducing an external modulator inside the cavity.

3.5.2 *Light-Emitting Diodes*

Lasers are expensive devices and are not affordable for many applications where the data rates are low and distances are short. This is the case in many data communications applications (see Chapter 6) and in some access networks (Chapter 11). In such cases, *light-emitting diodes* (LEDs) provide a cheaper alternative.

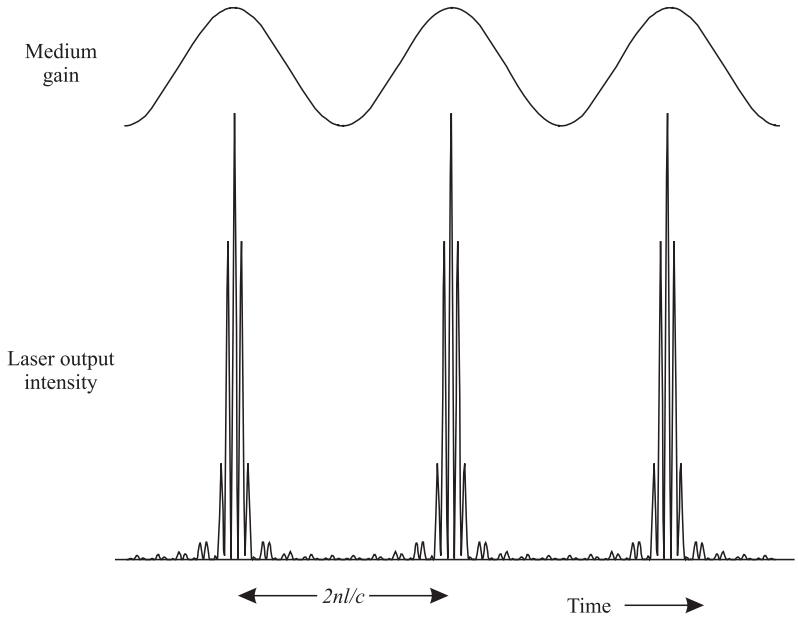

Figure 3.50 Illustration of mode locking by amplitude modulation of the cavity gain.

An LED is a forward-biased *pn*-junction in which the recombination of the injected minority carriers (electrons in the *p*-type region and holes in the *n*-type region) by the spontaneous emission process produces light. (Unwanted nonradiative recombination is also possible and is an important factor affecting the performance of LEDs.) Because spontaneous emission occurs within the entire bandwidth of the gain medium (corresponding to all energy differences between the valence and conduction bands for an LED), the light output of an LED has a broad spectrum, unlike that of a laser. We can crudely think of an LED as a laser with facets that are not very reflective. Increasing the pump current simply increases the spontaneous emission, and there is no chance to build up stimulated emission due to the poor reflectivity of the facets. For this reason, LEDs are also not capable of producing high-output powers like lasers, and typical output powers are on the order of −20 dBm. They cannot be directly modulated (see Section 3.5.4) at data rates higher than a few hundred megabits per second.

In some low-speed, low-budget applications, there is a requirement for a source with a narrow spectral width. DFB lasers provide narrow spectral widths but may be too expensive for these applications. In such cases, *LED slicing* provides a cheaper

alternative. An LED slice is the output of a narrow passband optical filter placed in front of the LED. The optical filter selects a portion of the LED's output. Different filters can be used to select (almost) nonoverlapping spectral slices of the LED output. Thus one LED can be shared by a number of users. We will see an application for this technique in Chapter 11.

3.5.3 *Tunable Lasers*

Tunable lasers are highly desirable components for WDM networks for several reasons. Fixed-wavelength DFB lasers work very well for today's applications. However, each wavelength requires a different, unique laser. This implies that in order to supply a 100-channel WDM system, we need to stock 100 different laser types. The inventory and sparing issues associated with this are expensive and affect everybody from laser manufacturers to network operators. Laser manufacturers need to set up multiple production and test lines for each laser wavelength (or time-share the same production and test line but change the settings each time a different laser is made). Equipment suppliers need to stock these different lasers and keep inventories and spares for each wavelength. Finally, network operators need to stockpile spare wavelengths in the event transmitters fail in the field and need to be replaced. Having a tunable laser alleviates this problem dramatically.

Tunable lasers are also one of the key enablers of reconfigurable optical networks. They provide the flexibility to choose the transmit wavelength at the source of a lightpath. For instance, if we wanted to have a total of, say, four lightpaths starting at a node, we would equip that node with four tunable lasers. This would allow us to choose the four transmit wavelengths in an arbitrary manner. In contrast, if we were to use fixed-wavelength lasers, either we would have to preequip the node with a large number of lasers to cover all the possible wavelengths, or we would have to manually equip the appropriate wavelength as needed. We will see more of this application in Chapter 7. The tuning time required for such applications is on the order of milliseconds because the wavelength selection happens only at the times where the lightpath is set up, or when it needs to be rerouted in the event of a failure.

Another application for tunable lasers is in optical packet-switched networks, where data needs to be transmitted on different wavelengths on a packet-by-packet basis. These networks are primarily in their early stages of research today, but supporting such an application would require tuning times on the order of nanoseconds to microseconds, depending on the bit rate and packet size used.

Finally, tunable lasers are a staple in most WDM laboratories and test environments, where they are widely used for characterizing and testing various types of

optical equipment. These lasers are typically tabletop-type devices and are not suitable for use in telecom applications, which call for compact, low-cost semiconductor lasers.

The InGaAsP/InP material used for most long-wavelength lasers is enhanced by the use of *quantum well* structures and has an overall gain bandwidth of about 250 nm at 1.55 μm, large enough for the needs of current WDM systems. However, the tuning mechanisms available potentially limit the tuning range to a small fraction of this number. The following tuning mechanisms are typically used:

- Injecting current into a semiconductor laser causes a change in the refractive index of the material, which in turn changes the lasing wavelength. This effect is fairly small—about a 0.5–2% change in the refractive index (and the wavelength) is possible. This effect can be used to effect a tuning range of approximately 10–15 nm in the 1.55 μm wavelength window.
- Temperature tuning is another possibility. The wavelength sensitivity of a semiconductor laser to temperature is approximately 0.1 nm/°C. In practice, the allowed range for temperature tuning is about 1 nm, corresponding to a 10°C temperature variation. Operating the laser at significantly higher temperatures than room temperature causes it to age rapidly, degrading its lifetime.
- Mechanical tuning can be used to provide a wide tunable range in lasers that use a separate external cavity mechanism. Many of these lasers tend to be bulky. We will look at one laser structure of this type using a micro-electro-mechanical tuning mechanism, which is quite compact.

As we will see, the tuning mechanisms are complex and, in many cases, interact with the modulation mechanisms, making it difficult to directly modulate most of the tunable lasers that we will study here.

The ideal tunable laser is a device that can tune rapidly over a wide continuous tuning range of over 100 nm. It should be stable over its lifetime and easily controllable and manufacturable. Many of the tunable laser technologies described here have been around for many years, but we are only now beginning to see commercially available devices due to the complexity of manufacturing and controlling these devices and solving the reliability challenges. The strong market demand for these devices has stimulated a renewed effort to solve these problems.

External Cavity Lasers

External cavity lasers can be tuned if the center wavelength of the grating or other wavelength-selective mirror used can be changed. Consider the grating external cavity laser shown in Figure 3.46. The wavelength selected by the grating for reflection

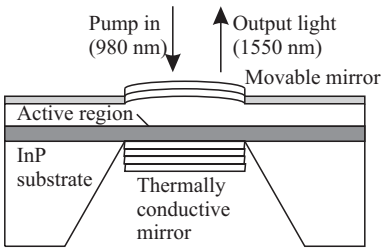

Figure 3.51 Structure of a tunable micro-electro-mechanical vertical cavity surface-emitting laser (MEM-VCSEL) (from [Vak99]).

to the gain cavity is determined by the pitch of the diffraction grating, its tilt angle with respect to the gain cavity, and its distance from the gain cavity (see Section 3.3.1, specifically, (3.9)). Thus by varying the tilt angle and the distance of the diffraction grating from the gain cavity (shown by the dotted arrows in Figure 3.46), the laser wavelength can be changed. This is a slow method of tuning since the tilt and position of the diffraction grating have to be changed by mechanical means. However, a very wide tuning range of about 100 nm can be obtained for semiconductor lasers by this method. This method of tuning is appropriate for test instruments but not for a compact light source for communication systems.

Tunable VCSELs

We studied VCSELs in Section 3.5.1. There we saw that the main challenges in realizing long-wavelength 1.55 μm VCSELs were in obtaining sufficient cavity gain, obtaining highly reflective mirror surfaces, dealing with the heat dissipation, and making the laser operate in a single-longitudinal mode. Figure 3.51 shows a VCSEL design [Vak99] that attempts to solve these problems, while also making the laser itself tunable. The tunability is achieved by having the upper mirror be a movable micro-electro-mechanical (MEM) membrane. The cavity spacing can be adjusted by moving the upper mirror by applying a voltage across the upper and lower mirrors. The upper mirror is curved to prevent beam walk-off in the cavity, leading to better stability of the lasing mode.

To conduct the heat away from the bottom mirror, a hole is etched in the InP substrate. The design uses a 980 nm pump laser to pump the VCSEL cavity. Any pump wavelength lower than the desired lasing wavelength can be used to excite the semiconductor electrons to the conduction band. For example, the 980 nm semiconductor pumps used to pump erbium-doped fiber amplifiers can be used here as well. By designing the pump spot size to match the size of the fundamental lasing mode,

the laser can be made single mode while suppressing the higher-order Fabry-Perot cavity modes. Using gain to perform this function is better than trying to design the cavity to provide higher loss at the higher-order modes. The high gain also allows the output coupling reflectivity to be reduced, while still maintaining sufficient inversion inside the cavity to prevent excessive recombination.

The laser described in [Vak99] was able to put out about 0 dBm of power in continuous-wave (CW) mode over a tuning range of 50 nm.

Two- and Three-Section DBR Lasers

We saw earlier that we can change the refractive index of a semiconductor laser by injecting current into it. This can result in an overall tuning range of about 10 nm. The DFB laser shown in Figure 3.44 can be tuned by varying the forward-bias current, which changes the refractive index, which in turn changes the effective pitch of the grating inside the laser cavity. However, changing the forward-bias current also changes the output power of the device, making this technique unsuitable for use in a DFB laser.

A conventional DBR laser also has a single gain region, which is controlled by injecting a forward-bias current I_g, as shown in Figure 3.44(b). Varying this current only changes the output power and does not affect the wavelength. This structure can be modified by adding another electrode to inject a separate current I_b into the Bragg region that is decoupled from the gain region, as shown in Figure 3.52(a). This allows the wavelength to be controlled independently of the output power.

As in a conventional DBR laser, the laser has multiple closely spaced cavity modes corresponding to the cavity length, of which the one that lases corresponds to the wavelength peak of the Bragg grating. As the wavelength peak of the grating is varied by varying I_b, the laser hops from one cavity mode to another. This effect is shown in Figure 3.52(a). As the current I_b is varied, the Bragg wavelength changes. At the same time, there is also a small change in the cavity mode spacing due to the change in refractive index in the grating portion of the overall cavity. The two changes do not track each other, however. As a result, as I_b is varied and the Bragg wavelength changes, the laser wavelength changes, with the laser remaining on the same cavity mode for some time. As the current is varied further, the laser hops to the next cavity mode. By careful control over the cavity length, we can make the wavelength spacing between the cavity modes equal to the WDM channel spacing.

In order to obtain continuous tuning over the entire wavelength range, an additional third *phase* section can be added to the DBR, as shown in Figure 3.52(b). Injecting a third current I_p into this section allows us to obtain control of the cavity mode spacing, independent of the other effects that are present in the laser. Recall from Section 3.3.5 that it is sufficient to vary the effective cavity length by half a

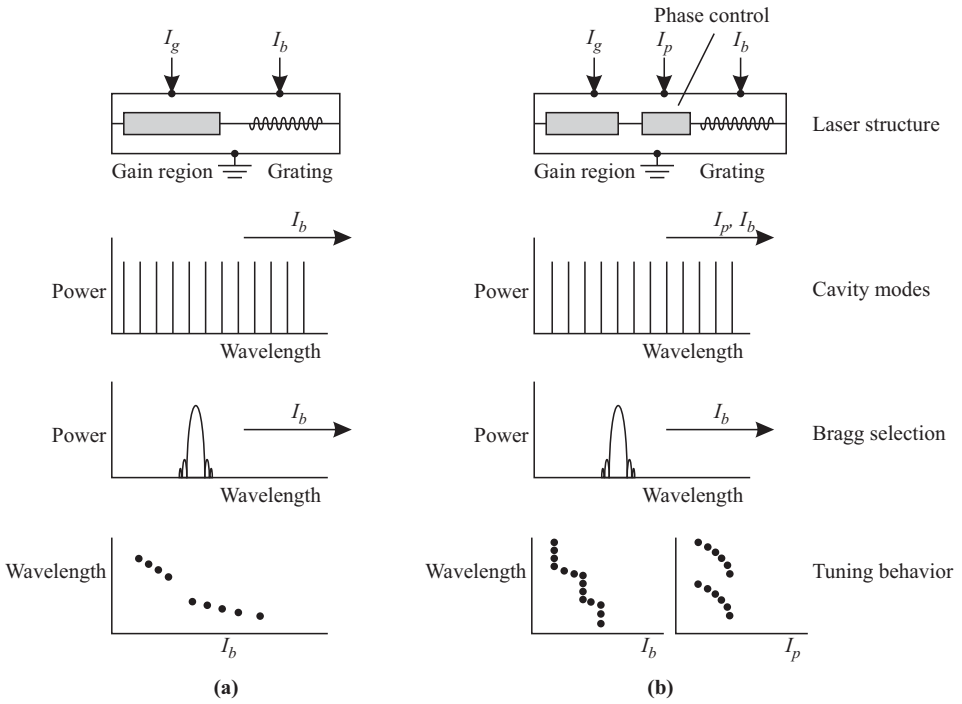

Figure 3.52 Two- and three-section DBR lasers and their principle of wavelength selection. (a) Two-section DBR showing separate control of the gain and Bragg sections. (c) Three-section DBR, which adds an additional control for the cavity phase.

wavelength (or equivalently, the phase by π) in order to obtain tuning across an entire free spectral range. This is a small fraction of the overall cavity length and is easily achieved by current injection into the phase section. By carefully controlling I_p to line up a cavity mode to correspond to the wavelength peak of the Bragg grating determined by I_b, the wavelength can be tuned continuously over the tunable range.

Two- and three-section DBRs capable of tuning over 32 channels in 50 GHz increments were demonstrated several years ago [KK90, Kam96] and are nearing commercial availability.

Clearly, a major problem that needs to be solved is in the control of these lasers, which can be quite complicated. As the laser ages, or temperature changes, the control currents may need to be recalibrated; otherwise the laser could end up hopping to another wavelength. The hopping could happen back and forth rapidly, and could

manifest itself as relative intensity noise (RIN) at the laser output. In a sense, we are eliminating the very fact that made DFB lasers so wavelength stable—a fixed grating. These problems are only compounded further in the more complex laser structures that we will discuss next.

The DBRs that we have looked at so far are all limited to about a 10–15 nm tuning range by the 0.5–2% change in refractive index possible. Increasing the tuning range beyond this value requires a new bag of tricks. One trick makes the laser wavelength dependent on the *difference* between the refractive indices of two different regions. The overall variation possible is much higher than the variation of each of the individual regions. The so-called vertical grating-assisted coupler filter (VGF) lasers [AKB+92, AI93] make use of this principle. The second trick is to make use of the Vernier effect, where we have two combs of wavelengths, each with slightly different wavelength spacing. The combination of the two combs yields another periodic comb with a much higher wavelength spacing between its peaks. Problem 3.28 explains this effect in more detail. Even if each comb can be tuned only to a small extent, the combination of the two combs yields a much higher tuning range. The *sampled grating* (SG) DBRs and the *super-structure grating* (SSG) DBRs [JCC93, Toh93] use this approach. Finally, the *grating-coupled sampled reflector* (GCSR) laser [WMB92, Rig95] is a combination of both approaches.

VGF Lasers

Figure 3.53 shows the schematic of a VGF laser. It consists of two waveguides, with a coupling region between them. Its operation is similar to that of the acousto-optic tunable filter of Section 3.3.9. Using (3.17), wavelength λ is coupled from one waveguide of refractive index n_1 to the other of refractive index n_2 if

$$\lambda = \Lambda_B(n_1 - n_2)$$

where Λ_B is the period of the Bragg grating. Changing the refractive index of one region, say, n_1 by Δn_1, therefore results in a wavelength tuning of $\Delta\lambda$ where

$$\frac{\Delta\lambda}{\lambda} \approx \frac{\Delta n_1}{n_1 - n_2}.$$

This is significantly larger than the $\Delta n_1/n_1$ ratio that is achievable in the two- and three-section DBRs that we studied earlier.

In Figure 3.53, current I_c controls the index n_1, and current I_g provides the current to the gain region in the other waveguide. Just as with the two- and three-section DBRs, in order to obtain continuous tuning, the cavity mode spacing needs to be controlled by a third current I_p. Lasers with tuning ranges over 70 nm have been demonstrated using this approach.

One major problem with this approach is that the cavity length needs to be fairly long (typically 800–1000 μm) to get good coupling between the waveguides. This

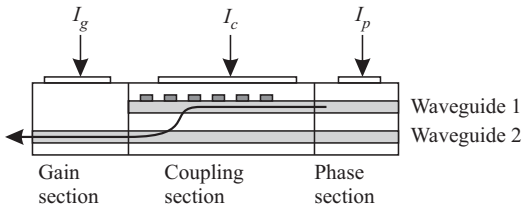

Figure 3.53 A vertical grating-assisted coupler filter tunable laser.

causes the cavity modes to be spaced very closely together. The laser therefore tends to hop fairly easily from one cavity mode to another, even though all the control currents are held steady. This effectively results in a poor side-mode suppression, making the laser not as suitable for high-bit-rate long-distance transmission.

Sampled Grating and Super-Structure Grating DBR Lasers

A sampled grating DBR laser is shown in Figure 3.54. It has two gratings, one in the front and one in the back. The Bragg grating in front is interrupted periodically (or *sampled*) with a period Λ_1. This results in a periodic set of Bragg reflector peaks, spaced apart in wavelength by $\lambda^2/2n_{\text{eff}}\Lambda_1$, as shown in Figure 3.54, where λ is the nominal center wavelength. The peaks gradually taper off in reflectivity, with the highest reflection occurring at the Bragg wavelength $2n_{\text{eff}}\Lambda$, where Λ is the period of the grating. The grating in the back is sampled with a different period Λ_2, which results in another set of reflection peaks spaced apart in wavelength by $\lambda^2/2n_{\text{eff}}\Lambda_2$. In order for lasing to occur, we need to have an overlap between the two reflection peaks of the Bragg gratings and a cavity mode. Even though the tuning range of each reflection peak is limited to 10–15 nm, combining the two sets of reflection peaks results in a large tuning range. Just as with the two- and three-section DBR lasers, a separate phase section controls the cavity mode spacing to ensure continuous tuning. An additional complication with this approach is that because the reflection peaks taper off, the current in the gain region needs to be increased to compensate for the poorer reflectivity as the laser is tuned away from the primary Bragg reflection peak.

Another way of getting the same effect is to use periodically chirped gratings instead of the gratings shown in Figure 3.54. This structure is called a super-structure grating DBR laser. The advantage of this structure is that the chirped gratings provide a highly reflective set of peaks over a wider wavelength range than the sampled grating structure.

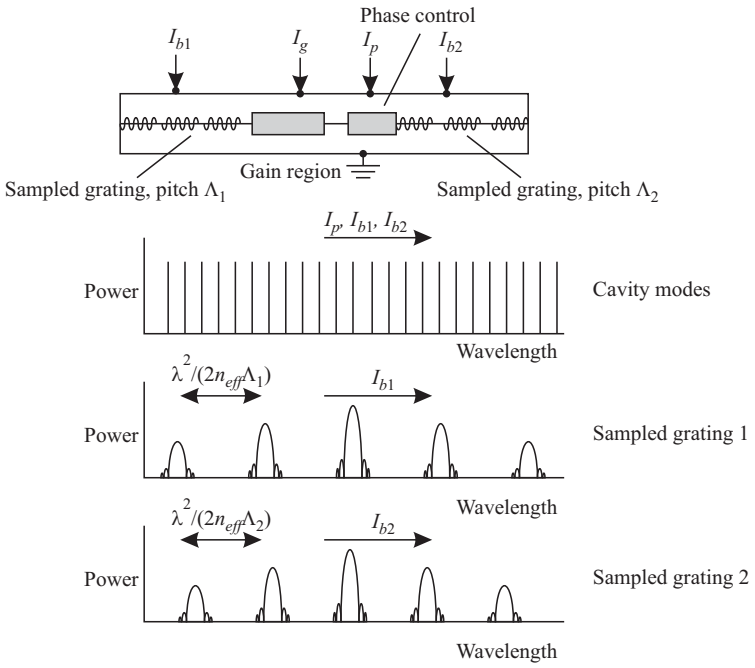

Figure 3.54 A sampled grating DBR laser and its principle of wavelength selection.

Grating-Coupled Sampled Reflector Laser

The GCSR laser is a combination of a VGF and a sampled or super-structure grating, as shown in Figure 3.55. The VGF provides a wide tuning range, and the SSG grating provides high selectivity to eliminate side modes. In a sense, the VGF provides coarse tuning to select a wavelength band with multiple cavity modes in the band, and the SSG grating provides the wavelength selection within the band. Just as in the two- and three-section DBR lasers, an additional phase section provides the fine control over the cavity modes to provide continuous tuning within the band to suppress side modes.

Laser Arrays

Another way to obtain a tunable laser source is to use an array of wavelength-differentiated lasers and turn one of them on at any time. Arrays could also be used to replace individual light sources.

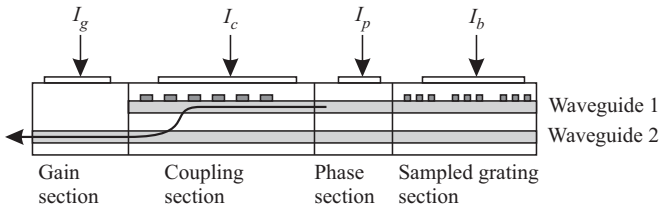

Figure 3.55 A grating coupled sampled reflector laser.

One approach is to fabricate an array of DFB lasers, each of them at a different wavelength. Combined with temperature tuning, we can use this method to obtain fairly continous tuning. A major problem with this approach is in the wavelength accuracy of the individual lasers in the array, making it difficult to obtain a comb of accurately spaced wavelengths out of the array. However, if only one laser is to be used at any given time, we can use temperature tuning to make up for this inaccuracy. Lasers using this approach have been demonstrated and used in system experiments [Zah92, You95].

Another approach is to use Fabry-Perot–type laser arrays and use an external mechanism for selecting the lasing wavelength. Several structures have been proposed [Soo92, ZJ94], one using an external waveguide grating and the other using an external arrayed waveguide grating. With these structures, the wavelength accuracy is determined by the external grating. The long cavity length results in potentially a large number of cavity modes within the grating wavelength selection window, which could cause the laser to hop between cavity modes during operation.

3.5.4 Direct and External Modulation

The process of imposing data on the light stream is called *modulation*. The simplest and most widely used modulation scheme is called *on-off keying* (OOK), where the light stream is turned on or off, depending on whether the data bit is a 1 or 0. We will study this in more detail in Chapter 4.

OOK modulated signals are usually realized in one of two ways: (1) by *direct modulation* of a semiconductor laser or an LED, or (2) by using an *external modulator*. The direct modulation scheme is illustrated in Figure 3.56. The drive current into the semiconductor laser is set well above threshold for a 1 bit and below (or slightly above) threshold for a 0 bit. The ratio of the output powers for the 1 and 0 bits is called the *extinction ratio*. Direct modulation is simple and inexpensive since no other components are required for modulation other than the light source

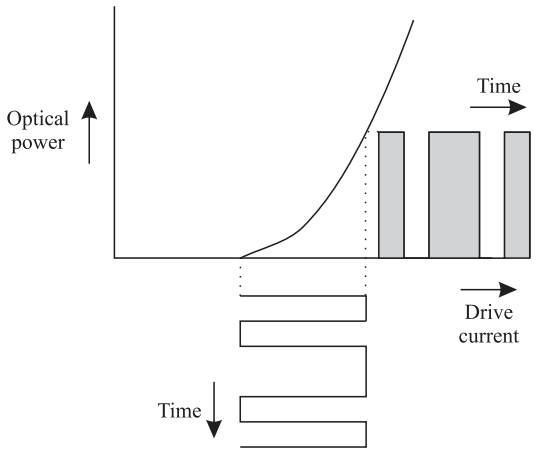

Figure 3.56 Direct modulation of a semiconductor laser.

(laser/LED) itself. In fact, a major advantage of semiconductor lasers is that they can be directly modulated. In contrast, many other lasers are continuous wave sources and cannot be modulated directly at all. These lasers require an external modulator. For example, because of the long lifetime of the erbium atoms at the E_2 level in Figure 3.35, erbium lasers cannot be directly modulated even at speeds of a few kilobits per second.

The disadvantage of direct modulation is that the resulting pulses are considerably *chirped*. Chirp is a phenomenon wherein the carrier frequency of the transmitted pulse varies with time, and it causes a broadening of the transmitted spectrum. As we saw in Section 2.4, chirped pulses have much poorer dispersion limits than unchirped pulses. The amount of chirping can be reduced by increasing the power of a 0 bit so that the laser is always kept well above its threshold; the disadvantage is that this reduces the extinction ratio, which in turn, degrades the system performance, as we will see in Section 5.3. In practice, we can realize an extinction ratio of around 7 dB while maintaining reasonable chirp performance. This enhanced pulse broadening of chirped pulses is significant enough to warrant the use of *external modulators* in high-speed, dispersion-limited communication systems.

An OOK external modulator is placed in front of a light source and turns the light signal on or off based on the data to be transmitted. The light source itself is continuously operated. This has the advantage of minimizing undesirable effects, particularly chirp. Several types of external modulators are commercially available and are increasingly being integrated with the laser itself inside a single package

to reduce the packaging cost. In fact, transmitter packages that include a laser, external modulator, and wavelength stabilization circuits are becoming commercially available for use in WDM systems.

External modulators become essential in transmitters for communication systems using solitons or return-to-zero (RZ) modulation (see Section 2.6). As shown in Figure 3.57(a), to obtain a modulated train of RZ pulses, we can use a laser generating a train of periodic pulses, such as a mode-locked laser (see Section 3.5.1) followed by an external modulator. The modulator blocks the pulses corresponding to a 0 bit. (Usually we cannot directly modulate a pulsed laser emitting periodic pulses.) Unfortunately, cost-effective and compact solid-state lasers for generating periodic pulses are not yet commercially available. More commonly, as shown in Figure 3.57(b), practical RZ systems today use a continuous-wave DFB laser followed by a two-stage external modulator. The first stage creates a periodic train of short (RZ) pulses, and the second stage imposes the modulation by blocking out the 0 bits. Dispersion-managed soliton systems (see Section 2.6.1) require the generation of RZ pulses with a carefully controlled amount and sign of chirp. This can be accomplished by using another phase modulation stage.

Two types of external modulators are widely used today: lithium niobate modulators and semiconductor electro-absorption (EA) modulators. The lithium niobate modulator makes use of the electro-optic effect, where an applied voltage induces a change in refractive index of the material. The device itself is configured either as a directional coupler or as a Mach-Zehnder interferometer (MZI). Figure 3.58 shows the directional coupler configuration. Applying a voltage to the coupling region changes its refractive index, which in turn determines how much power is coupled from the input waveguide 1 to the output waveguide 1 in the figure.

Figure 3.59 shows the MZI configuration, which operates on the principles that we studied in Section 3.3.7. Compared to a directional coupler, the MZI offers a higher modulation speed for a given drive voltage and provides a higher extinction ratio. For these reasons, it is the more popular configuration. In one state, the signals in the two arms of the MZI are in phase and interfere constructively and appear at the output. In the other state, applying a voltage causes a π phase shift between the two arms of the MZI, leading to destructive interference and no output signal. These modulators have very good extinction ratios ranging from 15 to 20 dB, and we can control the chirp very precisely. Due to the high polarization dependence of the device, a polarization maintaining fiber is used between the laser and the modulator.

The EA modulator is an attractive alternative to lithium niobate modulators because it can be fabricated using the same material and techniques used to fabricate semiconductor lasers. This allows an EA modulator to be integrated along with a DFB laser in the same package and results in a very compact, lower-cost solution, compared to using an external lithium niobate modulator. In simple terms, the EA

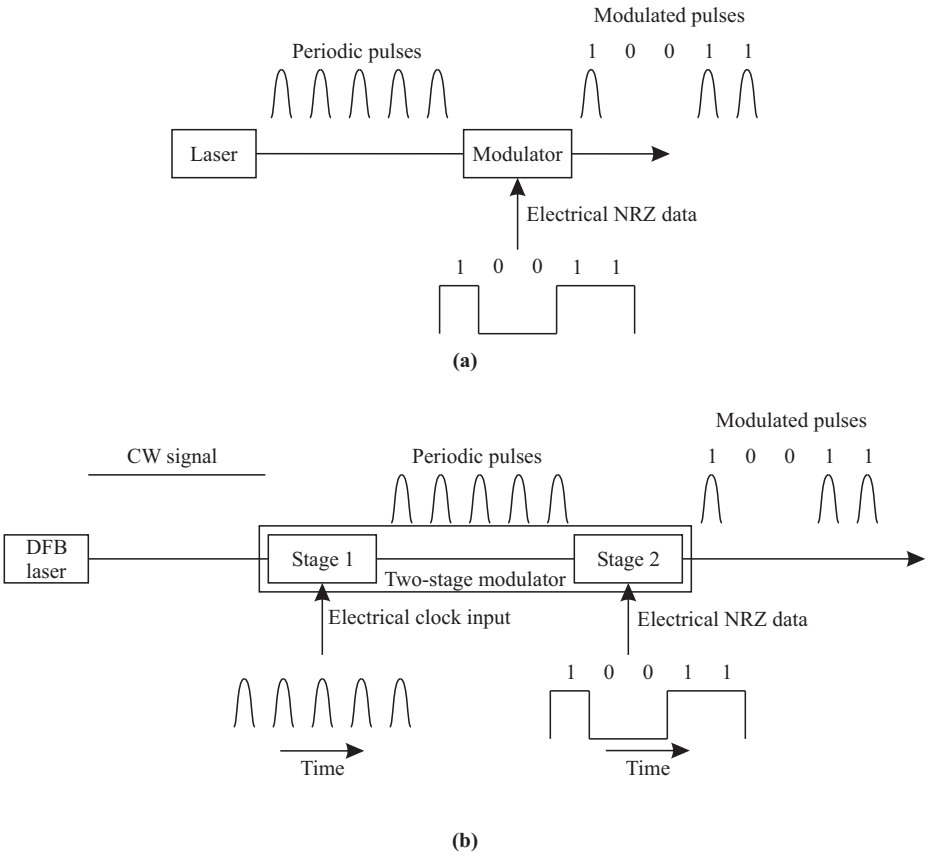

Figure 3.57 Using external modulators to realize transmitters for systems using RZ or soliton pulses. (a) A laser emitting a periodic pulse train, with the external modulator used to block the 0 bits and pass through the 1 bits. (b) A more common approach using a continuous-wave (CW) DFB laser followed by a two-stage modulator.

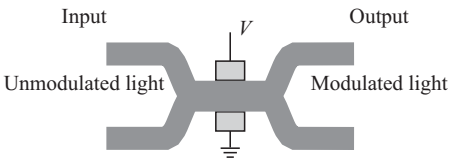

Figure 3.58 A lithium niobate external modulator using a directional coupler configuration.

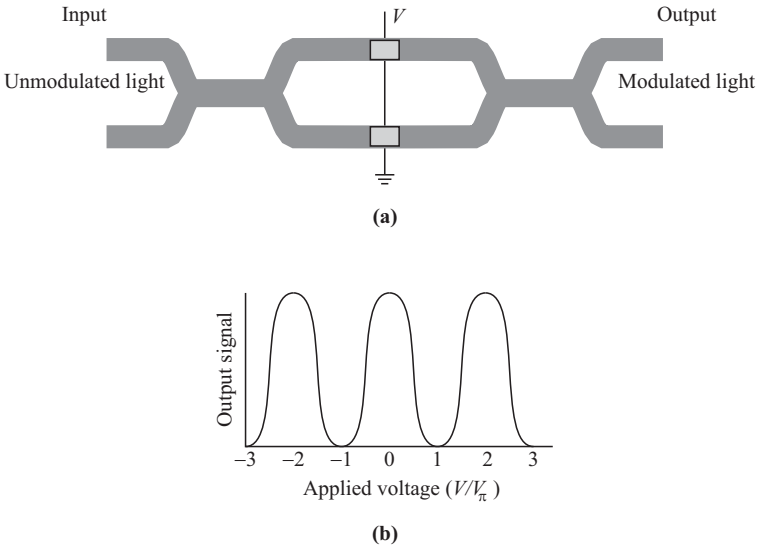

Figure 3.59 A lithium niobate external modulator using a Mach-Zehnder interferometer (MZI) configuration. (a) Device configuration. (b) Theoretical switching response as a function of applied voltage, V. V_π denotes the voltage required to achieve a π phase shift between the two arms. Note that the MZI has a periodic response.

modulator uses a material such that under normal conditions, its bandgap is higher than the photon energy of the incident light signal. This allows the light signal to propagate through. Applying an electric field to the modulator results in shrinking the bandgap of the material, causing the incident photons to be absorbed by the material. This effect is called the *Franz-Keldysh effect* or the *Stark effect*. The response time of this effect is sufficiently fast to enable us to realize 2.5 Gb/s and 10 Gb/s modulators. The chirp performance of EA modulators, though much better than directly modulated lasers, is not as good as that of lithium niobate MZI modulators. (While ideally there is no chirp in an external modulator, in practice, some chirp is induced in EA modulators because of residual phase modulation effects. This chirp can be controlled precisely in lithium niobate modulators.)

3.5.5 Pump Sources for Raman Amplifiers

One of the biggest challenges in realizing the Raman amplifiers that we discussed in Section 3.4.4 is a practical high-power pump source at the right wavelength. Since

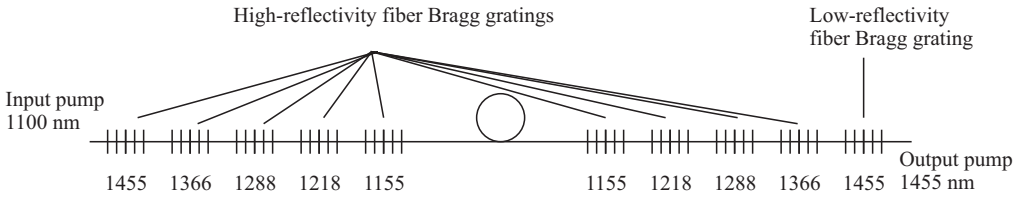

Figure 3.60 A high-power pump laser obtained by cascading resonators (after [Gru95]).

the Raman effect is only seen with very high powers in the fiber, pump powers on the order of several watts are required to provide effective amplification.

Several approaches have been proposed to realize high-power pump sources. One method is to combine a number of high-power semiconductor pump lasers. The power that can be extracted from a single semiconductor pump laser diode is limited to a few hundred milliwatts. Multiple semiconductor pump lasers can be combined using a combination of wavelength and/or polarization multiplexing to obtain a composite pump with sufficiently high power.

The other challenge lies in realizing the laser at the desired pump wavelength. One interesting approach is the cascaded Raman laser, shown in Figure 3.60.

Starting with a high-power pump laser at a conveniently available wavelength, we can generate pump sources at higher wavelengths using the Raman effect itself in fiber, by successively cascading a series of resonator structures. The individual resonators can be realized conveniently using fiber Bragg gratings or other filter structures. In Figure 3.60, a pump input at 1100 nm provides Raman gain into a fiber. A Fabry-Perot resonator is created in the fiber between by using a pair of matched fiber Bragg gratings that serve as wavelength-selective mirrors (see Section 3.3.5 for how the resonator works). The innermost resonator converts the initial pump signal into another pump signal at 1155 nm. It passes through signals at other wavelengths. The next resonator converts the 1155 nm pump into a 1218 nm pump. In principle, we can obtain any desired pump wavelength by cascading the appropriate series of resonators. The figure shows a series of resonators cascaded to obtain a 1455 nm pump output. The fiber Bragg grating at the end is designed to have lower reflectivity, allowing the 1455 nm pump signal to be output. This pump signal can then be used to provide Raman gain around 1550 nm. Due to the low fiber loss and high reflectivity of the fiber Bragg gratings, 80% of the input light is converted to the output.

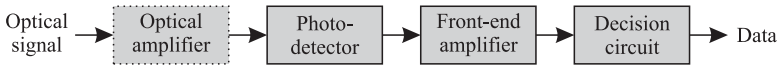

Figure 3.61 Block diagram of a receiver in a digital communication system.

3.6 Detectors

A receiver converts an optical signal into a usable electrical signal. Figure 3.61 shows the different components within a receiver. The *photodetector* generates an electrical current proportional to the incident optical power. The *front-end amplifier* increases the power of the generated electrical signal to a usable level. In digital communication systems, the front-end amplifier is followed by a *decision circuit* that estimates the data from the output of the front-end amplifier. The design of this decision circuit depends on the modulation scheme used to transmit the data and will be discussed in Section 4.4. An optical amplifier may be optionally placed before the photodetector to act as a *preamplifier*. The performance of optically preamplified receivers will be discussed in Chapter 4. This section covers photodetectors and front-end amplifiers.

3.6.1 Photodetectors

The basic principle of photodetection is illustrated in Figure 3.62. Photodetectors are made of semiconductor materials. Photons incident on a semiconductor are absorbed by electrons in the valence band. As a result, these electrons acquire higher energy and are excited into the conduction band, leaving behind a hole in the valence band. When an external voltage is applied to the semiconductor, these electron-hole pairs give rise to an electrical current, termed the *photocurrent*.

It is a principle of quantum mechanics that each electron can absorb only one photon to transit between energy levels. Thus the energy of the incident photon must be at least equal to the bandgap energy in order for a photocurrent to be generated. This is also illustrated in Figure 3.62. This gives us the following constraint on the frequency f_c or the wavelength λ at which a semiconductor material with bandgap E_g can be used as a photodetector:

$$hf_c = \frac{hc}{\lambda} \geq eE_g. \tag{3.19}$$

Here, c is the velocity of light, and e is the electronic charge.

The largest value of λ for which (3.19) is satisfied is called the *cutoff wavelength* and is denoted by λ_{cutoff}. Table 3.2 lists the bandgap energies and the corresponding

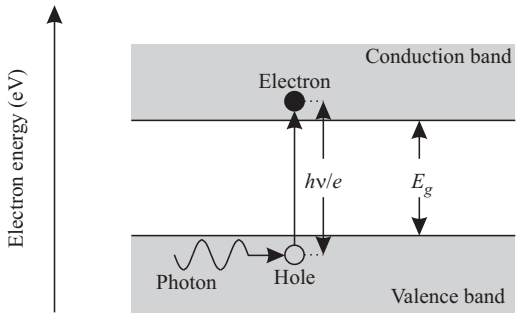

Figure 3.62 The basic principle of photodetection using a semiconductor. Incident photons are absorbed by electrons in the valence band, creating a free or mobile electron-hole pair. This electron-hole pair gives rise to a photocurrent when an external voltage is applied.

cutoff wavelengths for a number of semiconductor materials. We see from this table that the well-known semiconductors silicon (Si) and gallium arsenide (GaAs) cannot be used as photodetectors in the 1.3 and 1.55 μm bands. Although germanium (Ge) can be used to make photodetectors in both these bands, it has some disadvantages that reduce its effectiveness for this purpose. The new compounds indium gallium arsenide (InGaAs) and indium gallium arsenide phosphide (InGaAsP) are commonly used to make photodetectors in the 1.3 and 1.55 μm bands. Silicon photodetectors are widely used in the 0.8 μm band.

The fraction of the energy of the optical signal that is absorbed and gives rise to a photocurrent is called the *efficiency* η of the photodetector. For transmission at high bit rates over long distances, optical energy is scarce, and thus it is important to design the photodetector to achieve an efficiency η as close to 1 as possible. This can be achieved by using a semiconductor slab of sufficient thickness. The power absorbed by a semiconductor slab of thickness L μm can be written as

$$P_{\text{abs}} = (1 - e^{-\alpha L}) P_{\text{in}}, \tag{3.20}$$

where P_{in} is the incident optical signal power, and α is the absorption coefficient of the material; therefore,

$$\eta = \frac{P_{\text{abs}}}{P_{\text{in}}} = 1 - e^{-\alpha L}. \tag{3.21}$$

The absorption coefficient depends on the wavelength and is zero for wavelengths $\lambda > \lambda_{\text{cutoff}}$. Thus a semiconductor is transparent to wavelengths greater than its cutoff

Table 3.2 Bandgap energies and cutoff wavelengths for a number of semiconductor materials. $In_{1-x}Ga_xAs$ is a ternary compound semiconductor material where a fraction $1-x$ of the Ga atoms in GaAs are replaced by In atoms. $In_{1-x}Ga_xAs_yP_{1-y}$ is a quaternary compound semiconductor material where, in addition, a fraction $1-y$ of the As atoms are replaced by P atoms. By varying x and y, the bandgap energies and cutoff wavelengths can be varied.

Material	E_g (eV)	λ_{cutoff} (μm)
Si	1.17	1.06
Ge	0.775	1.6
GaAs	1.424	0.87
InP	1.35	0.92
$In_{0.55}Ga_{0.45}As$	0.75	1.65
$In_{1-0.45y}Ga_{0.45y}As_yP_{1-y}$	0.75–1.35	1.65–0.92

wavelength. Typical values of α are on the order of 10^4/cm, so to achieve an efficiency $\eta > 0.99$, a slab of thickness on the order of 10 μm is needed. The area of the photodetector is usually chosen to be sufficiently large so that all the incident optical power can be captured by it. Photodetectors have a very wide operating bandwidth since a photodetector at some wavelength can also serve as a photodetector at all smaller wavelengths. Thus a photodetector designed for the 1.55 μm band can also be used in the 1.3 μm band.

Photodetectors are commonly characterized by their *responsivity* $\mathcal{R}$. If a photodetector produces an average current of I_p amperes when the incident optical power is P_{in} watts, the responsivity

$$\mathcal{R} = \frac{I_p}{P_{\text{in}}} \text{ A/W}.$$

Since an incident optical power P_{in} corresponds to an incidence of P_{in}/hf_c photons/s on the average, and a fraction η of these incident photons are absorbed and generate an electron in the external circuit, we can write

$$\mathcal{R} = \frac{e\eta}{hf_c} \text{ A/W}.$$

The responsivity is commonly expressed in terms of λ; thus

$$\mathcal{R} = \frac{e\eta\lambda}{hc} = \frac{\eta\lambda}{1.24} \text{ A/W},$$

where λ in the last expression is expressed in μm. Since η can be made quite close to 1 in practice, the responsivities achieved are on the order of 1 A/W in the 1.3 μm band and 1.2 A/W in the 1.55 μm band.

In practice, the mere use of a slab of semiconductor as a photodetector does not realize high efficiencies. This is because many of the generated conduction band electrons recombine with holes in the valence band before they reach the external circuit. Thus it is necessary to sweep the generated conduction band electrons rapidly out of the semiconductor. This can be done by imposing an electric field of sufficient strength in the region where the electrons are generated. This is best achieved by using a semiconductor *pn*-junction (see Section 3.4.5) instead of a homogeneous slab and applying a *reverse-bias* voltage (positive bias to the *n*-type and negative bias to the *p*-type) to it, as shown in Figure 3.63. Such a photodetector is called a *photodiode*.

The depletion region in a *pn*-junction creates a built-in electric field. Both the depletion region and the built-in electric field can be enhanced by the application of a reverse-bias voltage. In this case, the electrons that are generated by the absorption of photons within or close to the depletion region will be swept into the *n*-type semiconductor before they recombine with the holes in the *p*-type semiconductor. This process is called *drift* and gives rise to a current in the external circuit. Similarly, the generated holes in or close to the depletion region drift into the *p*-type semiconductor because of the electric field.

Electron-hole pairs that are generated far away from the depletion region travel primarily under the effect of diffusion and may recombine without giving rise to a current in the external circuit. This reduces the efficiency η of the photodetector. More importantly, since diffusion is a much slower process than drift, the *diffusion current* that is generated by these electron-hole pairs will not respond quickly to changes in the intensity of the incident optical signal, thus reducing the frequency response of the photodiode.

pin Photodiodes

To improve the efficiency of the photodetector, a very lightly doped *intrinsic* semiconductor is introduced between the *p*-type and *n*-type semiconductors. Such photodiodes are called *pin* photodiodes, where the *i* in *pin* is for intrinsic. In these photodiodes, the depletion region extends completely across this intrinsic semiconductor (or region). The width of the *p*-type and *n*-type semiconductors is small compared to the intrinsic region, so that much of the light absorption takes place in this region. This increases the efficiency and thus the responsivity of the photodiode.

A more efficient method of increasing the responsivity is to use a semiconductor material for the *p*-type and *n*-type regions that is *transparent* at the wavelength

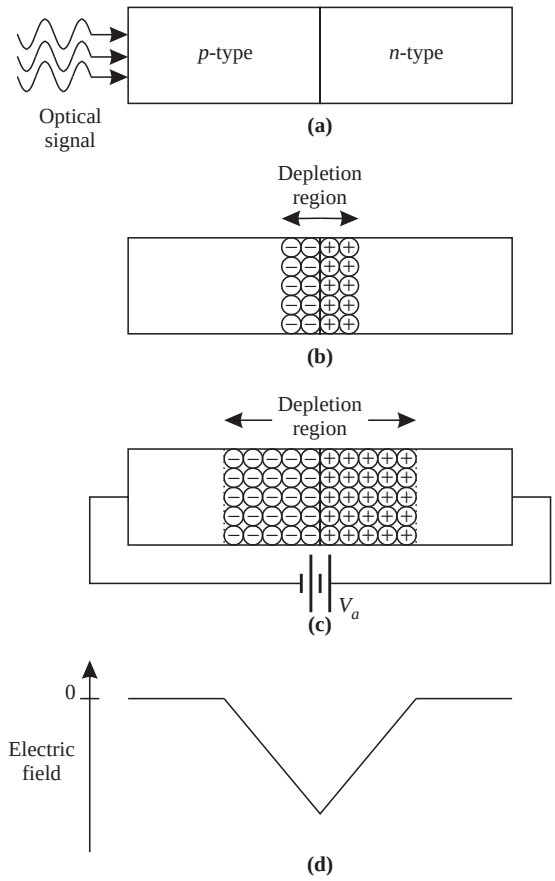

Figure 3.63 A reverse-biased *pn*-junction used as a photodetector. (a) A *pn*-junction photodiode. (b) Depletion region with no bias voltage applied. (c) Depletion region with a reverse-bias voltage, V_a. (d) Built-in electric field on reverse bias.

of interest. Thus the wavelength of interest is larger than the cutoff wavelength of this semiconductor, and no absorption of light takes place in these regions. This is illustrated in Figure 3.64, where the material InP is used for the *p*-type and *n*-type regions, and InGaAs for the intrinsic region. Such a *pin* photodiode structure is termed a *double heterojunction* or a *heterostructure* since it consists of two junctions of completely different semiconductor materials. From Table 3.2, we see that the cutoff wavelength for InP is 0.92 μm and that for InGaAs is 1.65 μm. Thus the

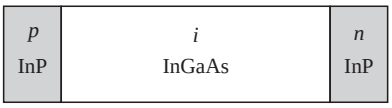

Figure 3.64 A *pin* photodiode based on a heterostructure. The *p*-type and *n*-type regions are made of InP, which is transparent in the 1.3 and 1.55 μm wavelength bands. The intrinsic region is made of InGaAs, which strongly absorbs in both these bands.

p-type and *n*-type regions are transparent in the 1.3–1.6 μm range, and the diffusion component of the photocurrent is completely eliminated.

Avalanche Photodiodes

The responsivities of the photodetectors we have described thus far have been limited by the fact that one photon can generate only one electron when it is absorbed. However, if the generated electron is subjected to a very high electric field, it can acquire sufficient energy to knock off more electrons from the valence band to the conduction band. These secondary electron-hole pairs can generate even further electron-hole pairs when they are accelerated to sufficient levels. This process is called *avalanche multiplication*. Such a photodiode is called an *avalanche photodiode,* or simply an *APD*.

The number of secondary electron-hole pairs generated by the avalanche multiplication process by a single (primary) electron is random, and the mean value of this number is termed the *multiplicative gain* and denoted by G_m. The multiplicative gain of an APD can be made quite large and even infinite—a condition called *avalanche breakdown*. However, a large value of G_m is also accompanied by a larger variance in the generated photocurrent, which adversely affects the noise performance of the APD. Thus there is a trade-off between the multiplicative gain and the noise factor. APDs are usually designed to have a moderate value of G_m that optimizes their performance. We will study this issue further in Section 4.4.

3.6.2 Front-End Amplifiers

Two kinds of front-end amplifiers are used in optical communication systems: the *high-impedance* front end and the *transimpedance* front end. The equivalent circuits for these amplifiers are shown in Figure 3.65.

The capacitances C in this figure include the capacitance due to the photodiode, the amplifier input capacitance, and other parasitic capacitances. The main design issue is the choice of the load resistance R_L. We will see in Chapter 4 that the *thermal*

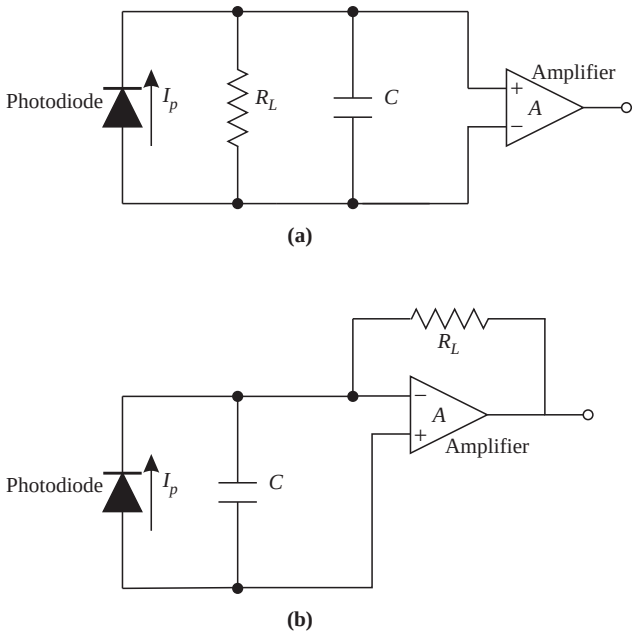

Figure 3.65 (a) Equivalent circuit for a high-impedance front-end amplifier. (b) Equivalent circuit for a transimpedance front-end amplifier.

noise current that arises due to the random motion of electrons and contaminates the photocurrent is inversely proportional to the load resistance. Thus, to minimize the thermal noise, we must make R_L large. However, the bandwidth of the photodiode, which sets the upper limit on the usable bit rate, is inversely proportional to the output load resistance seen by the photodiode, say, R_p. First consider the high-impedance front end. In this case, $R_p = R_L$, and we must choose R_L small enough to accommodate the bit rate of the system. Thus there is a trade-off between the bandwidth of the photodiode and its noise performance. Now consider the transimpedance front end for which $R_p = R_L/(A + 1)$, where A is the gain of the amplifier. The bandwidth is increased by a factor of $A + 1$ for the same load resistance. However, the thermal noise current is also higher than that of a high-impedance amplifier with the same R_L (due to considerations beyond the scope of this book), but this increase is quite moderate—a factor usually less than two. Thus the transimpedance front end is chosen over the high-impedance one for most optical communication systems.

There is another consideration in the choice of a front-end amplifier: *dynamic range*. This is the difference between the largest and smallest signal levels that the

front-end amplifier can handle. This may not be an important consideration for many optical communication links since the power level seen by the receivers is usually more or less fixed. However, dynamic range of the receivers is a very important consideration in the case of networks where the received signal level can vary by a few orders of magnitude, depending on the location of the source in the network. The transimpedance amplifier has a significantly higher dynamic range than the high-impedance one, and this is another factor in favor of choosing the transimpedance amplifier. The higher dynamic range arises because large variations in the photocurrent I_p translate into much smaller variations at the amplifier input, particularly if the amplifier gain is large. This can be understood with reference to Figure 3.65(b). A change ΔI_p in the photocurrent causes a change in voltage $\Delta I_p R_L$ across the resistance R_L (ignoring the current through the capacitance C). This results in a voltage change across the inputs of the amplifier of only $\Delta I_p R_L/(A+1)$. Thus if the gain, A, is large, this voltage change is small. In the case of the high-impedance amplifier, however, the voltage change across the amplifier inputs would be $\Delta I_p R_L$ (again ignoring the current through the capacitance C).

A *field-effect transistor* (FET) has a very high input impedance and for this reason is often used as the amplifier in the front end. A *pin* photodiode and an FET are often integrated on the same semiconductor substrate, and the combined device is called a *pinFET.*

3.7 Switches

Optical switches are used in optical networks for a variety of applications. The different applications require different switching times and number of switch ports, as summarized in Table 3.3. One application of optical switches is in the *provisioning* of lightpaths. In this application, the switches are used inside wavelength crossconnects to reconfigure them to support new lightpaths. In this application, the switches are replacements for manual fiber patch panels, but with significant added software for end-to-end network management, a subject that we will cover in detail in Chapters 8 and 9. Thus, for this application, switches with millisecond switching times are acceptable. The challenge here is to realize large switch sizes.

Another important application is that of *protection switching*, the subject of Chapter 9. Here the switches are used to switch the traffic stream from a primary fiber onto another fiber in case the primary fiber fails. The entire operation must typically be completed in several tens of milliseconds, which includes the time to detect the failure, communicate the failure to the appropriate network elements handling the switching, and the actual switch time. Thus the switching time required is on the order of a few milliseconds. Different types of protection switching are

Table 3.3 Applications for optical switches and their switching time and port count requirements.

Application	Switching Time Required	Number of Ports
Provisioning	1–10 ms	> 1000
Protection switching	1–10 ms	2–1000
Packet switching	1 ns	> 100
External modulation	10 ps	1

possible, and based on the scheme used, the number of switch ports needed may vary from two ports to several hundreds to thousands of ports when used in a wavelength crossconnect.

Switches are also important components in high-speed optical *packet-switched* networks. In these networks, switches are used to switch signals on a packet-by-packet basis. For this application, the switching time must be much smaller than a packet duration, and large switches will be needed. For example, ordinary Ethernet packets have lengths between about 60 to 1500 bytes. At 10 Gb/s, the transmission time of a 60-byte packet is 48 ns. Thus, the switching time required for efficient operation is on the order of a few nanoseconds. Optical packet switching is the subject of Chapter 12.

Yet another use for switches is as external modulators to turn on and off the data in front of a laser source. In this case, the switching time must be a small fraction of the bit duration. So an external modulator for a 10 Gb/s signal (with a bit duration of 100 ps) must have a switching time (or, equivalently, a rise and fall time) of about 10 ps.

In addition to the switching time and the number of ports, the other important parameters used to characterize the suitability of a switch for optical networking applications are the following:

1. The *extinction ratio* of an on-off switch is the ratio of the output power in the on state to the output power in the off state. This ratio should be as large as possible and is particularly important in external modulators. Whereas simple mechanical switches have extinction ratios of 40–50 dB, high-speed external modulators tend to have extinction ratios of 10–25 dB.
2. The *insertion loss* of a switch is the fraction of power (usually expressed in decibels) that is lost because of the presence of the switch and must be as small as possible. Some switches have different losses for different input-output connections. This is an undesirable feature because it increases the dynamic range of the

signals in the network. With such switches, we may need to include variable optical attenuators to equalize the loss across different paths. This *loss uniformity* is determined primarily by the architecture used to build the switch, rather than the inherent technology itself, as we will see in several examples below.

3. Switches are not ideal. Even if input x is nominally connected to output y, some power from input x may appear at the other outputs. For a given switching state or interconnection pattern, and output, the *crosstalk* is the ratio of the power at that output from the desired input to the power from all other inputs. Usually, the *crosstalk of a switch* is defined as the worst-case crosstalk over all outputs and interconnection patterns.
4. As with other components, switches should have a low polarization-dependent loss (PDL). When used as external modulators, polarization dependence can be tolerated since the switch is used immediately following the laser, and the laser's output state of polarization can be controlled by using a special polarization-preserving fiber to couple the light from the laser into the external modulator.
5. A *latching* switch maintains its switch state even if power is turned off to the switch. This is a somewhat desirable feature because it enables traffic to be passed through the switch even in the event of power failures.
6. The switch needs to have a readout capability wherein its current state can be monitored. This is important to verify that the right connections are made through the switch.
7. The reliability of the switch is an important factor in telecommunications applications. The common way of establishing reliability is to cycle the switch through its various states a large number of times, perhaps a few million cycles. However, in the provisioning and protection-switching applications discussed above, the switch remains in one state for a long period, say, even a few years, and is then activated to change state. The reliability issue here is whether the switch will actually switch after it has remained untouched for a long period. This property is more difficult to establish without a long-term history of deployment.

3.7.1 *Large Optical Switches*

Switches with port counts ranging from a few hundred to a few thousand are being sought by carriers for their next-generation networks. Given that a single central office handles multiple fibers, with each fiber carrying several tens to hundreds of wavelengths, it is easy to imagine the need for large-scale switches to provision and

protect these wavelengths. We will study the use of such switches as wavelength crossconnects in Chapter 7.

The main considerations in building large switches are the following:

Number of switch elements required. Large switches are made by using multiple switch elements in some form or the other, as we will see below. The cost and complexity of the switch to some extent depends on the number of switch elements required. However, this is only one of the factors that affects the cost. Other factors include packaging, splicing, and ease of fabrication and control.

Loss uniformity. As we mentioned in the context of switch characteristics earlier, switches may have different losses for different combinations of input and output ports. This situation is exacerbated for large switches. A measure of the loss uniformity can be obtained by considering the minimum and maximum number of switch elements in the optical path, for different input and output combinations.

Number of crossovers. Some of the optical switches that we will study next are fabricated by integrating multiple switch elements on a single substrate. Unlike integrated electronic circuits (ICs), where connections between the various components can be made at multiple layers, in integrated optics, all these connections must be made in a single layer by means of waveguides. If the paths of two waveguides cross, two undesirable effects are introduced: power loss and crosstalk. In order to have acceptable loss and crosstalk performance for the switch, it is thus desirable to minimize, or completely eliminate, such waveguide crossovers. Crossovers are not an issue with respect to free-space switches, such as the MEMS switches that we will describe later in this section.

Blocking characteristics. In terms of the switching function achievable, switches are of two types: *blocking* or *nonblocking*. A switch is said to be *nonblocking* if an unused input port can be connected to any unused output port. Thus a nonblocking switch is capable of realizing every interconnection pattern between the inputs and the outputs. If some interconnection pattern(s) cannot be realized, the switch is said to be *blocking*. Most applications require nonblocking switches. However, even nonblocking switches can be further distinguished in terms of the effort needed to achieve the nonblocking property. A switch is said to be *wide-sense nonblocking* if any unused input can be connected to any unused output, without requiring any existing connection to be rerouted. Wide-sense nonblocking switches usually make use of specific routing algorithms to route connections so that future connections will not be blocked. A *strict-sense nonblocking* switch allows any unused input to be connected to any unused output regardless of how previous connections were made through the switch.

Table 3.4 Comparison of different switch architectures. The switch count for the Spanke architecture is made in terms of $1 \times n$ switches, whereas 2×2 switches are used for the other architectures.

	Nonblocking Type	No. Switches	Max. Loss	Min. Loss
Crossbar	Wide sense	n^2	$2n-1$	1
Clos	Strict sense	$4\sqrt{2}n^{1.5}$	$5\sqrt{2n}-5$	3
Spanke	Strict sense	$2n$	2	2
Beneš	Rearrangeable	$\frac{n}{2}(2\log_2 n - 1)$	$2\log_2 n - 1$	$2\log_2 n - 1$
Spanke-Beneš	Rearrangeable	$\frac{n}{2}(n-1)$	n	$\frac{n}{2}$

A nonblocking switch that may require rerouting of connections to achieve the nonblocking property is said to be *rearrangeably nonblocking*. Rerouting of connections may or may not be acceptable depending on the application since the connection must be interrupted, at least briefly, in order to switch it to a different path. The advantage of rearrangeably nonblocking switch architectures is that they use fewer small switches to build a larger switch of a given size, compared to the wide-sense nonblocking switch architectures.

While rearrangeably nonblocking architectures use fewer switches, they require a more complex control algorithm to set up connections, but this control complexity is not a significant issue, given the power of today's microprocessors used in these switches that would execute such an algorithm. The main drawback of rearrangeably nonblocking switches is that many applications will not allow existing connections to be disrupted, even temporarily, to accommodate a new connection.

Usually, there is a trade-off between these different aspects. We will illustrate this when we study different architectures for building large switches next. Table 3.4 compares the characteristics of these architectures.

Crossbar

A 4×4 crossbar switch is shown in Figure 3.66. This switch uses 16 2×2 switches, and the interconnection between inputs and outputs is achieved by appropriately setting the states of these 2×2 switches. The settings of the 2×2 switches required to connect input 1 to output 3 are shown in Figure 3.66. This connection can be viewed as taking a path through the network of 2×2 switches making up the 4×4 switch. Note that there are other paths from input 1 to output 3; however, this is the preferred path as we will see next.

The crossbar architecture is wide-sense nonblocking. To connect input i to output j, the path taken traverses the 2×2 switches in row i till it reaches column j and then

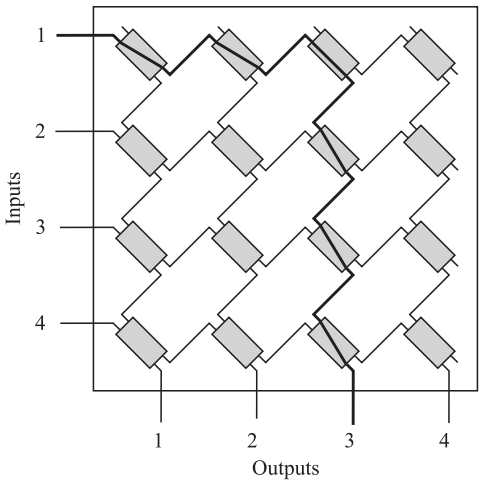

Figure 3.66 A 4×4 crossbar switch realized using 16 2×2 switches.

traverses the switches in column j till it reaches output j. Thus the 2×2 switches on this path in row i and column j must be set appropriately for this connection to be made. We leave it to you to be convinced that *if this connection rule is used,* this switch is nonblocking and does not require existing connections to be rerouted.

In general, an $n \times n$ crossbar requires n^2 2×2 switches. The shortest path length is 1 and the longest path length is $2n - 1$, and this is one of the main drawbacks of the crossbar architecture. The switch can be fabricated without any crossovers.

Clos

The Clos architecture provides a strict-sense nonblocking switch and is widely used in practice to build large port count switches. A three-stage 1024-port Clos switch is shown in Figure 3.67. An $n \times n$ switch is constructed as follows. We use three parameters, m, k, and p. Let $n = mk$. The first and third stage consist of k $(m \times p)$ switches. The middle stage consists of p $(k \times k)$ switches. Each of the k switches in the first stage is connected to all the switches in the middle stage. (Each switch in the first stage has p outputs. Each output is connected to the input of a different switch in the middle stage.) Likewise, each of the k switches in the third stage is connected to all the switches in the middle stage. We leave it to you to verify that if $p \geq 2m - 1$, the switch is strictly nonblocking (see Problem 3.29).

To minimize the cost of the switch, let us pick $p = 2m - 1$. Usually, the individual switches in each stage are designed using crossbar switches. Thus each of the $m \times$

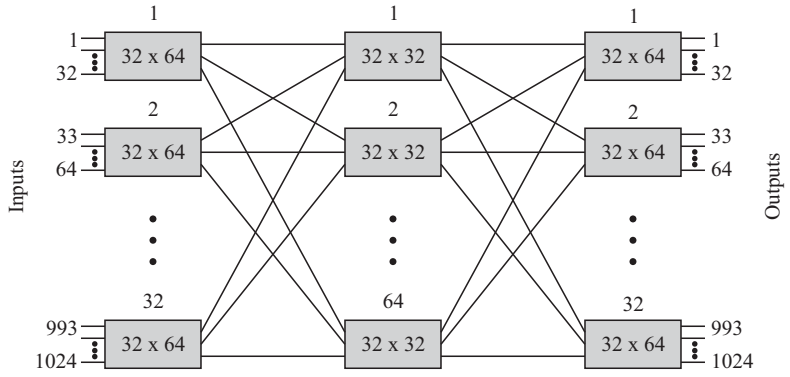

Figure 3.67 A strict-sense nonblocking 1024 × 1024 switch realized using 32 × 64 and 32 × 32 switches interconnected in a three-stage Clos architecture.

$(2m - 1)$ switches requires $m(2m - 1)$ 2×2 switch elements, and each of the $k \times k$ switches in the middle stage requires k^2 2×2 switch elements. The total number of switch elements needed is therefore

$$2km(2m - 1) + (2m - 1)k^2.$$

Using $k = n/m$, we leave it to you to verify that the number of switch elements is minimized when

$$m \approx \sqrt{\frac{n}{2}}.$$

Using this value for m, the number of switch elements required for the minimum cost configuration is approximately

$$4\sqrt{2}n^{3/2} - 4n,$$

which is significantly lower than the n^2 required for a crossbar.

The Clos architecture has several advantages that make it suitable for use in a multistage switch fabric. The loss uniformity between different input-output combinations is better than a crossbar, and the number of switch elements required is significantly smaller than a crossbar.

Spanke

The Spanke architecture shown in Figure 3.68 is turning out to be a popular architecture for building large switches. An $n \times n$ switch is made by combining n $1 \times n$ switches along with n $n \times 1$ switches, as shown in the figure. The architecture is

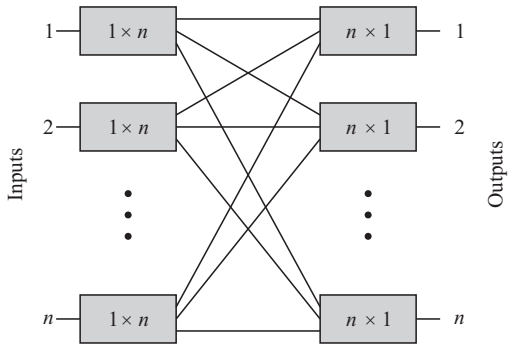

Figure 3.68 A strict-sense nonblocking $n \times n$ switch realized using $2n$ $1 \times n$ switches interconnected in the Spanke architecture.

strict-sense nonblocking. So far we have been counting the number of 2×2 switch elements needed to build large switches as a measure of the switch cost. What makes the Spanke architecture attractive is that, in many cases, a $1 \times n$ optical switch can be built using a single switch element and does not need to be built out of 1×2 or 2×2 switch elements. This is the case with the MEMS analog beam steering mirror technology that we will discuss later in this section. Therefore, only $2n$ such switch elements are needed to build an $n \times n$ switch. This implies that the switch cost scales linearly with n, which is significantly better than other switch architectures. In addition, each connection passes through two switch elements, which is significantly smaller than the number of switch elements in the path for other multistage designs. This approach provides a much lower insertion loss than the multistage designs. Moreover, the optical path length for all the input–output combinations can be made essentially the same, so that the loss is the same regardless of the specific input–output combination.

Beneš

The Beneš architecture is a rearrangeably nonblocking switch architecture and is one of the most efficient switch architectures in terms of the number of 2×2 switches it uses to build larger switches. A rearrangeably nonblocking 8×8 switch that uses only 20 2×2 switches is shown in Figure 3.69. In comparison, an 8×8 crossbar switch requires 64 2×2 switches. In general, an $n \times n$ Beneš switch requires $(n/2)(2 \log_2 n - 1)$ 2×2 switches, n being a power of two. The loss is the same through every path in the switch—each path goes through $2 \log_2 n - 1$ 2×2 switches.

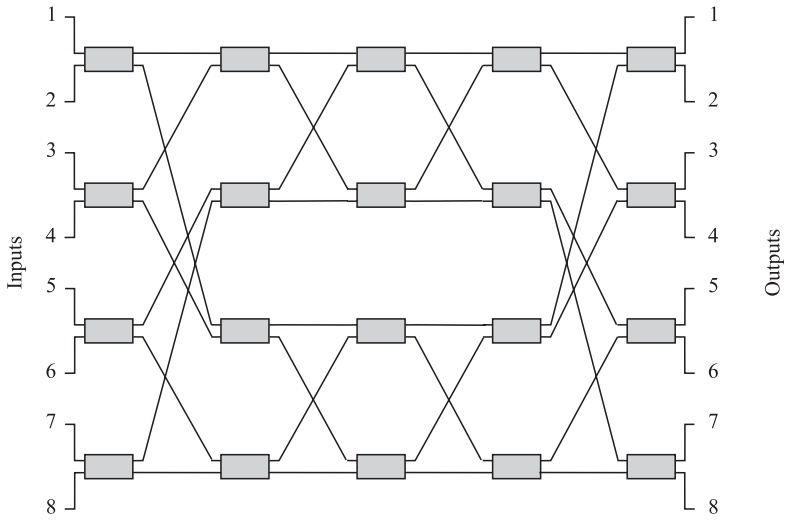

Figure 3.69 A rearrangeably nonblocking 8×8 switch realized using 20 2×2 switches interconnected in the Beneš architecture.

Its two main drawbacks are that it is not wide-sense nonblocking and that a number of waveguide crossovers are required, making it difficult to fabricate in integrated optics.

Spanke-Beneš

A good compromise between the crossbar and Beneš switch architectures is shown in Figure 3.70, which is a rearrangeably nonblocking 8×8 switch using 28 2×2 switches and *no* waveguide crossovers. This switch architecture was discovered by Spanke and Beneš [SB87] and is called the *n-stage planar architecture* since it requires n stages (columns) to realize an $n \times n$ switch. It requires $n(n-1)/2$ switches, the shortest path length is $n/2$, and the longest path length is n. There are no crossovers. Its main drawbacks are that it is not wide-sense nonblocking and the loss is nonuniform.

3.7.2 Optical Switch Technologies

Many different technologies are available to realize optical switches. These are compared in Table 3.5. With the exception of the large-scale MEMS switch, the switch elements described in the next section all use the crossbar architecture.

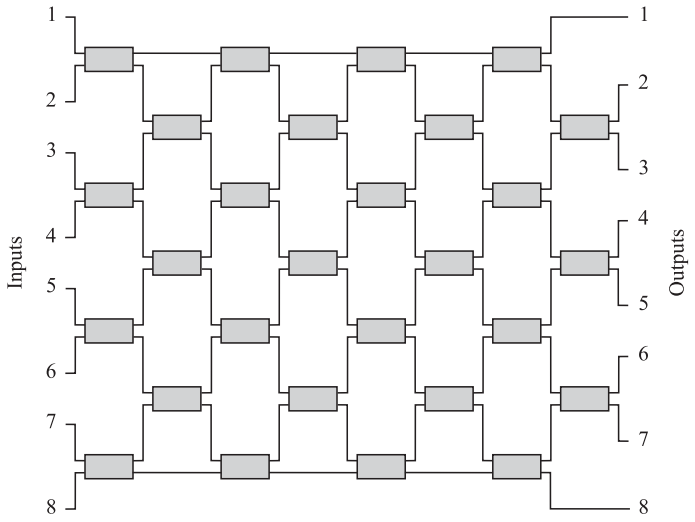

Figure 3.70 A rearrangeably nonblocking 8×8 switch realized using 28 2×2 switches and no waveguide crossovers interconnected in the n-stage planar architecture.

Table 3.5 Comparison of different optical switching technologies. The mechanical, MEMS, and polymer-based switches behave in the same manner for 1.3 and 1.55 μm wavelengths, but other switches are designed to operate at only one of these wavelength bands. The numbers represent parameters for commercially available switches in early 2001.

Type	Size	Loss (dB)	Crosstalk (dB)	PDL (dB)	Switching Time
Bulk mechanical	8×8	3	55	0.2	10 ms
2D MEMS	32×32	5	55	0.2	10 ms
3D MEMS	1000×1000	5	55	0.5	10 ms
Thermo-optic silica	8×8	8	40	Low	3 ms
Liquid crystal	2×2	1	35	0.1	4 ms
Polymer	8×8	10	30	Low	2 ms
Electro-optic $LiNbO_3$	4×4	8	35	1	10 ps
SOA	4×4	0	40	Low	1 ns

Bulk Mechanical Switches

In mechanical switches, the switching function is performed by some mechanical means. One such switch uses a mirror arrangement whereby the switching state is controlled by moving a mirror in and out of the optical path. Another type of mechanical switch uses a directional coupler. Bending or stretching the fiber in the interaction region changes the coupling ratio of the coupler and can be used to switch light from an input port between different output ports.

Bulk mechanical switches have low insertion losses, low PDL, and low crosstalk, and are relatively inexpensive devices. In most cases, they are available in a crossbar configuration, which implies somewhat poor loss uniformity. However, their switching speeds are on the order of a few milliseconds and the number of ports is fairly small, say, 8 to 16. For these reasons, they are particularly suited for use in small wavelength crossconnects for provisioning and protection-switching applications but not for the other applications discussed earlier. As with most mechanical components, long-term reliability for these switches is of some concern. Larger switches can be realized by cascading small bulk mechanical switches, as we saw in Section 3.7.1, but there are better ways of realizing larger port count switches, as we will explore next.

Micro-Electro-Mechanical System (MEMS) Switches

Micro-electro-mechanical systems (MEMS) are miniature mechanical devices typically fabricated using silicon substrates. In the context of optical switches, MEMS usually refers to miniature movable mirrors fabricated in silicon, with dimensions ranging from a few hundred micrometers to a few millimeters. A single silicon wafer yields a large number of mirrors, which means that these mirrors can be manufactured and packaged as arrays. Moreover, the mirrors can be fabricated using fairly standard semiconductor manufacturing processes. These mirrors are deflected from one position to another using a variety of electronic actuation techniques, such as electromagnetic, electrostatic, or piezoelectric methods, hence the name MEMS. Of these methods, electrostatic deflection is particularly power efficient but is relatively hard to control over a wide deflection range.

The simplest mirror structure is a so-called two-state pop-up mirror, or 2D mirror, shown in Figure 3.71. In one state, the mirror is flat in line with the substrate. In this state, the light beam is not deflected. In the other state, the mirror pops up to a vertical position, and the light beam, if present, is deflected. Such a mirror can be used in a crossbar arrangement discussed below to realize an $n \times n$ switch. Practical switch module sizes are limited by wafer sizes and processing constraints to be around 32×32. These switches are particularly easy to control through digital means, as only two mirror positions need to be supported.

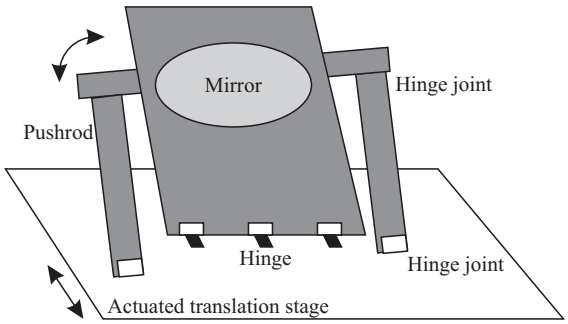

Figure 3.71 A two-state pop-up MEMS mirror, from [LGT98], shown in the popped-up position. The mirror can be moved to fold flat in its other position.

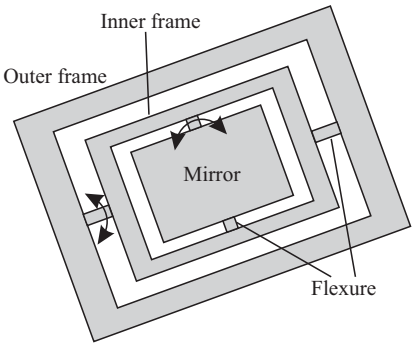

Figure 3.72 An analog beam steering mirror. The mirror can be freely rotated on two axes to deflect an incident light beam.

Another type of mirror structure is shown in Figure 3.72. The mirror is connected through flexures to an inner frame, which in turn is connected through another set of flexures to an outer frame. The flexures allow the mirror to be rotated freely on two distinct axes. This mirror can be controlled in an analog fashion to realize a continuous range of angular deflections. This type of mirror is sometimes referred to as an analog beam steering mirror, a gimbel mirror, or a 3D mirror. A mirror of this type can be used to realize a $1 \times n$ switch. The control of these mirrors is not a trivial matter, with fairly sophisticated servo control mechanisms required to deflect the mirrors to their correct position and hold them there.

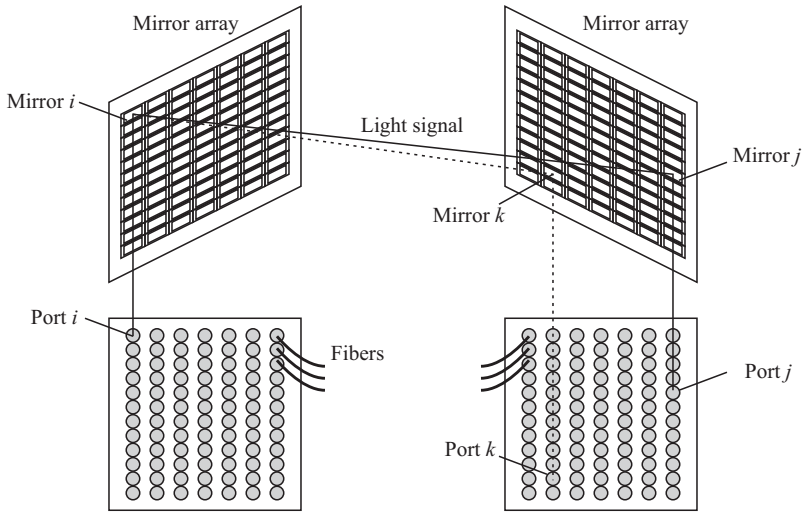

Figure 3.73 An $n \times n$ switch built using two arrays of analog beam steering MEMS mirrors.

Figure 3.73 shows a large $n \times n$ switch using two arrays of analog beam steering mirrors. This architecture corresponds to the Spanke architecture, which we discussed in Section 3.7.1. Each array has n mirrors, one associated with each switch port. An input signal is coupled to its associated mirror in the first array using a suitable arrangement of collimating lenses. The first mirror can be deflected to point the beam to any of the mirrors in the second array. To make a connection from port i to port j, the mirror i in the first array is pointed to mirror j in the second array and vice versa. Mirror j then allows the beam to be coupled out of port j. To make a connection from port i to another port, say, port k, mirror i in the first array and mirror k in the second array are pointed at each other. Note that in order to switch this connection from port i to port k, the beam is scanned from output mirror j to output mirror k, passing over other mirrors along the way. This does not lead to additional crosstalk because a connection is established only when the two mirrors are pointed at each other and not under any other circumstances. Note also that beams corresponding to multiple connections cross each other inside the switch but do not interfere.

There are two types of fabrication techniques used to make MEMS structures: *surface micromachining* and *bulk micromachining*. In surface micromachining, multiple layers are deposited on top of a silicon substrate. These layers are partially

etched away, and pieces are left anchored to the substrate to produce various structures. In bulk micromachining, the MEMS structures are crafted directly from the bulk of the silicon wafer. The type of micromachining used and the choice of the appropriate type of silicon substrate directly influence the properties of the resulting structure. For a more detailed discussion on some of the pros and cons of these approaches, see [NR01]. Today we are seeing the simple 2D MEMS mirrors realized using surface micromachining and the 3D MEMS mirrors realized using bulk micromachining.

Among the various technologies discussed in this section, the 3D MEMS analog beam steering mirror technology offers the best potential for building large-scale optical switches, for example, 256 to 1000 ports. These switches are compact, have very good optical properties (low loss, good loss uniformity, negligible dispersion), and can have extremely low power consumption. Most of the other technologies are limited to small switch sizes.

Liquid Crystal Switches

Liquid crystal cells offer another way for realizing small optical switches. These switches typically make use of polarization effects to perform the switching function. By applying a voltage to a suitably designed liquid crystal cell, we can cause the polarization of the light passing through the cell either to be rotated or not. This can then be combined with passive polarization beam splitters and combiners to yield a polarization-independent switch, as shown in Figure 3.74. The principle of operation is similar to the polarization-independent isolator of Figure 3.5. Typically, the passive polarization beam splitter, combiner, and active switch element can all be realized using an array of liquid crystal cells. The polarization rotation in the liquid crystal cell does not have to be digital in nature—it can be controlled in an analog fashion by controlling the voltage. Thus this technology can be used to realize a variable optical attentuator (VOA) as well. In fact, the VOA can be incorporated in the switch itself to control the output power being coupled out. The switching time is on the order of a few milliseconds. Like the bubble-based waveguide switch, a liquid crystal switch is a solid-state device. Thus, it can be better manufactured in volume and low cost.

Electro-Optic Switches

A 2×2 electro-optic switch can be realized using one of the external modulator configurations that we studied in Section 3.5.4. One commonly used material is lithium niobate ($LiNbO_3$). In the directional coupler configuration, the coupling ratio is varied by changing the voltage and thus the refractive index of the material in the coupling region. In the Mach-Zehnder configuration, the relative path length

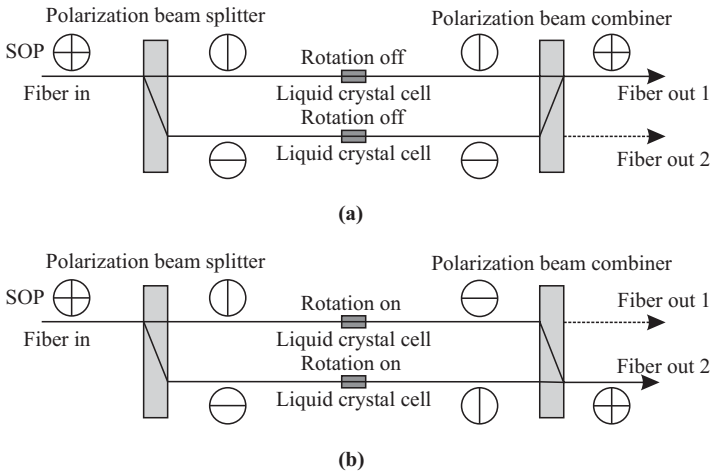

Figure 3.74 A 1 × 2 liquid crystal switch. (a) The rotation is turned off, causing the light beam to exit on output port 1. (b) The rotation is turned on by applying a voltage to the liquid crystal cell, causing the light beam to exit on output port 2.

between the two arms of the Mach-Zehnder is varied. An electro-optic switch is capable of changing its state extremely rapidly—typically, in less than 1 ns. This switching time limit is determined by the capacitance of the electrode configuration.

Among the advantages of lithium niobate switches are that they allow modest levels of integration, compared to mechanical switches. Larger switches can be realized by integrating several 2 × 2 switches on a single substrate. However, they tend to have a relatively high loss and PDL, and are more expensive than mechanical switches.

Thermo-Optic Switches

These switches are essentially 2 × 2 integrated-optic Mach-Zehnder interferometers, constructed on waveguide material whose refractive index is a function of the temperature. By varying the refractive index in one arm of the interferometer, the relative phase difference between the two arms can be changed, resulting in switching an input signal from one output port to another. These devices have been made on silica as well as polymer substrates, but have relatively poor crosstalk. Also the thermo-optic effect is quite slow, and switching speeds are on the order of a few milliseconds.

Semiconductor Optical Amplifier Switches

The SOA described in Section 3.4.5 can be used as an on-off switch by varying the bias voltage to the device. If the bias voltage is reduced, no population inversion is achieved, and the device absorbs input signals. If the bias voltage is present, it amplifies the input signals. The combination of amplification in the on state and absorption in the off state makes this device capable of achieving very large extinction ratios. The switching speed is on the order of 1 ns. Larger switches can be fabricated by integrating SOAs with passive couplers. However, this is an expensive component, and it is difficult to make it polarization independent because of the highly directional orientation of the laser active region, whose width is almost always much greater than its height (except for VCSELs).

3.7.3 *Large Electronic Switches*

We have focused primarily on optical switch technologies in this section. However, many of the practical "optical" or wavelength crossconnects actually use electronic switch fabrics.

Typically, a large electronic switch uses a multistage design, and in many cases, the Clos approach is the preferred approach as it provides a strict-sense nonblocking architecture with a relatively small number of crosspoint switches. Two approaches are possible. In the first approach, the input signal at 2.5 Gb/s or 10 Gb/s is converted into a parallel bit stream at a manageable rate, say, 51 Mb/s, and all the switching is done at the latter bit rate. This approach makes sense if we need to switch the signal in units of 51 Mb/s for other reasons. Also in many cases, the overall cost of an electronic switch is dominated by the cost of the optical to electrical converters, rather than the switch fabric itself. This implies that once the signal is available in the electrical domain, it makes sense to switch signals at a fine granularity.

The other approach is to design the switch to operate at the line rate in a serial fashion without splitting the signal into lower-speed bit streams. The basic unit of this serial approach is a crossbar fabricated as a single integrated circuit (IC). The practical considerations related to building larger switches using these ICs have to do with managing the power dissipation and the interconnects between switch stages. For example, suppose a 64×64 switch IC dissipates 25 W. About 100 such switches are required to build a 1024×1024 switch. The total power dissipated is therefore around 25 kW. (In contrast, a 1024×1024 optical switch using 3D MEMS may consume only about 3 kW and is significantly more compact overall, compared to an equivalent electrical switch.) Cooling such a switch is a significant problem. The other aspect has to do with the high-speed interconnect required between switch modules. As long as the switch modules are within a single printed circuit board,

the interconnections are not difficult. However, practical considerations of power dissipation and board space dictate the necessity for having multiple printed circuit boards and perhaps multiple racks of equipment. The interconnects between these boards and racks need to operate at the line rate, which is typically 2.5 Gb/s or higher. High-quality electrical interconnects or optical interconnects can be used for this purpose. The drivers required for the electrical interconnects also dissipate a significant amount of power, and the distances possible are limited, typically to 5–6 m. Optical interconnects make use of arrayed lasers and receivers along with fiber optic ribbon cables. These offer lower power dissipation and significantly longer reach between boards, typically to about 100 m or greater.

3.8 Wavelength Converters

A wavelength converter is a device that converts data from one incoming wavelength to another outgoing wavelength. Wavelength converters are useful components in WDM networks for three major reasons. First, data may enter the network at a wavelength that is not suitable for use within the network. For example, the first-generation networks of Chapter 6 commonly transmit data in the 1310 nm wavelength window, using LEDs or Fabry-Perot lasers. Neither the wavelength nor the type of laser is compatible with WDM networks. So at the inputs and outputs of the network, data must be converted from these wavelengths to narrow-band WDM signals in the 1550 nm wavelength range. A wavelength converter used to perform this function is sometimes called a *transponder.*

Second, wavelength converters may be needed within the network to improve the utilization of the available wavelengths on the network links. This topic is studied in detail in Chapter 10.

Finally, wavelength converters may be needed at boundaries between different networks if the different networks are managed by different entities and these entities do not coordinate the allocation of wavelengths in their networks.

Wavelength converters can be classified based on the range of wavelengths that they can handle at their inputs and outputs. A fixed-input, fixed-output device always takes in a fixed-input wavelength and converts it to a fixed-output wavelength. A variable-input, fixed-output device takes in a variety of wavelengths but always converts the input signal to a fixed-output wavelength. A fixed-input, variable-output device does the opposite function. Finally, a variable-input, variable-output device can convert any input wavelength to any output wavelength.

In addition to the range of wavelengths at the input and output, we also need to consider the range of input optical powers that the converter can handle, whether the converter is transparent to the bit rate and modulation format of the input signals,

and whether it introduces additional noise or phase jitter to the signal. We will see that the latter two characteristics depend on the type of regeneration used in the converter. For all-optical wavelength converters, polarization-dependent loss should also be kept to a minimum.

There are four fundamental ways of achieving wavelength conversion: (1) optoelectronic, (2) optical gating, (3) interferometric, and (4) wave mixing. The latter three approaches are all-optical but not yet mature enough for commercial use. Optoelectronic converters today offer substantially better performance at lower cost than comparable all-optical wavelength converters.

3.8.1 Optoelectronic Approach

This is perhaps the simplest, most obvious, and most practical method today to realize wavelength conversion. As shown in Figure 3.75, the input signal is first converted to electronic form, regenerated, and then retransmitted using a laser at a different wavelength. This is usually a variable-input, fixed-output converter. The receiver does not usually care about the input wavelength, as long as it is in the 1310 or 1550 nm window. The laser is usually a fixed-wavelength laser. A variable output can be obtained by using a tunable laser.

The performance and transparency of the converter depend on the type of regeneration used. Figure 3.75 shows the different types of regeneration possible. In the simplest case, the receiver simply converts the incoming photons to electrons, which get amplified by an analog RF (radio-frequency) amplifier and drive the laser. This is called 1R regeneration. This form of conversion is truly transparent to the modulation format (provided the appropriate receiver is used to receive the signal) and can handle analog data as well. However, noise is added at the converter, and the effects of nonlinearities and dispersion (see Chapter 5) are not reset.

Another alternative is to use regeneration with reshaping but without retiming, also called 2R regeneration. This is applicable only to digital data. The signal is reshaped by sending it through a logic gate, but not retimed. The additional phase jitter introduced because of this process will eventually limit the number of stages that can be cascaded.

The final alternative is to use regeneration with reshaping and retiming (3R). This completely resets the effects of nonlinearities, fiber dispersion, and amplifier noise; moreover, it introduces no additional noise. However, retiming is a bit-rate-specific function, and we lose transparency. If transparency is not very important, this is a very attractive approach. (Note that in Chapter 8 we will discuss another way of maintaining some transparency with 3R using the so-called digital wrapper.) These types of regenerators often include circuitry to perform performance monitoring and

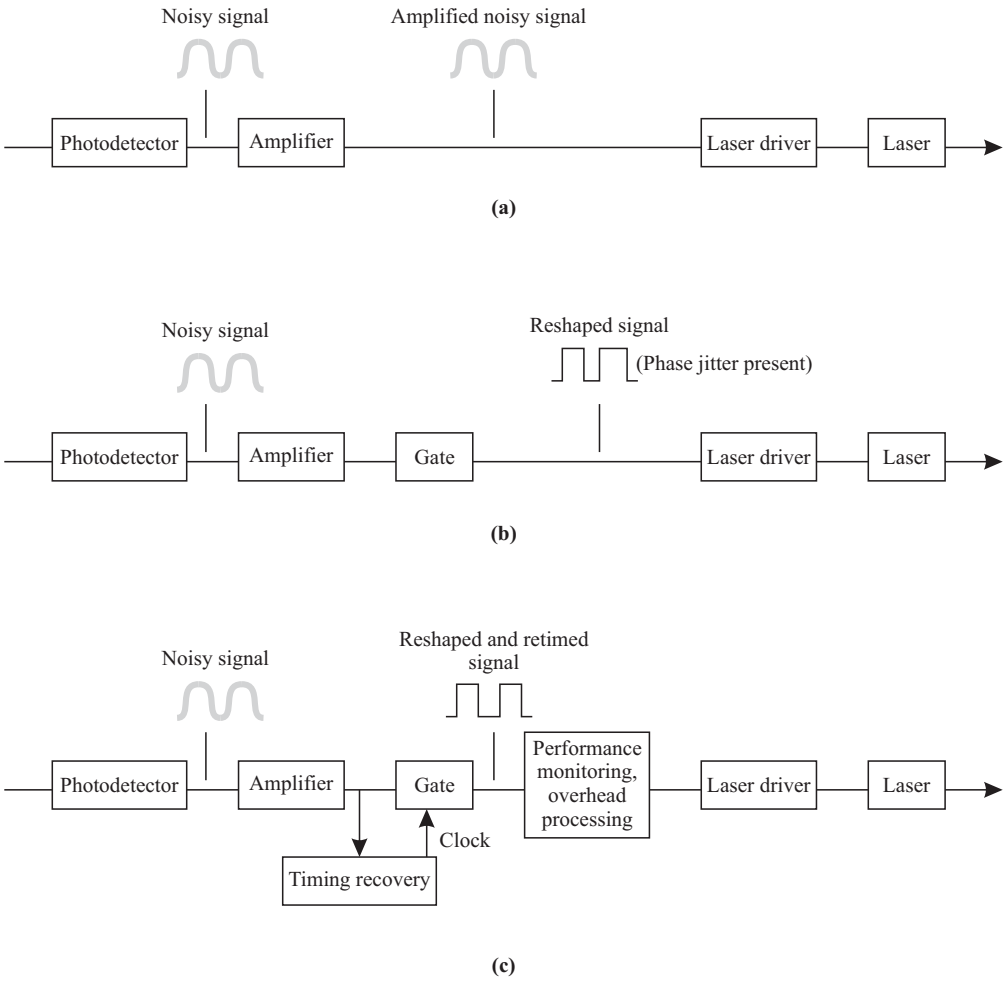

Figure 3.75 Different types of optoelectronic regeneration. (a) 1R (regeneration without reshaping or retiming. (b) 2R (regeneration with reshaping). (c) 3R (regeneration with reshaping and retiming).

process and modify associated management overheads associated with the signal. We will look at some of these overheads in Sections 6.1 and 8.5.7.

3.8.2 Optical Gating

Optical gating makes use of an optical device whose characteristics change with the intensity of an input signal. This change can be transferred to another unmodulated *probe* signal at a different wavelength going through the device. At the output, the probe signal contains the information that is on the input signal. Like the optoelectronic approach, these devices are variable-input and either fixed-output or variable-output devices, depending on whether the probe signal is fixed or tunable. The transparency offered by this approach is limited—only intensity-modulated signals can be converted.

The main technique using this principle is cross-gain modulation (CGM), using a nonlinear effect in a semiconductor optical amplifier (SOA). This approach works over a wide range of signal and probe wavelengths, as long as they are within the amplifier gain bandwidth, which is about 100 nm. Early SOAs were polarization sensitive, but by careful fabrication, it is possible to make them polarization insensitive. SOAs also add spontaneous emission noise to the signal.

CGM makes use of the dependence of the gain of an SOA on its input power, as shown in Figure 3.76. As the input power increases, the carriers in the gain region of the SOA get depleted, resulting in a reduction in the amplifier gain. What makes this interesting is that the carrier dynamics within the SOA are very fast, happening on a picosecond time scale. Thus the gain responds in tune with the fluctuations in input power on a bit-by-bit basis. The device can handle bit rates as high as 10 Gb/s. If a low-power probe signal at a different wavelength is sent into the SOA, it will experience a low gain when there is a 1 bit in the input signal and a higher gain when there is a 0 bit. This very same effect produces crosstalk when multiple signals at different wavelengths are amplified by a single SOA and makes the SOA relatively unsuitable for amplifying WDM signals.

The advantage of CGM is that it is conceptually simple. However, there are several drawbacks. The achievable extinction ratio is small (less than 10) since the gain does not really drop to zero when there is an input 1 bit. The input signal power must be high (around 0 dBm) so that the amplifier is saturated enough to produce a good variation in gain. This high-powered signal must be eliminated at the amplifier output by suitable filtering, unless the signal and probe are counterpropagating. Moreover, as the carrier density within the SOA varies, it changes the refractive index as well, which in turn affects the phase of the probe and creates a large amount of pulse distortion.

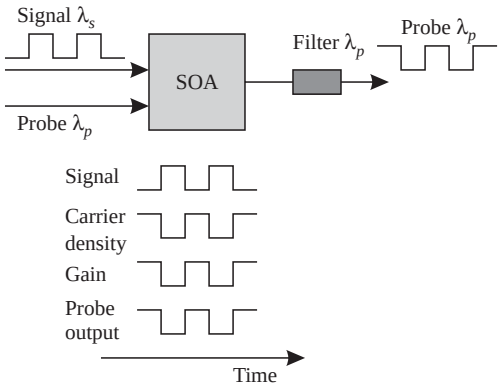

Figure 3.76 Wavelength conversion by cross-gain modulation in a semiconductor optical amplifier.

3.8.3 *Interferometric Techniques*

The same phase-change effect that creates pulse distortion in CGM can be used to effect wavelength conversion. As the carrier density in the amplifier varies with the input signal, it produces a change in the refractive index, which in turn modulates the phase of the probe. Hence we use the term *cross-phase* modulation for this approach. This phase modulation can be converted into intensity modulation by using an interferometer such as a Mach-Zehnder interferometer (MZI) (see Section 3.3.7). Figure 3.77 shows one possible configuration of a wavelength converter using cross-phase modulation. Both arms of the MZI have exactly the same length, with each arm incorporating an SOA. The signal is sent in at one end (A) and the probe at the other end (B). If no signal is present, then the probe signal comes out unmodulated. The couplers in the MZI are designed with an asymmetric coupling ratio $\gamma \neq 0.5$. When the signal is present, it induces a phase change in each amplifier. The phase change induced by each amplifier on the probe is different because different amounts of signal power are present in the two amplifiers. The MZI translates this relative phase difference between its two arms on the probe into an intensity-modulated signal at the output.

This approach has a few interesting properties. The natural state of the MZI (when no input signal is present) can be arranged to produce either destructive or constructive interference on the probe signal. Therefore we can have a choice of whether the data coming out is the same as the input data or is complementary.

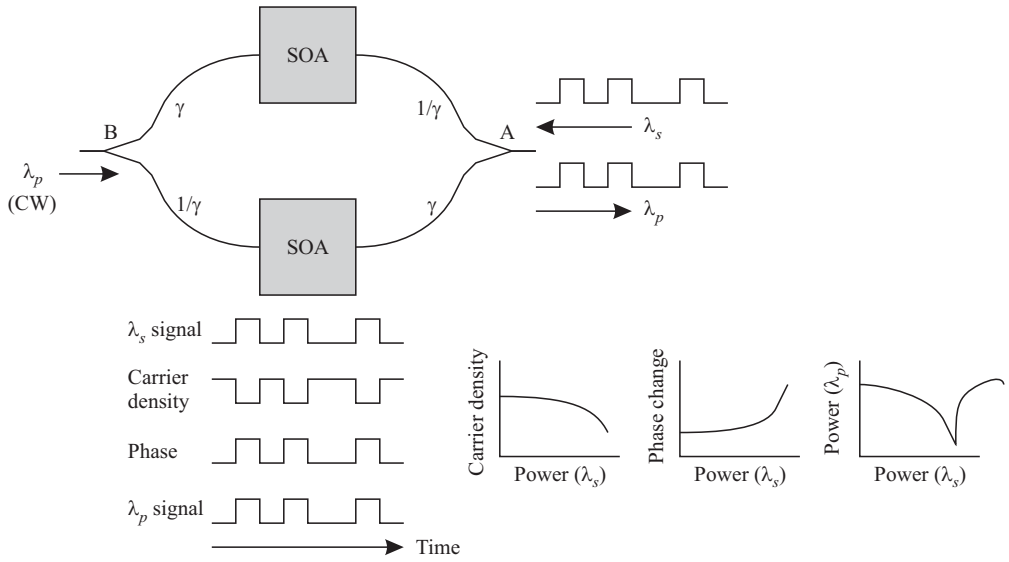

Figure 3.77 Wavelength conversion by cross-phase modulation using semiconductor optical amplifiers embedded inside a Mach-Zehnder interferometer.

The advantage of this approach over CGM is that much less signal power is required to achieve a large phase shift compared to a large gain shift. In fact, a low signal power and a high probe power can be used, making this method more attractive than CGM. This method also produces a better extinction ratio because the phase change can be converted into a "digital" amplitude-modulated output signal by the interferometer. So this device provides regeneration with reshaping (2R) of the pulses. Depending on where the MZI is operated, the probe can be modulated with the same polarity as the input signal, or the opposite polarity. Referring to Figure 3.77, where we plot the power coupled out at the probe wavelength versus the power at the signal wavelength, depending on the slope of the demultiplexer, a signal power increase can either decrease or increase the power coupled out at the probe wavelength. Like CGM, the bit rate that can be handled is at most 10 Gb/s and is limited by the carrier lifetime. This approach, however, requires very tight control of the bias current of the SOA, as small changes in the bias current produce refractive index changes that significantly affect the phase of signals passing through the device.

We have seen that the CPM interferometric approach provides regeneration with reshaping (2R) of the pulses. As we saw earlier, while 2R cleans up the signal shape,

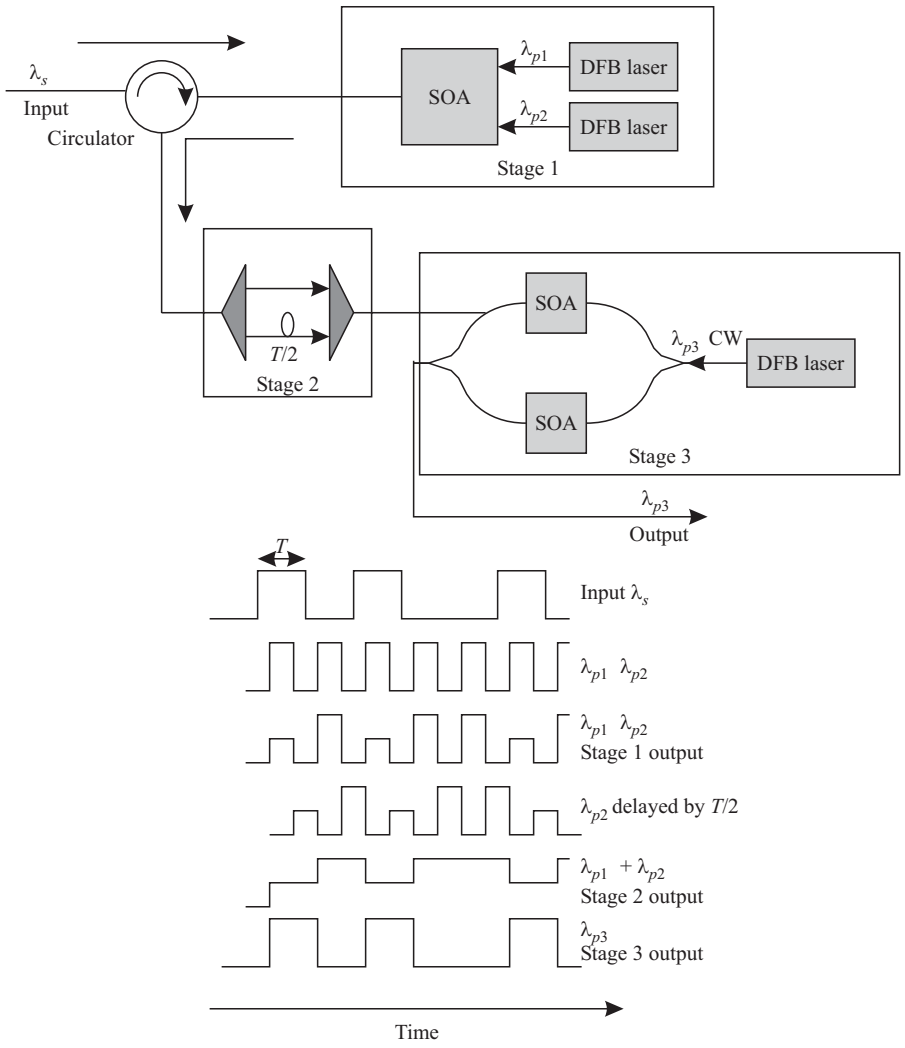

Figure 3.78 All-optical regeneration with reshaping and retiming (3R) using a combination of cross-gain modulation and cross-phase modulation in semiconductor optical amplifiers. (After [Chi97].)

it does not eliminate phase (or equivalently timing) jitter in the signal, which would accumulate with each such 2R stage. In order to completely clean up the signal, including its temporal characteristics, we need regeneration with reshaping and retiming (3R). Figure 3.78 shows one proposal for accomplishing this in the optical domain without resorting to electronic conversion [Chi97, Gui98]. The approach uses a combination of CGM and CPM. We assume that a local clock is available to sample the incoming data. This clock needs to be recovered from the data; we will study ways of doing this in Section 12.2. The regenerator consists of three stages. The first stage samples the signal. It makes use of CGM in an SOA. The incoming signal is probed using two separate signals at different wavelengths. The two probe signals are synchronized and modulated at twice the data rate of the incoming signal. Since the clock is available, the phase of the probe signals is adjusted to sample the input signal in the middle of the bit interval. At the output of the first stage, the two probe signals have reduced power levels when the input signal is present and higher power levels when the input signal is absent. In the second stage, one of the probe signals is delayed by half a bit period with respect to the other. At the output of this stage, the combined signal has a bit rate that matches the bit rate of the input signal and has been regenerated and retimed. This signal is then sent through a CPM-based interferometric converter stage, which then regenerates and reshapes the signal to create an output signal that has been regenerated, retimed, and reshaped.

3.8.4 Wave Mixing

The four-wave mixing phenomenon that occurs because of nonlinearities in the transmission medium (discussed in Section 2.5.8) can also be utilized to realize wavelength conversion. Recall that four-wave mixing causes three waves at frequencies f_1, f_2, and f_3 to produce a fourth wave at the frequency $f_1 + f_2 - f_3$; when $f_1 = f_2$, we get a wave at the frequency $2f_1 - f_3$. What is interesting about four-wave mixing is that the resulting waves can lie in the same band as the interacting waves. As we have seen in Section 2.5.8, in optical fibers, the generated four-wave mixing power is quite small but can lead to crosstalk if present (see Section 5.8.4).

For the purposes of wavelength conversion, the four-wave mixing power can be enhanced by using an SOA because of the higher intensities within the device. If we have a signal at frequency f_s and a probe at frequency f_p, then four-wave mixing will produce signals at frequencies $2f_p - f_s$ and $2f_s - f_p$, as long as all these frequencies lie within the amplifier bandwidth (Figure 3.79).

The main advantage of four-wave mixing is that it is truly transparent because the effect does not depend on the modulation format (since both amplitude and phase are preserved during the mixing process) and the bit rate. The disadvantages are that the other waves must be filtered out at the SOA output, and the conversion

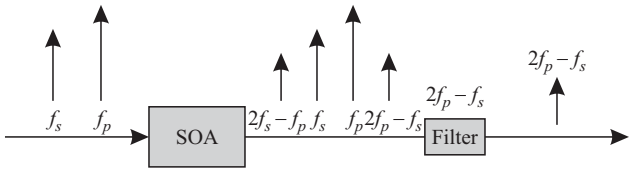

Figure 3.79 Wavelength conversion by four-wave mixing in a semiconductor optical amplifier.

efficiency goes down significantly as the wavelength separation between the signal and probe is increased. We will study the conversion efficiency of four-wave mixing in Section 5.8.4.

Summary

We have studied many different optical components in this chapter. Couplers, isolators, and circulators are all commodity components. Many of the optical filters that we studied are commercially available, with fiber gratings, thin-film multicavity filters, and arrayed waveguide gratings used in commercial WDM systems.

Erbium-doped fiber amplifiers (EDFAs) are widely deployed and indeed served as a key enabler for WDM. EDFA designs can incorporate multiple stages and gain-flattening filters and provide midstage access between the multiple stages to insert other elements such as dispersion compensating modules and wavelength add/drop multiplexers. Distributed Raman amplifiers are being used in conjunction with EDFAs in ultra-long-haul systems.

Semiconductor lasers are available commercially. Semiconductor DFB lasers are used in most high-speed communication systems as well as compact semiconductor tunable lasers. High-speed APDs and pinFET receivers are both available today.

There are a variety of technologies available to build switches. MEMS-based optical switches are suited for wavelength selective switches with moderate or large numbers of ports. For smaller-scale switches, most switch technologies can be applied. The switch technologies can be applied to other systems, for example, MEMS and liquid crystal technologies are used in variable optical attenuators.

All-optical wavelength converters are still in the research laboratories, awaiting significant cost reductions and performance improvements before they can become practical.