
Nature and Properties of Light

Linda J. Vandergriff
Director of Photonics System Engineering
Science Applications International Corporation
McLean, Virginia

Light plays a vital role in our daily lives. It is used in compact disc (CD) players, in which a laser reflecting off a CD transforms the returning signal into music. It is used in grocery store checkout lines, where laser beams read bar codes for prices. It is used by laser printers to record images on paper. It is used in digital cameras that capture our world and allow pictures to be displayed on the Internet. It is the basis of the technology that allows computers and telephones to be connected to one another over fiber-optic cables. And light is used in medicine, to produce images used in hospitals and in lasers that perform eye surgery.

The generation, transport, manipulation, detection, and use of light are at the heart of photonics. Photonics is a critical part of the future and a growing career field. In 1997 it was a \$50 billion market with a projected growth of 10 to 20 percent over the next decade. Photonics technicians and engineers must master new concepts, learn new techniques, and develop new skills. To work in photonics it is necessary to have a basic understanding of the nature of light and its properties.

Prerequisites

This module requires a basic understanding of high school algebra, trigonometry, general scientific nomenclature, the scientific process, units conversions, and basic concepts in elementary physics and chemistry.

Objectives

When you finish this module you will be able to:

- Describe the wide variety of opportunities for photonics technicians.
- Define and use basic radiometric units.
- Define the following properties of light:
 - Speed
 - Frequency
 - Wavelength
 - Energy
- Describe the dual nature of light, as a continuous wave and a discrete particle (photon), and give examples of light exhibiting both natures.
- Describe the six properties of electromagnetic waves and give everyday examples.
- Explain the mechanism that causes light to be polarized, explain the use of polarizing material, and give an example of the use of polarizers.
- Describe Huygens' principle and the superposition principle.
- Define the terms *reflection*, *refraction*, and *index of refraction* and explain how they are related.
- Explain diffraction and interference in terms of Huygens' principle.
- List the three types of emission and identify the material properties that control the emission type.
- Describe in a short paragraph the electromagnetic spectrum and sketch a diagram of the key optical regions and uses.
- Give a basic explanation of atoms and molecules and their ability to absorb, store, and emit quanta of energy.
- Define the primary equations describing the relationships between temperature of, wavelength of, and energy emitted by a blackbody and a graybody.
- Describe the mechanisms that affect light propagating in a medium and its transmission.

True Life Scenario

Holly works as a photonics assembly technician. The factory where she is employed creates laser diode assemblies for a variety of commercial uses. As an assembly technician, she creates and aligns laser diode components and ensures that the bonded products meet the tight quality standard she and her company require of all their products.

At the beginning of the shift, Holly gets her assignment for the production run. She prepares for entering the clean room where the work will take place. Then, after entering, she logs on at her assembly station. She selects the correct workspace file for the devices to be manufactured during that shift. The workspace file selection is based on wavelength, other optical properties, and pass/fail criteria set for this particular device.

With the components provided, Holly first sets up the lateral shearing interferometer and the microlens/laser diode product. Then she collimates and directs the laser diode beam into the beam analysis tool for testing.

After that she reviews the results on the computer monitor and adjusts the alignment until the device is acceptable. Finally, she bonds the microlens and laser diode and stores the final measurements to a log file on the network server. These measurements, along with those of the other devices created during this production run, will serve as a statistical base for evaluation of the production process and the product quality.

Holly will create a batch of microlens/laser diode pairs during her shift. Depending on the order and application, this run may require days or weeks. Then a new product will be designed for production, and Holly and her counterparts on the production floor will create it. This will require an evolving understanding of light and its uses to allow flexibility in the manufacturing process and keep Holly and her company competitive in the growing marketplace.



Figure 1-1 Photonics assembly technician assembling LEDs¹

Opening Demonstration

Note: The interactive exercise that follows is to be used as a short introduction to the wide range of photonics applications. It is intended to stimulate interest in the study of photonics.

Instructions: Create two or three groups. Have each group, with notes and manuals closed and without repeating, name a use of light technology. When no one in a group can think of a use that has not already been added to the list, that group drops out. Compare the groups' lists with the one following. Did the groups miss any areas? Are there any areas that should be on the following list but are not?

¹ Source: *Laser Focus World*, May 1999, 291. Reprinted by permission.

Photonics in Our Daily Lives

Home

- Energy-saving fluorescent lamps
- Infrared remote controls
- TV flat panel / large screen
- Optical fibers for cable TV
- Compact disc players
- IR motion sensors for home security
- Video disk players
- Alarm clock radio with LED display
- IR noncontact “ear” thermometers
- Infrared remote headphones

Office

- Optical scanners
- Fax machines
- Optical fiber telephone cables
- Optical data storage
- Laser printers
- Photocopiers
- Overhead slide projectors
- Video teleconferences
- Laser pointers
- Computer active matrix displays
- Computer displays
- Infrared remote connections
- Special optical computers

Car

- Infrared security systems
- Optical monitors for antilock brakes
- Optical fiber dashboard displays
- LED traffic signals
- Laser traffic radar
- Solar-powered emergency services

Store

- Supermarket bar-code scanners
- Credit card holograms

Medical

- Laser surgery
- Medical diagnosis tools
- Microscopes

Manufacturing

- Laser welding and cutting
- Optical stereo-lithography
- Machine vision
- Image recognition for quality control
- Nondestructive testing
- Precision measurement
- Optical inspection of labeling and packaging
- Laser fabric cutting machines

Other

- Laser light shows
- Digital cameras
- Night vision goggles
- Missile guidance
- Laser weapons
- Surveillance cameras
- Surveying—alignment and range finders
- Computer-generated optical elements
- Art gallery holography exhibits

Basic Concepts

I. NATURE AND PROPERTIES OF LIGHT

A. Introduction to Photonics

Photonics is defined as the generation, manipulation, transport, detection, and use of light information and energy whose quantum unit is the photon.

Photonics is based on the science of optics and electronics. The origins of optical technology (photonics) date back to the remote past. Exodus 38:8 (ca 1200 BCE) tells of “the looking glasses of the women.” In the coming century, photonics promises to be a vital part of the information revolution.

To enable us to understand and apply photonics, it is necessary to have a basic understanding of the behavior and properties of light. This course focuses on these fundamentals of photonics and prepares you for an exciting future as a photonics technician.

B. Photonics Opportunities

There are ten broad areas of employment that are likely to need increasing numbers of photonics technicians:

- Medicine-biomedical
- Environmental
- Energy
- Transportation
- Defense
- Public safety
- Aerospace
- Computers
- Manufacturing with photonics and test and analysis
- Communication and information technology

Medicine has seen significant growth in the use of photonics devices in laser surgery and in noninvasive diagnostic tools. This growth translates into ever-growing opportunities for biomedical photonics technicians.

On the environmental front, photonics devices can measure the pollutants in our air and water remotely. Photonics devices can harness renewable energy from the sun, augment other energy sources, and preserve our natural resources. Transportation will be undergoing significant changes, such as the introduction of the Intelligent Vehicle Highway System, which provides guidance, collision avoidance, and continuous tuning of engines based on driving conditions.

The defense industry and public safety agencies need the ability to see and understand the environment, whether it is an enemy’s movement or a tornado’s path. The defense industry, in addition, has identified several photonics devices that can neutralize enemy weapons. These fields will grow and will provide significant potential for technicians who work in the areas of remote sensing, image processing, and high-energy laser operation.

In the Information Age, photonics will be essential in gathering, manipulating, storing, routing, and displaying information. New optical computers are proposed for some functions, and charged coupled devices (CCD) cameras digitize artwork. Compact discs, digital video devices (DVD), and other media are used for data storage and retrieval using lasers. The links between nodes of the Internet or telephone lines make use of fiber optics. Data can be printed with laser printers or displayed on plasma panels. This area of photonics application is growing at an incredible rate, and the potential for technicians who work in this area is very high.

Automation of manufacturing relies heavily on photonics. Fabrication is performed mainly by industrial lasers that cut, weld, trim, drill holes, and heat-treat products. To ensure product quality, inspection is performed using spectroscopy, interferometry, machine vision, and image processing. As manufacturing becomes more sophisticated in its use of photonics, the demand for skilled photonics technicians is expected to grow explosively.

C. Properties of Light

What is light? This question has been debated for many centuries. The sun radiates light, electric lights brighten our darkness, and many other uses of light impact our lives daily. The answer, in short, is light is a special kind of electromagnetic energy.

The speed of light, although quite fast, is not infinite. The speed of light in a vacuum is expressed as $c = 2.99 \times 10^8$ m/s. Light travels in a vacuum at a constant speed, and this speed is considered a universal constant. It is important to note that speed changes for light traveling through nonvacuum media such as air (0.03% slower) or glass (30.0% slower).

For most purposes, we may represent light in terms of its magnitude and direction. In a vacuum, light will travel in a straight line at fixed speed, carrying energy from one place to another. Two key properties of light interacting with a medium are:

1. It can be deflected upon passing from one medium to another (refraction).
2. It can be bounced off a surface (reflection).

The aspects of light interaction with media other than a vacuum will be addressed further in Modules 1.3 and 1.4, which deal with geometrical and physical optics, respectively.

The field of detection and measurement of light energy is called *radiometry*. It uses a standardized system for characterizing radiant energy. Table 1-1 defines the standard terms used in this course.

Table 1-1: Radiometric Definition and Units

Term	Definition	Symbol	Units
Quantity	Radiant energy	Q	Joule (J)
Flux	Rate of radiant energy	Φ	Watt (W); Joule/second (J/s)
Flux density (irradiance)	Flux per unit area	E	Watts per square meter (W/m^2)
Intensity	Flux per solid angle	I	Watts per steradian (W/sr)
Radiance	Flux per unit area per unit solid angle	L	Watts per square meter per steradian ($\text{W/m}^2 \cdot \text{sr}$)
Spectral radiance	Radiance per unit wavelength	L_λ	Watts per square meter per steradian per nanometer $\frac{\text{W}}{\text{m}^2 \cdot \text{sr} \cdot \Delta\lambda}$

Dual Nature of Light

Scientists build models of physical processes to help them understand and predict behavior. So it is with light energy. It is through seeing the effects of light that the models are developed. Scientists have observed that light energy can behave like a wave as it moves through space, or it can behave like a discrete particle with a discrete amount of energy (quantum) that can be absorbed and emitted. As we study and use light, both models are helpful.

Concept of a photon

The particle-like nature of light is modeled with photons. A photon has no mass and no charge. It is a carrier of electromagnetic energy and interacts with other discrete particles (e.g., electrons, atoms, and molecules).

A beam of light is modeled as a stream of photons, each carrying a well-defined energy that is dependent upon the wavelength of the light. The energy of a given photon can be calculated by:

$$\text{Photon energy } (E) = hc/\lambda \quad (1-1)$$

where E is in joules

$$h = \text{Planck's constant} = 6.625 \times 10^{-34} \text{ J}\cdot\text{s}$$

$$c = \text{Speed of light} = 2.998 \times 10^8 \text{ m/s}$$

$$\lambda = \text{Wavelength of the light in meters}$$

Example 1-1

Photons in a pale blue light have a wavelength of 500 nm. (The symbol nm is defined as a nanometer = 10^{-9} m.) What is the energy of this photon?

Solution:

$$\begin{aligned}
 E &= hc / \lambda = [6.625 \times 10^{-34} \text{ J}\cdot\text{s} \times 2.998 \times 10^8 \text{ m/s}] / 500 \times 10^{-9} \text{ m} \\
 &= \frac{6.625 \times 10^{-34} \times 2.998 \times 10^8}{500 \times 10^{-9}} \\
 &= 3.97 \times 10^{-19} \text{ J}
 \end{aligned}$$

When ultraviolet light shines on some metal surfaces, it causes electrons to be emitted. This effect is shown in Figure 1-2. The *photoelectric effect* did not produce results that matched the early predictions of wave theory. Two concerns were:

1. More intense radiation (larger-amplitude waves) did not cause emitted electrons to have more energy.
2. The energy of the emitted electron was dependent on the wavelength of the light, not the amplitude of the wave.

In the photoelectric effect experiment shown in Figure 1-2, light strikes a metal plate. Electrons are immediately released. The flow of electricity in the external circuit can be measured and the number of electrons generated for a given light signal can be determined.

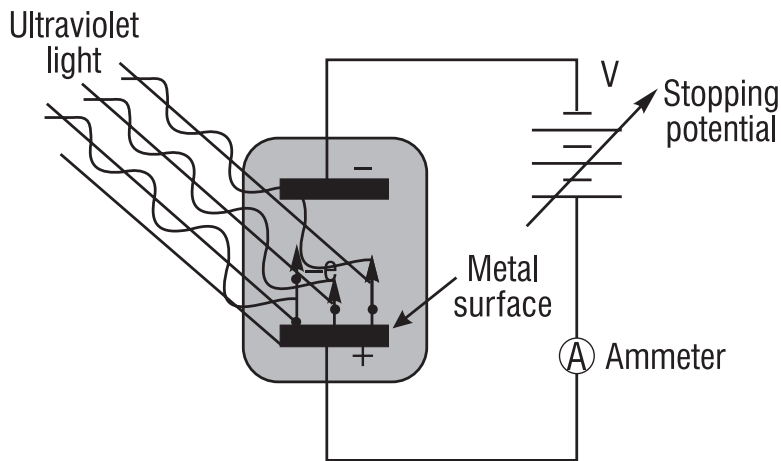


Figure 1-2 Photoelectric effect experiment

If light were a continuous wave, it might wash over the metal surface and interact with the electrons to give them the needed energy to escape at lower light levels (intensities), but only after long delays. However, faint light at high frequencies (short wavelengths) caused the immediate release of electrons. Thus, light knocked the electrons out of the metal surface as if the light were made of particles—*photons*.

There is a minimum energy threshold for an electron to escape from the metal. Photons with frequencies below a given threshold eject no electrons, no matter how intense the light. Photons with frequencies above the threshold do eject electrons, no matter how low the intensity. The energy of the released electrons can be calculated from Equation 1-2:

$$E_{e-} = hc/\lambda - p \quad (1-2)$$

where: p = characteristic escape energy for the metal
 E_{e-} = the kinetic energy of an escaping electron
 hc/λ = the energy of the photon of wavelength λ

Example 1-2

We can calculate the threshold wavelength of light needed to just release electrons from gold. This corresponds to E_{e-} equal to zero. Solve Equation 1-2 for λ .

Solution:

Let $hc/\lambda = p$, so that

$$hc/p = \lambda$$

The escape energy for gold is $p_{\text{gold}} = 7.68 \times 10^{-19} \text{ J}$

$$\begin{aligned} \lambda &= \frac{6.625 \times 10^{-34} \text{ J} \cdot \text{s} \times 2.998 \times 10^8 \text{ m/s}}{7.68 \times 10^{-19} \text{ J}} \\ &= 2.59 \times 10^{-7} \text{ m or } 0.259 \mu\text{m} \end{aligned}$$

The photon model, although quite useful in explaining some properties of light, is still closely related to the wave model discussed below.

Wave Model

The particle-like model of light describes large-scale effects such as light passing through lenses or bouncing off mirrors (dealt with in Module 1-3, *Basic Geometrical Optics*). However, a wavelike model must be used to describe fine-scale effects such as interference and diffraction that occur when light passes through small openings or by sharp edges (dealt with in Module 1-4, *Basic Physical Optics*). The propagation of light or electromagnetic energy through space can be described in terms of a traveling wave motion. The wave moves energy—without moving mass—from one place to another at a speed independent of its intensity or wavelength.

This wave nature of light is the basis of physical optics and describes the interaction of light with media. Many of these processes require calculus and quantum theory to describe them rigorously. For this text it is sufficient to provide the resulting equations and models to be used by the photonics technician in real applications.

Characteristics of light waves

To understand light waves, it is important to understand basic wave motion itself. Water waves are sequences of crests (high points) and troughs (low points) that “move” along the surface of the water. When ocean waves roll in toward the beach, the line of crests and troughs is seen as profiles parallel to the beach. An electromagnetic wave is made of an electric field and a magnetic field that alternately get weaker and stronger. The directions of the fields are at right

angles to the direction the wave is moving, just as the motion of the water is up and down while a water wave moves horizontally. Figure 1-3 is a one-dimensional representation of the electric field.

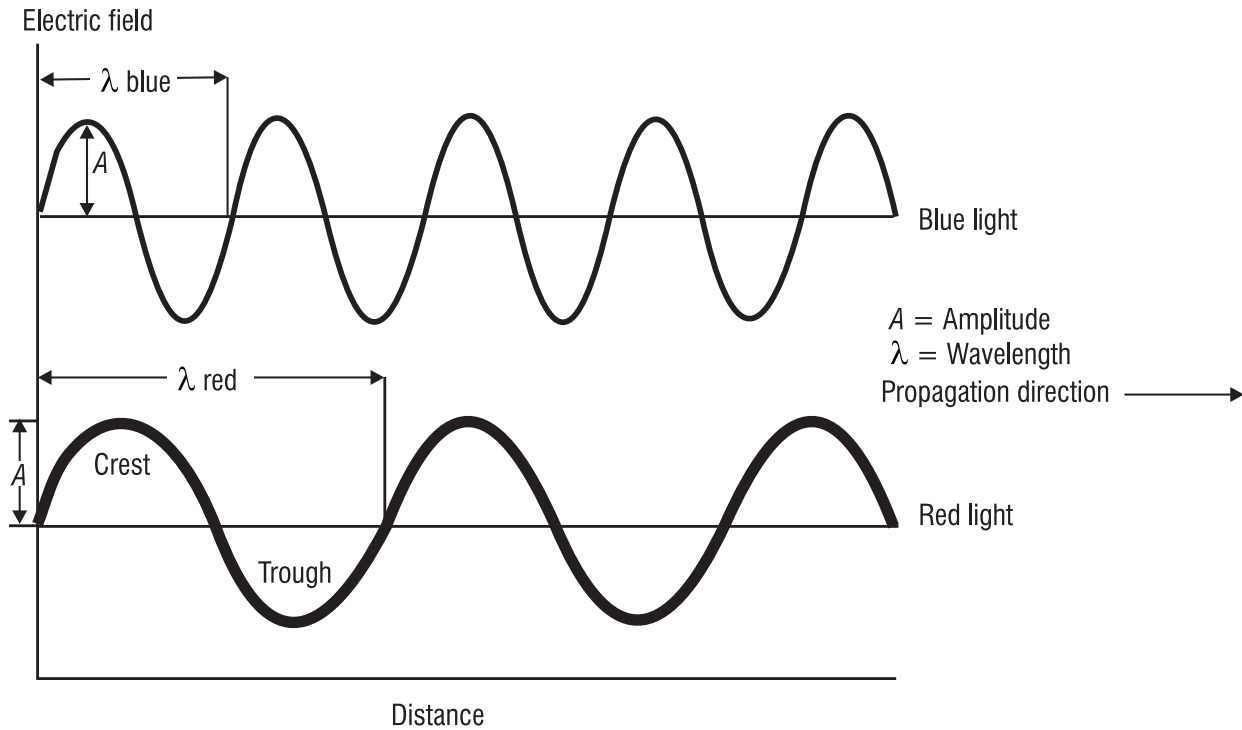


Figure1-3 One-dimensional representation of the electromagnetic wave

The maximum value of the wave displacement is called the **amplitude** (A) of the wave. The cycle starts at zero and repeats after a distance. This distance is called the **wavelength** (λ). Light can have different wavelengths, such as the blue light and red light shown in Figure 1-3. The inverse of the wavelength ($1/\lambda$) is the **wave number** (ν), which is expressed in cm^{-1} . The wave propagates at a **wave speed** (v). This wave speed in a vacuum is equal to c , and is less than c in a medium. At a stationary point along the wave, the wave passes by in a repeating cycle. The time to complete one cycle is called the **cycle time or period** (τ) and can be calculated using Equation 1-3.

$$\tau = \lambda/v$$

(1-3)

Another important measure of a wave is its **frequency** (f). It is measured as the number of waves that pass a given point in one second. The unit for frequency is cycles per second, also called hertz (Hz). As you can see, the frequency and the period are reciprocals of one another. If the wave speed and wavelength are known, the frequency can be calculated with Equation 1-4.

$$f = 1/\tau = v/\lambda$$

(1-4)

Example 1-3

For blue light in a vacuum, we can calculate the cycle time and frequency. From a previous example, we know that the wavelength of blue light is 500 nm and the velocity of light in a vacuum is c . Plugging in the numbers in Equation 1-3 we get:

$$\tau = \lambda/v = \frac{500 \times 10^{-9} \text{ m}}{2.998 \times 10^8 \text{ m/s}} = 1.667 \times 10^{-15} \text{ s}$$

Then we can calculate the frequency using Equation 1-4.

$$f = 1/\tau = 1/1.667 \times 10^{-15} \text{ s} = 5.996 \times 10^{14} \text{ Hz}$$

It is possible for a wave to have other than sinusoidal shapes; however, the important concept to remember is that light waves are transverse electric and magnetic fields changing in space and time and propagating at the speed of light in a given medium, as we show below.

Concept of light waves—Oscillating electric and magnetic fields

Light waves are complex. They are not one-dimensional waves but rather are composed of mutually perpendicular electric and magnetic fields with wave motion at right angles to both fields, as illustrated in Figure 1-4. The wave carries light energy with it. The amount of energy that flows per second across a unit area perpendicular to the direction of travel is called the irradiance (flux density) of the wave.

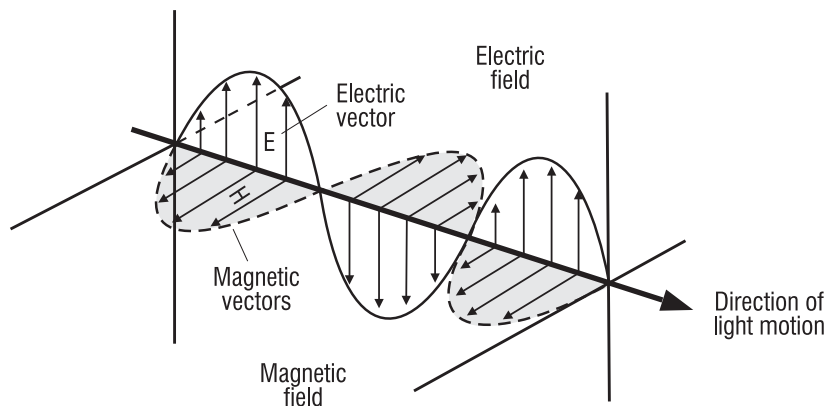


Figure 1-4 *Electric and magnetic fields in a light wave*

Electromagnetic waves share six properties with all forms of wave motion:

- Polarization
- Superposition
- Reflection
- Refraction
- Diffraction
- Interference

Polarization

Up to this point we have discussed the direction of light's propagation and its associated electric and magnetic fields. Polarization arises from the direction of the E-field vector with respect to the direction of the light's propagation. Since a light wave's electric field vibrates in a direction perpendicular to its propagation motion, it is called a *transverse* wave and is polarizable.

A sound wave, by contrast, vibrates back and forth along its propagation direction and thus is not polarizable.

Light is *unpolarized* if it is composed of vibrations in many different directions, with no preferred orientation. See Figure 1-5(a). Many light sources (e.g., incandescent bulbs, arc lamps, the sun) produce unpolarized light. Vertically polarized light is shown in Figure 1-5(b) and horizontally polarized light in Figure 1-5(c). Each is an example of *linearly* polarized light.

Figure 1-5(d) shows linearly polarized light making an angle of θ with the vertical. In this case, the tilted E-vector can be described by its components, E_x and E_y .

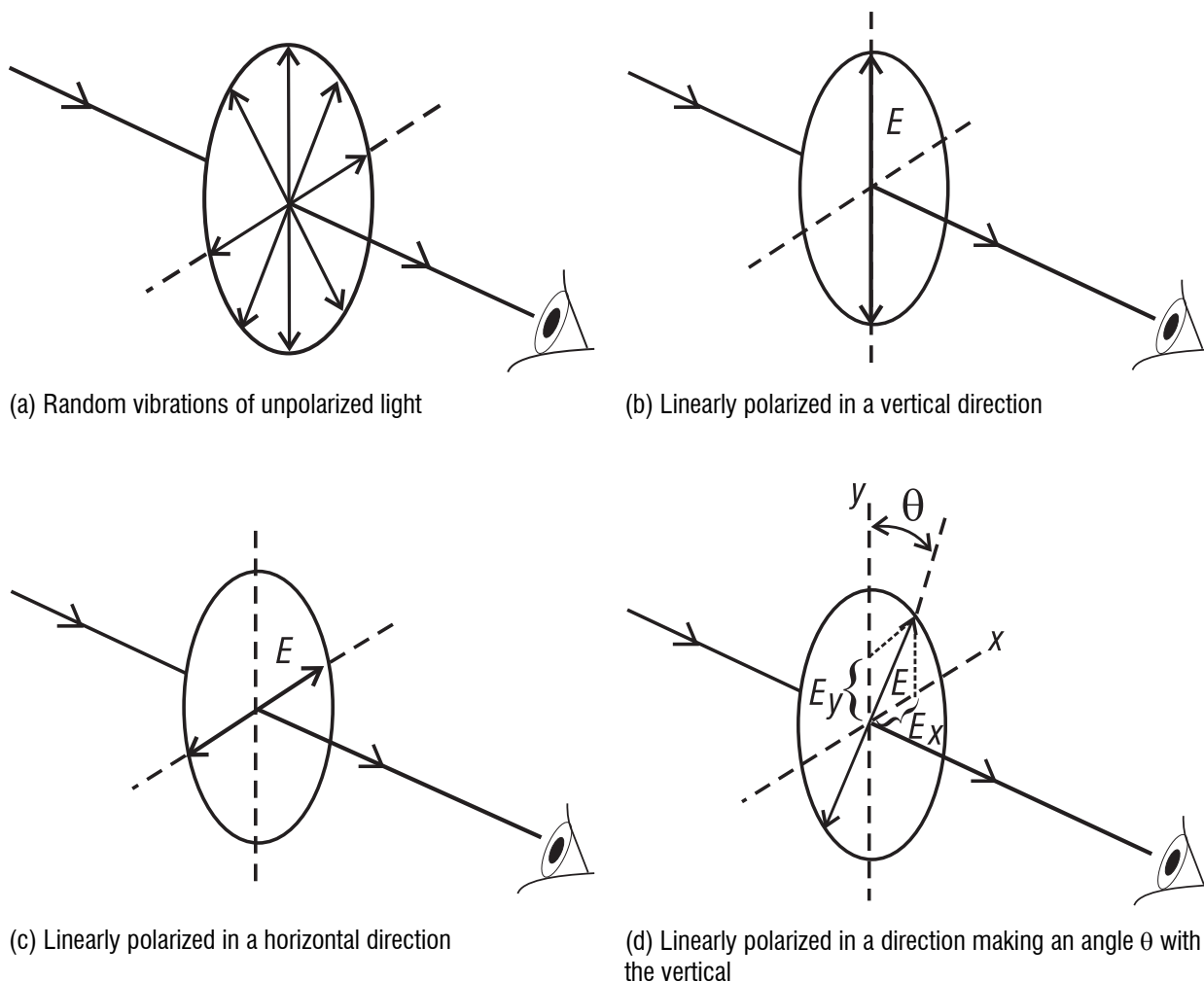


Figure 1-5 Unpolarized and linearly polarized light

When it happens, as in some cases, that E_x and E_y are not in the same phase—that is, they do not reach their maxima and minima at the same time—the E -field does not remain oriented in a fixed, linear direction. Rather, the amplitude maxima of the two components do not occur at the same time and so-called *elliptically* polarized light is exhibited. This means that, over time, light exhibits differing polarization orientations. A special case of elliptical polarization—called *circular* polarization—occurs when E_x equals E_y and they are out of phase by 90° .

Certain materials will transmit only selected polarizations. They are called polarizers—or analyzers—and have many uses. With randomly polarized light, a polarizer will pass light of one polarization and absorb or reflect other polarizations. A common example of the use of polarization in our daily life is found in polarizing sunglasses. The material in the lenses passes light whose electric field vibrations are perpendicular to certain molecular alignments and absorbs light whose electric field vibrations are parallel to the molecular alignments. The major component of light reflecting from a surface, such as a lake or car hood, is horizontally polarized, parallel to the surface. Thus, polarization in sunglasses, with the *transmission axis* in a vertical direction, rejects horizontally polarized light and therefore reduces glare. However, if you consider a sunbather lying on his or her side, wearing such sunglasses, the usual vertical polarization (transmission axis) will now be at 90° and parallel to the surface and will therefore pass the horizontally polarized light reflected off the water or the land.

The intensity of light passing through a linear polarizer can be calculated using Equation 1-5.

$$I(\theta) = I_0 \cos^2 (\theta) \quad (1-5)$$

where $I(\theta)$ is the light intensity passed by the polarizer

I_0 is the incident light intensity.

The angle of the E -field with respect to the transmission axis is defined as θ .

Example 1-4

(a) Given horizontally polarized light, what would be the ratio of the light intensity output to the light intensity input for $\theta = 0^\circ$, 45° , and 90° ?

Solution: Use Equation 1-5 to solve for $I(\theta)/I_0$ and plug in the numbers.

$$I(\theta)/I_0 = \cos^2 (\theta)$$

$$I(0)/I_0 = \cos^2 (0) = 1$$

$$I(45)/I_0 = \cos^2 (45) = 0.5$$

$$I(90)/I_0 = \cos^2 (90) = 0$$

(b) Given two polarizers and incident vertically polarized light, what is the ratio of the resultant light intensity to the incident light intensity if the polarizers' transmission axes are both vertical and parallel? What is the ratio if the axes are crossed, that is, one vertical and one horizontal?

Solution: First, for the parallel polarizers, calculate the $I(\theta)/I_0$ for the first polarizer assuming θ is 0. Then take the ratio of the two and repeat for the second polarizer. The resulting ratio is 1. Now, for the perpendicular polarizers, calculate $I(\theta)/I_0$ for the first polarizer, assuming θ is 0. Then take

the ratio of the two and repeat for the second polarizer, this time assuming that θ is 90. The resulting value is 0, as should be expected from crossed polarizers.

Huygens' Principle

In the seventeenth century, Christian Huygens proposed a principle that can be used to predict where a given wave front will be at any time in the future if you know the current location. His principle assumes that each point along a wave front can be considered a point source for production of secondary spherical wavelets. After a period of time, the new position of the wave front will be the surface tangent to these secondary wavelets. Huygens' principle is illustrated in Figure 1-6, for five point sources on a wave front.

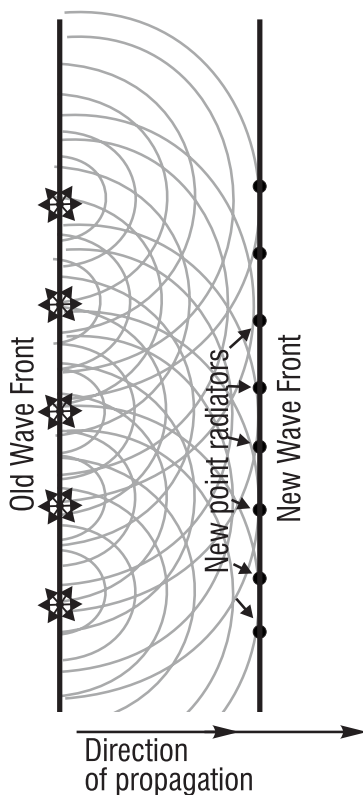
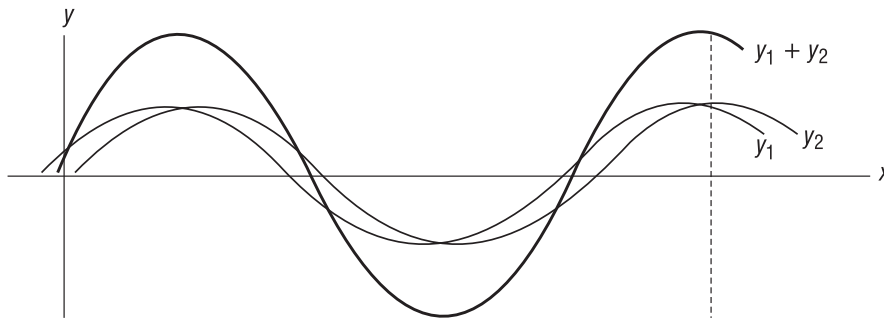


Figure 1-6 Using Huygens' principle to establish new wave fronts

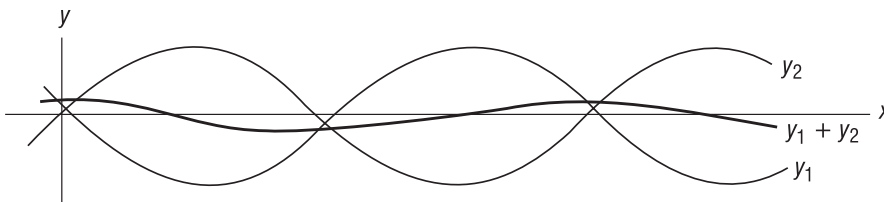
Superposition

For many kinds of waves, including electromagnetic, two or more waves can traverse the same space at the same time *independently* of one another. This means that the electric field at any point in space is simply the vector sum of the electric fields that the individual waves alone produce at the point. This is the *superposition principle*. Both the electric and magnetic fields of an electromagnetic wave satisfy the superposition principle. Thus, given multiple waves, the field at any given point can be calculated by summing each of the individual wave vectors.

When two or more waves are superimposed, the resulting physical effect is called interference. Suppose two waves, y_1 and y_2 , have nearly the same wavelength and phase (i.e., the maxima occur at nearly the same time and place). Superposition of these waves results in a wave ($y_1 + y_2$) of almost twice the amplitude of the individual waves. See Figure 1-7a. This is called constructive interference. If the maximum of one wave is near the minimum of the other wave, the resultant ($y_1 + y_2$) has almost no amplitude, as shown in Figure 1-7b. This is called destructive interference.



(a) Mostly constructive interference

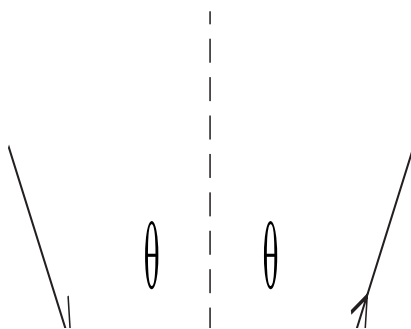


(b) Mostly destructive interference

Figure 1-7 Using the principle of superposition to add individual waves

Reflection

When a ray of light reflects off a surface (such as a mirror), its new direction depends on only the angle of incidence. The law of reflection states that the angle of incidence on a reflecting basic surface is equal to the angle of reflection. This is discussed in further detail in Module 1-3, *Basic Geometrical Optics*.



Law of reflection: Angle of incidence = Angle of reflection

Refraction

When a ray of light passes from one medium to another, it changes direction (bends) at the interface because of the difference in speed of the wave in the media. The ratio of this speed difference is called the **index of refraction** (n). The ratio of the indices of refraction and the direction of the two rays of light for the two media are expressed in Snell's law as shown in Figure 1-8 and Equation 1-6.

$$\frac{n_2}{n_1} = \frac{\sin \theta}{\sin \phi} \quad (1-6)$$

where n_1 and n_2 are the indices of refraction for the two media

θ is the angle of incidence

ϕ is the angle of refraction.

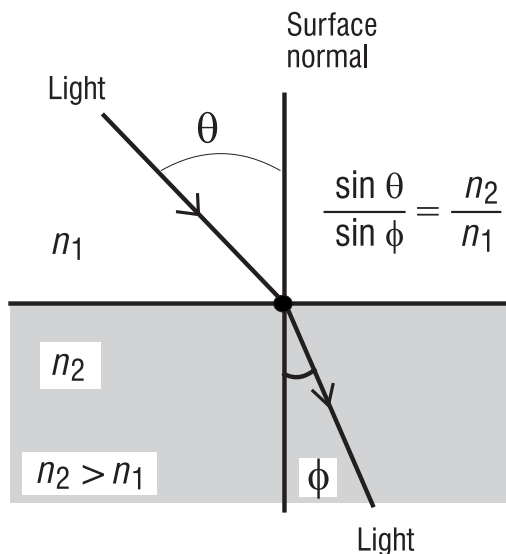


Figure 1-8 Refraction and Snell's law

Diffraction

Conclusive evidence of the correctness of a wave model came with the explanation of observed diffraction and interference. When light passes an obstacle, the shadow is not precise and sharp as geometrical ray theory would predict, but rather diffracted a little into the dark region behind the obstacle, thus giving the shadow a fuzzy edge. This property of light that causes it to spread out as it travels by sharp edges or through tiny holes can be explained by light having wavelike properties. Diffraction is predicted from Huygens' principle. In Figure 1-9, a wave is incident on a barrier from the left. The barrier has a slit. Every point on the incident wave front that arrives at the slit can be viewed as the site of an expanding spherical wavelet. For apertures that are small compared to the wavelength, the aperture becomes like a source and spherical waves result. As the slit width d increases, the diffracted wave becomes more and more like the incident plane wave except for the edges at the shadow.

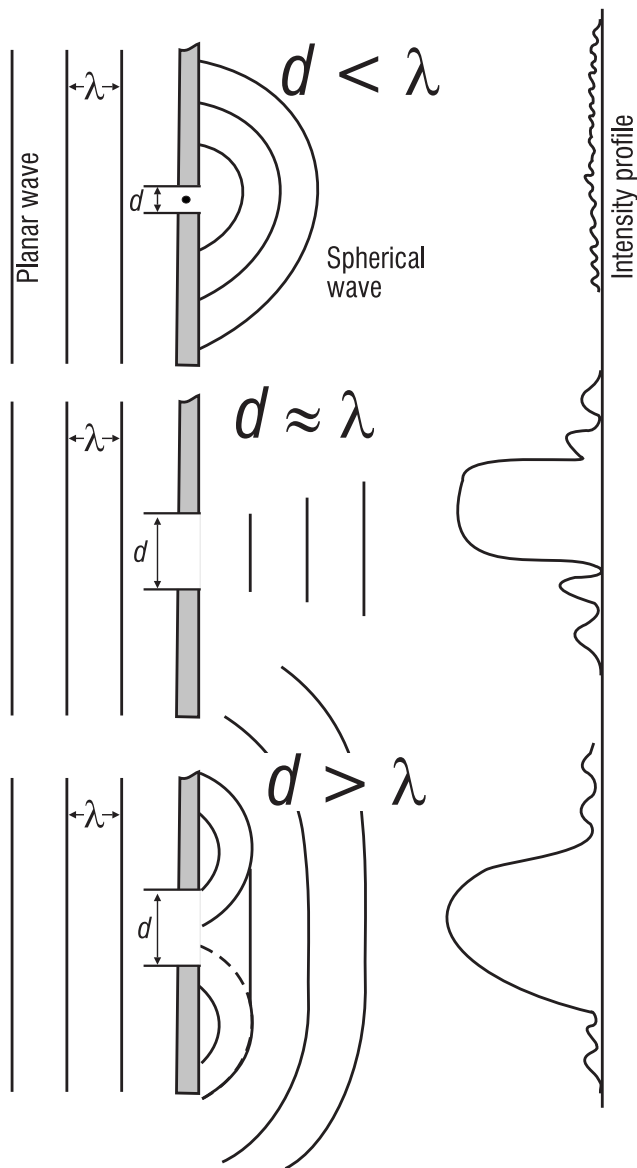


Figure 1-9 Diffraction of waves through slits of differing size

Interference

The first definitive demonstration of the wavelike nature of light was the classical two-slit experiment performed by Thomas Young in 1801. The two slits are very small compared to their separation distance. Thus, each slit produces diffracted spherical waves that overlap as they expand into the space to the right of the barrier. When they overlap, they interfere with each other, producing regions of mutually reinforcing waves. These appear on the screen as regions of maximum intensity. Between adjacent maxima is a region of minimum intensity. See Figure 1-10. The resulting pattern on the screen shows where constructive interference occurs (maxima, labeled B) and where destructive interference occurs (minima, labeled D). The experimental layout shown in Figure 1-10 can be used in practice to measure the wavelength of light. This experiment is covered with more rigor in Module 1-4, *Basic Physical Optics*.

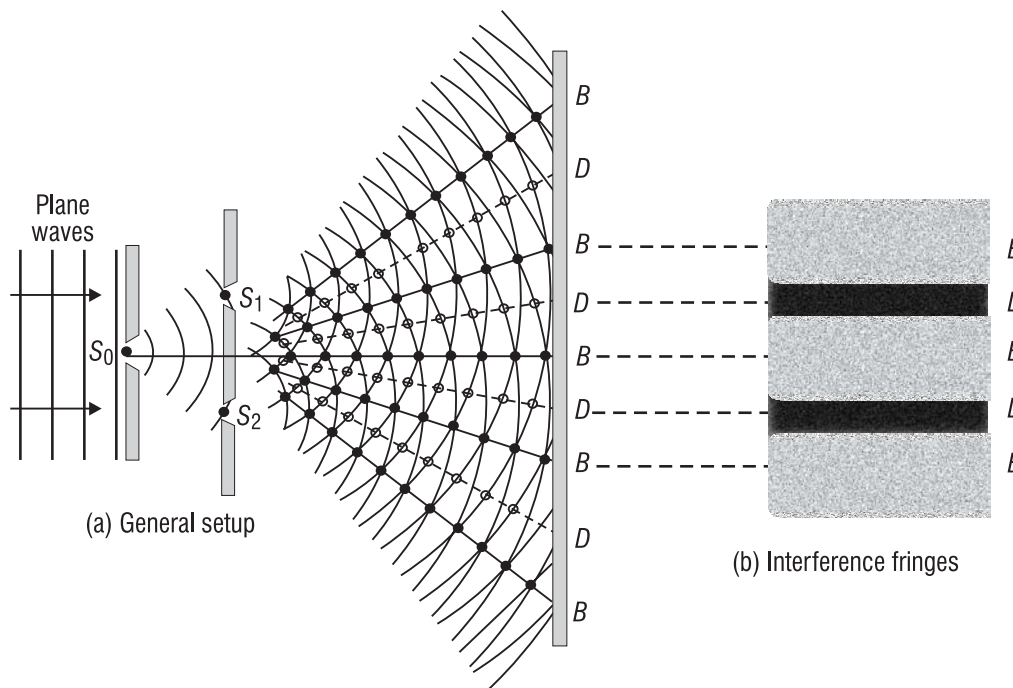


Figure 1-10 *Classic double-slit experiment*

The Electromagnetic Spectrum

All electromagnetic radiation has similar wavelike properties differing only in wavelength. Electromagnetic waves range in wavelength from very long (e.g., electric power line radiation at 60 Hz) to very short (e.g., gamma ray radiation). This entire range is called the **electromagnetic spectrum**. The spectrum shown in Figure 1-11 is divided by the practical applications for given ranges of frequencies that are set through convention by the sources and detection devices.

Of primary interest to photonics is the region from infrared to ultraviolet. However, each regime has some utility. Rotating generators and power lines generate low-frequency waves. These wavelengths are on the order of 10^5 to 10^8 meters. Heinrich Hertz produced radio waves in a very useful region of wavelengths ranging from 0.3 to 10^5 meters. Television and radio broadcasting bands are found in lower wavelengths. The microwave regime ranges from 0.01 to 0.3 meter and provides the radar and satellite communication bands. The infrared region, from 1 μm to 30 μm , was first detected by Sir William Herschel in 1800. This region is subdivided into five regions: very near (1–3 μm), near (3–5 μm), mid (5–6 μm), far (6–15 μm), and very long (15–30 μm) infrared. Just as the ear cannot hear above or below certain frequencies, the human eye cannot detect light outside a small range of wavelengths (0.76–0.49 μm). The ultraviolet region is a higher-energy region discovered by Johann Ritter. It triggers many chemical reactions and is what ionizes the upper atmosphere, creating the ionosphere. Wilhelm Röntgen discovered the X-ray regime in 1895. Its wavelength ranges from 10^{-8} to 10^{-11} meters. With its high energy, it can penetrate flesh and provide an image of higher-density material such as bones. Gamma rays represent the smallest wavelength (less than 10^{-13} meter). They exhibit

NATURE AND PROPERTIES OF LIGHT

particle-like properties with great energy and are emitted by the sun, linear and particle beam accelerators, and nuclear processes.

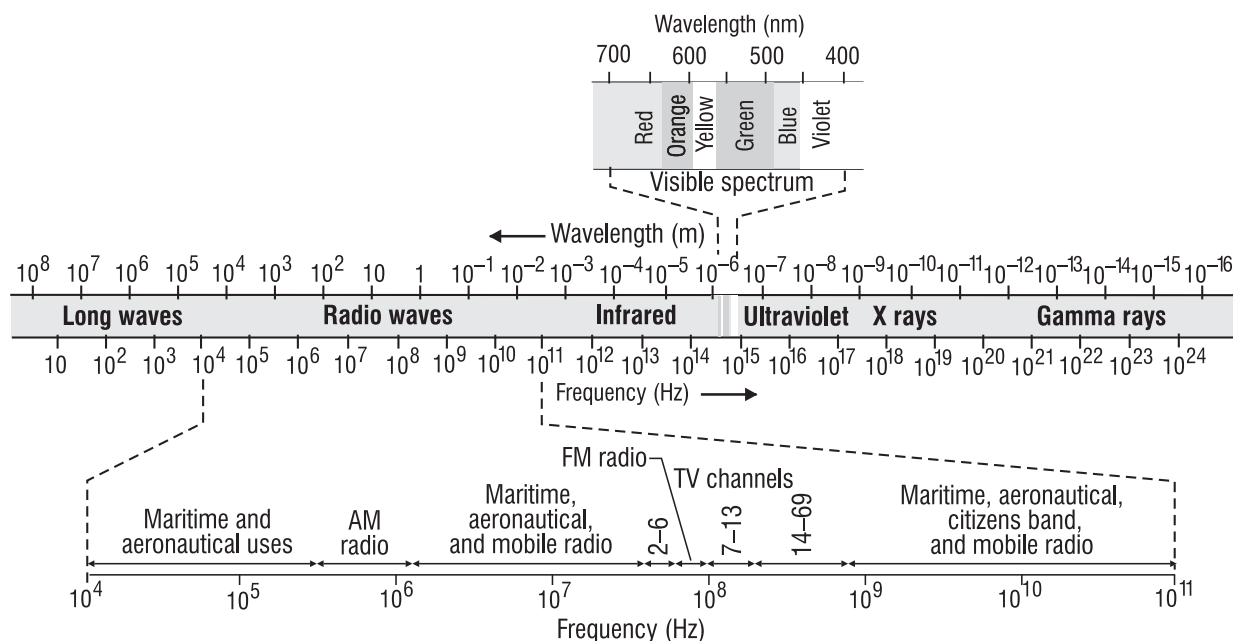


Figure 1-11 *Electromagnetic spectrum*

White light is a mixture of light of different colors. Each of these colors has a different wavelength and, when passed through a transparent medium, refracts differently. Thus, a prism can separate white light into its component colors, as shown in Figure 1-12.

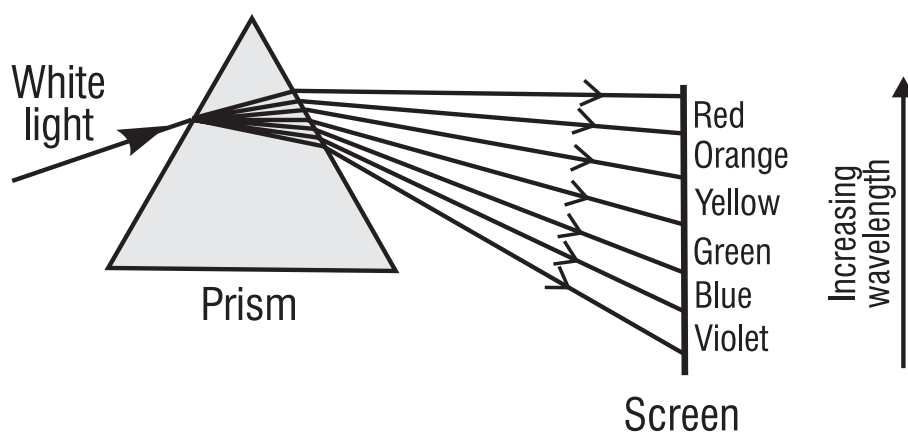


Figure 1-12 *Separation of light into component colors*

The colors displayed in visible light are categorized by wavelength. Table 1-2 gives the wavelengths of these colors. An arrangement showing the different components of light, with the wavelengths of the components in order, is called the spectrum of the light.

Table 1-2: Visible Spectrum Wavelengths

Color	Wavelength Band (μm)	Representative wavelength (μm)
Extreme violet	0.39–0.41	0.40
Violet	0.39–0.45	0.43
Dark blue	0.45–0.48	0.47
Light blue	0.48–0.50	0.49
Green	0.50–0.55	0.53
Yellow-green	0.55–0.57	0.56
Yellow	0.57–0.58	0.58
Orange	0.58–0.62	0.60
Red	0.62–0.70	0.64
Deep red	0.70–0.76	0.72

Spectra of Light Sources

The sources of electromagnetic radiation are many and varied. Usually sources are divided into two categories, natural and man-made. Examples of natural sources of radiation are the sun, observable stars, radio stars, lightning, and, in fact, any body that exists at a temperature over absolute zero. Some of the man-made sources of radiation are incandescent and fluorescent lights, heaters, lasers, masers, radio and television antennas, radars, and X-ray tubes.

Two types of spectra are important in photonics: the emission and the absorption spectra. An emission spectrum is from light emitted by a source. An absorption spectrum is from light that has passed through an absorbing medium.

All materials with temperatures above absolute zero emit electromagnetic radiation. Every atom and molecule has its own characteristic set of spectral lines. The understanding of the wavelength and energy that produce the spectral “fingerprint” is built on an explanation of the atomic and molecular structure. The line spectra observed early in the scientific age led to significant understanding of the nature of atoms. They even led to the development of modern quantum theory, which says that light emitted by an atom or molecule has a discrete wavelength, corresponding to a specific energy-level change within the atom or molecule.

These fingerprints can have any combination of spectral lines, bands, and continuums. Atoms changing states produce visible and ultraviolet radiation. Molecules changing vibrational and rotational states produce infrared radiation. For dense materials, many energy states are available; thus emission and absorption bands cover broad regimes for solids and liquids. For a less dense gas, the spectral bands are much narrower.

To observe a line or band spectrum, a light is passed through a slit. The image of this slit is then refracted by a prism or diffracted by a grating, based on the constituent wavelengths of the light. This is recorded on film or a spectrograph. The lines relate back to the atomic structure and the unique energy-level changes. Spectroscopy is the science that analyzes line spectra and identifies constituents of materials.

When atoms are close to each other, their electrons interact and the energy levels split. In a solid, there are so many levels that a continuous range of frequencies can be emitted or absorbed. Hot, dense materials emit *continuous spectra* containing bands of frequencies.

Atomic Structure

All matter is made up of atoms. An atom is the smallest unit that retains the characteristics of a chemical element. It consists of a positive nucleus surrounded by negative electrons arranged in distinct energy shells designated *K* through *O*, as shown in Figure 1-13. The notation *K*(2) indicates that the *K*-shell is complete when it has 2 electrons. Similarly, *L*(8) indicates that 8 electrons complete the *L*-shell, and *M*(18) indicates that 18 electrons complete the *M*-shell.

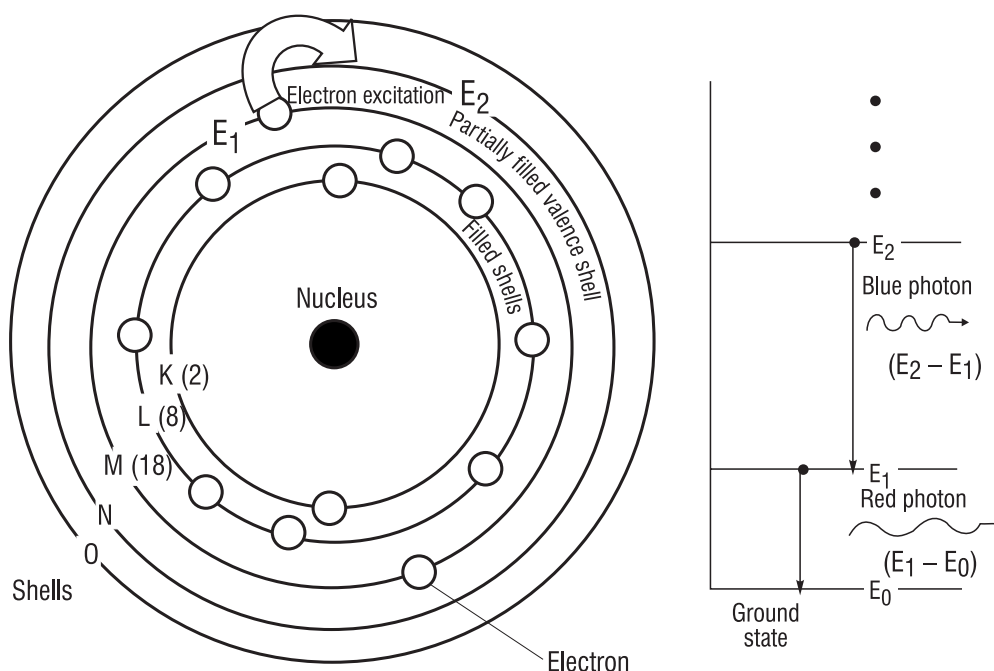


Figure 1-13 Atomic model

We model the energy of an atom with the electrons. When all the electrons are in an unexcited, or ground, state, the atom is assumed to be at its lowest energy level. When the atom absorbs energy, electrons can be “excited” and moved into higher-energy shells. As electrons move from one shell to another, unique amounts, or quanta, of energy are absorbed or emitted. This is how an atom can absorb or emit light. The light’s unique energy quanta are dependent on the electronic structure of the atom.

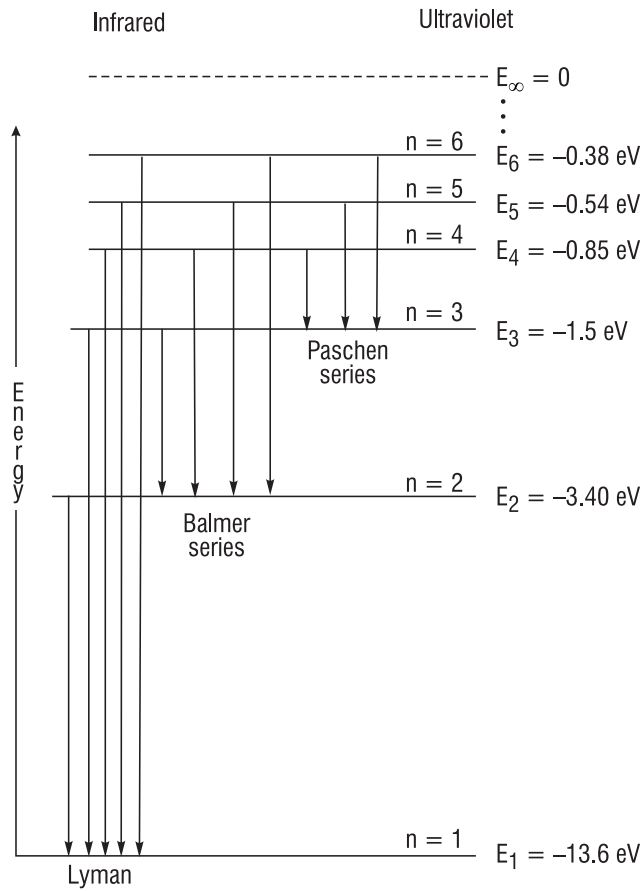


Figure 1-14 Energy-level diagram for a hydrogen atom

An atomic energy-level diagram shows the unique electron energies available in a given atom. An energy-level diagram for hydrogen is shown in Figure 1-14. Hydrogen has only one electron, and it can exist in only one of the energy levels shown at a time. The lowest level, E_1 , is the ground state. Energy must be added to the atom for the electron to move to a higher level. Note that energy levels range from a value of -13.6 eV (electron volts) for the lowest energy level ($n = 1$) to a value of 0 eV for the very highest energy level ($n = \infty$)—when the electron breaks free from the atom.

Suppose a hydrogen atom is in an excited state, say, the $n = 3$ level. The atom can make a transition to the ground state by emitting a photon. The energy of the photon equals the change in energy of the atom, as given by Equation 1-7.

$$\begin{aligned}
 E_{\text{photon}} &= E_3 - E_1 \\
 &= -1.51 \text{ eV} - (-13.6 \text{ eV}) \\
 &= 12.09 \text{ eV}
 \end{aligned}
 \tag{1-7}$$

The atom can also absorb photons whose energies exactly match differences between electron energy levels. For example, a hydrogen atom in the ground state can absorb a photon whose energy is 12.09 eV . The electron in the atom will move from energy level E_1 to energy level E_3 .

Molecular Structure

Molecules in gases or liquids can also absorb electromagnetic radiation. The photon energy must match a discrete rotational or vibrational energy level of the molecule. In solids, absorption is more complex, generally resulting from vibrational energy changes.

Blackbody Radiation

The first step toward developing an understanding of *blackbody radiation* is to describe the relationships between temperature, wavelength, and energy emitted by an ideal thermal radiator (blackbody). Based on our everyday observations, we know that bodies at different temperatures emit radiation (heat energy) of different wavelengths or colors. For example, the wires in a heater begin to glow red when heated.

Blackbody radiation is the theoretical maximum radiation expected for temperature-related thermal self-radiation. This radiation can have a peak energy distribution in the infrared, visible, or ultraviolet region of the electromagnetic spectrum. The hotter the emitter, the more energy emitted and the shorter the wavelength. An object at room temperature has its peak radiation in the infrared while the sun has its peak in the visible region.

The equations for calculating radiation based on temperature use the Kelvin temperature scale. (Be sure to use the Kelvin scale for all calculations.) The conversions between the different temperature scales are provided in Equations 1-8 and 1-9.

Fahrenheit (F) to Celsius (C)	$^{\circ}\text{C} = 5/9 (^{\circ}\text{F} - 32)$	(1-8)

Celsius (C) to Kelvin (K)	$\text{K} = ^{\circ}\text{C} + 273.15$	(1-9)

Example 1-5

Convert the following Fahrenheit temperatures to degrees Celsius and Kelvin: 212, 100, 32, 0, –100, –434.

Solution:

Use Equations 1-8 and 1-9 to complete Table 1-3 with the correct temperatures.

Table 1-3: Temperature Conversion

	Fahrenheit (°F)	Celsius (°C)	Kelvin (K)
Boiling water	212	100	373
	100	38	311
Freezing water	32	0	273
	0	–18	255
	–100	–73	200
Absolute zero	–434	–273	0

A waveband is a portion of the electromagnetic spectrum between defined upper and lower wavelengths. The energy radiated by a blackbody in a given waveband is the sum of all energies radiated at the wavelengths within the band. The rate of energy radiation is the power radiated. You can also add the power over all emitted wavelengths to find the total power radiated by a blackbody. For a blackbody at temperature T , the power radiated per unit surface area of the radiator is given by the Stefan-Boltzmann law in Equation 1-10.

$$W_s = \sigma_s T^4 \text{ watts/m}^2 \quad (1-10)$$

where $\sigma_s = 5.67 \times 10^{-8} \text{ watts/m}^2 \cdot \text{K}^4$ (Stefan-Boltzmann constant)
 T = Temperature (K)

The power per unit area, W_s , is called the emitted radiant flux density. A graybody is one that does not emit as a perfect “blackbody” but at a fraction of the theoretical maximum of a blackbody. The blackbody’s emitted radiant flux density is reduced by a factor called the emissivity. The emissivity (ϵ) is dependent on the material emitting and is less than 1. Thus, for a graybody the emitted radiant flux density is expressed in Equation 1-11.

$$W_s = \epsilon \sigma_s T^4 \text{ watts/m}^2 \quad (1-11)$$

Example 1-6

Calculate the radiant flux density emitted by a graybody (emissivity = 0.7) at room temperature (82°F).

Solution:

First we must convert 82°F to Kelvin. This is 301 K. We then use Equation 1-11 and plug in the numbers.

$$W_s = (0.7) \left(5.67 \times 10^{-8} \frac{\text{watts}}{\text{m}^2 \text{ K}^4} \right) (301 \text{ K})^4 = 325.5 \text{ W/m}^2$$

Spectral distribution

The radiation emitted by a blackbody is distributed over wavelength. The quantity W_λ is called the spectral flux density. It is defined so that $W_\lambda \Delta\lambda$ is the power radiated per unit area of surface for wavelengths in the waveband $\Delta\lambda$ (between λ and $\lambda + \Delta\lambda$). In 1900, Max Planck developed a formula that fits experimental measurements of W_λ extremely well. Planck’s radiation formula is given by Equation 1-12.

$$W_{\lambda} = \frac{C_1}{\lambda^5} \frac{1}{e^{\frac{C_2}{\lambda T}} - 1} \quad (1-12)$$

where λ = wavelength (m)

T = blackbody temperature (K)

$$C_1 = 2 \pi c^2 h = 3.75 \times 10^{-16} \text{ W} \cdot \text{m}^2$$

$$C_2 = hc/k = 1.44 \times 10^{-4} \text{ m} \cdot \text{K}$$

c = speed of light = 3.00×10^8 m/s

h = Planck's constant = 6.626×10^{-34} J·s

k = Boltzmann's constant = 1.38×10^{-23} J/K

The blackbody spectral flux density from Planck's formula is plotted in Figure 1-15 for five blackbody temperatures. The wavelengths are plotted in units of microns. Notice that the axes are logarithmic.

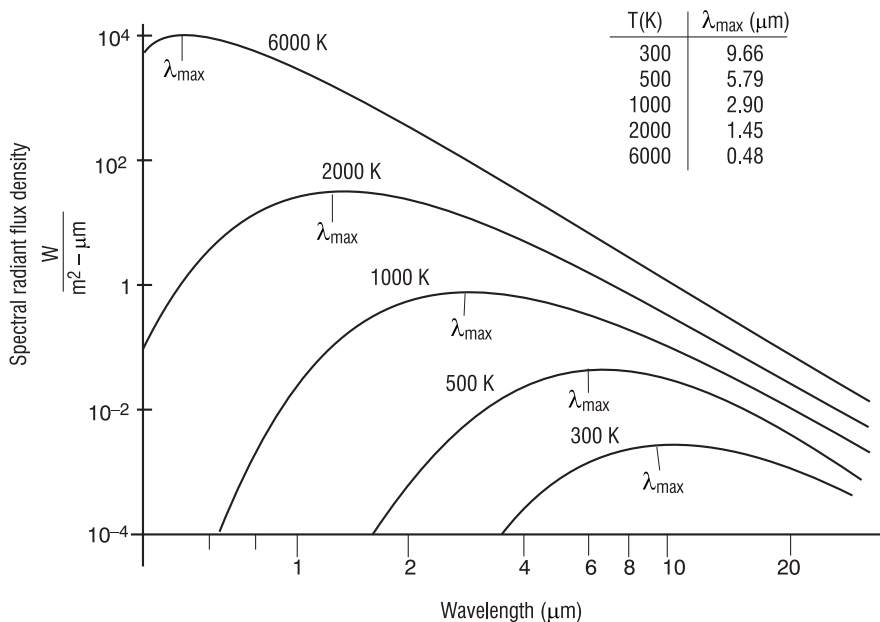


Figure 1-15 Spectral radiant blackbody flux density distributions at various temperatures

Wien's displacement law

The spectral distribution for each blackbody temperature has a maximum, or peak, emission wavelength. This maximum wavelength is related to the blackbody temperature. The relationship is given by Wien's displacement law:

$$\lambda_{\max} T = 2.898 \times 10^{-3} \text{ m} \cdot \text{K} \quad (1-13)$$

Wien's displacement law predicts that the peak wavelength decreases in value as the temperature of the blackbody increases.

Example 1-7

Calculate the apparent blackbody temperature of the sun. If it is observed that the peak spectral radiant flux density of the sun is near 490 nm, what is its effective blackbody temperature?

Solution:

By applying Wien's displacement law and solving for T , we can find the sun's effective temperature.

$$\lambda_{\max} T = 2.898 \times 10^{-3} \text{ m} \cdot \text{K}$$

$$T = \frac{2.898 \times 10^{-3} \text{ m} \cdot \text{K}}{490 \times 10^{-9} \text{ m}} = 5914 \text{ K}$$

This equation allows the choice of the wavelength most advantageous for detectors given an expected target temperature.

Interactions of Light with Matter

When light travels through a medium, it interacts with the medium. The important interactions are absorption and scattering.

Absorption

Absorption is a transfer of energy from the electromagnetic wave to the atoms or molecules of the medium. Energy transferred to an atom can excite electrons to higher energy states. Energy transferred to a molecule can excite vibrations or rotations. The wavelengths of light that can excite these energy states depend on the energy-level structures and therefore on the types of atoms and molecules contained in the medium. The spectrum of the light after passing through a medium appears to have certain wavelengths removed because they have been absorbed. This is called an absorption spectrum.

Selective absorption is also the basis for objects having color. A red apple is red because it absorbs the other colors of the visible spectrum and reflects only red light.

Scattering

Scattering is the redirection of light caused by the light's interaction with matter. The scattered electromagnetic radiation may have the same or longer wavelength (lower energy) as the incident radiation, and it may have a different polarization.

If the dimensions of the scatterer are much smaller than the wavelength of light, like a molecule, for example, the scatterer can absorb the incident light and quickly reemit the light in a different direction. If the reemitted light has the same wavelength as the incident light, the process is called *Rayleigh scattering*. If the reemitted light has a longer wavelength, the molecule is left in an excited state, and the process is called *Raman scattering*. In Raman scattering, secondary photons of longer wavelength are emitted when the molecule returns to the ground state.

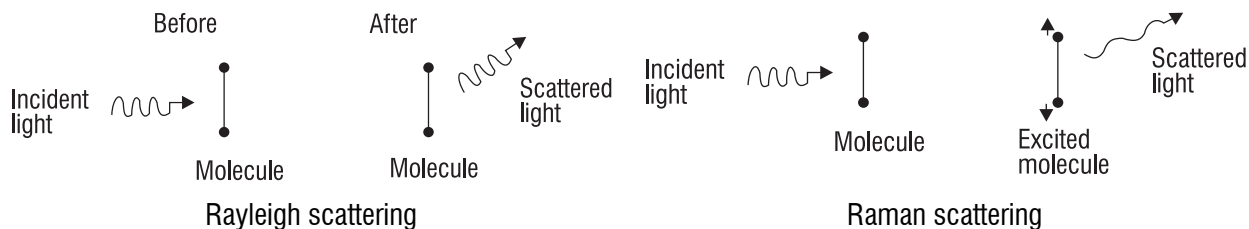


Figure 1-16 Rayleigh and Raman scattering

Air molecules (O_2 and N_2) are Rayleigh scatterers of visible light and are more effective at scattering shorter wavelengths (blue and violet). Can you use this information to explain why, on a clear day, the sky looks blue?

If the scatterer is similar in size to—or is much larger than—the wavelength of light, matching energy levels is not important. All wavelengths are equally scattered. This process is called *Mie scattering*. Water droplets effectively scatter all wavelengths of visible light in all directions. Can you use this information to explain the color of a cloud?

Laboratories

Purpose

Examine basic properties of light such as the following:

- speed
- wavelength
- color spectrum of visible light
- polarization

Laboratory 1.1A—*Finding the Speed of Red Light in Optical-Grade Plastic*

Theory

The speed of light in a vacuum, c , is exactly 299,792,485 m/s. Current physical theory asserts that nothing in our universe can have a speed greater than c . When light, or any electromagnetic wave, moves in any other medium it will have a speed less than c . In general, the speed of light through a medium depends upon both the medium and the wavelength of the light. The object of this experiment is to use the definition of index of refraction and Snell's law to determine the speed of red light in optical plastic.

Key Definitions and Relationships

1. The index of refraction for any medium, n , is defined as: $n_i = \frac{c}{v_i}$ where v_i is the velocity of light in medium, i .
2. Snell's law $\frac{n_1}{n_2} = \frac{\sin \theta_2}{\sin \theta_1}$
3. The speed of light in vacuum and air is the same to an accuracy of six significant figures, so we will use $n_{air} = 1.00000$.

Equipment

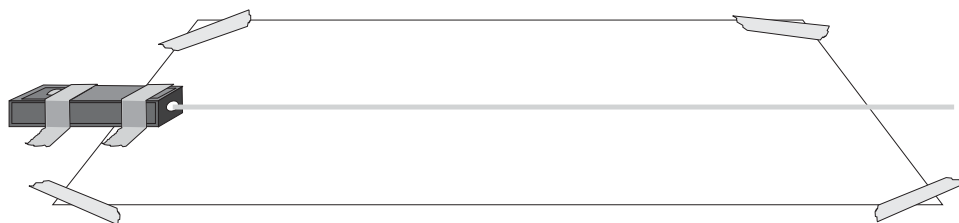
1. Laser, either a red laser pointer or a Class I or II HeNe laser
2. Laboratory stand and clamps to hold laser and gratings
3. Plastic optical block, approximately 8 cm $\times$ 6 cm $\times$ 2 cm

NATURE AND PROPERTIES OF LIGHT

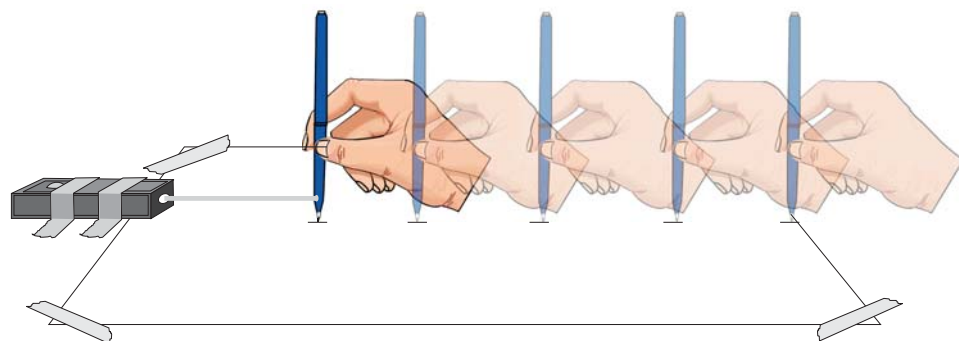
4. $8\frac{1}{2}" \times 14"$ white paper
5. Masking tape
6. Meterstick
7. Protractor

Procedure

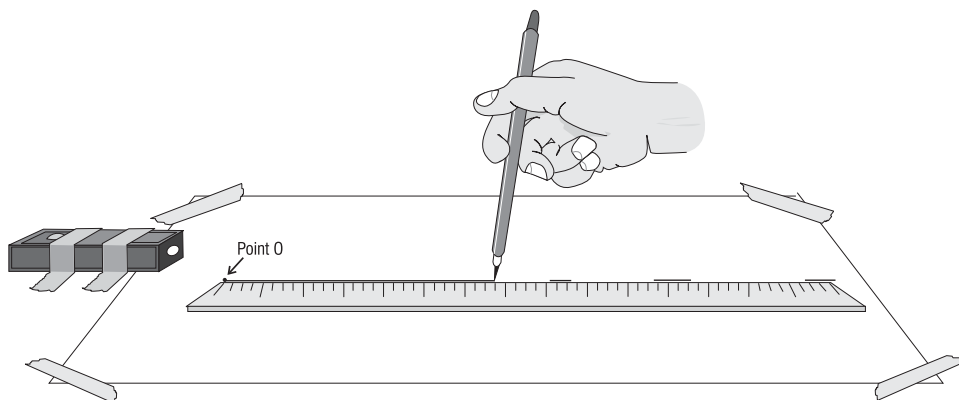
1. Tape a legal-size sheet of blank white paper to the tabletop. Tape a laser pointer on the left side so the beam is directed left to right across the paper.



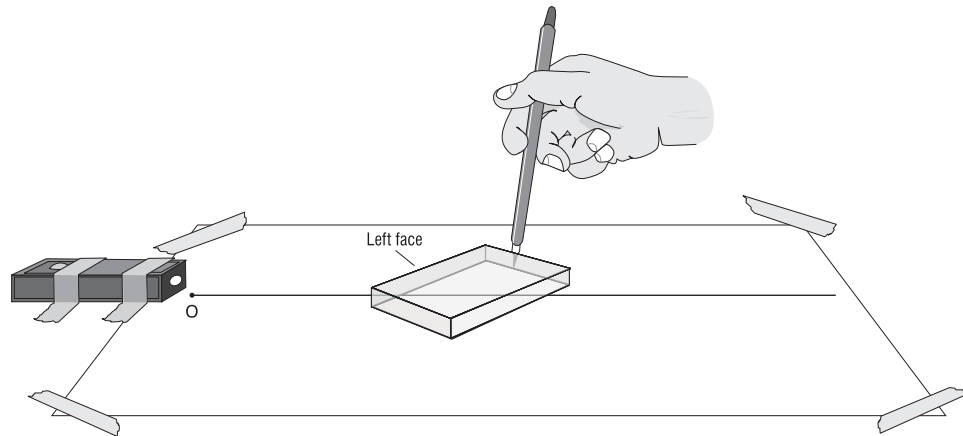
2. Turn on the laser. Hold a pencil in a vertical position at the right edge of the paper where the laser beam hits the center of the pencil. Mark the paper with a small dot or dash. See sketch.



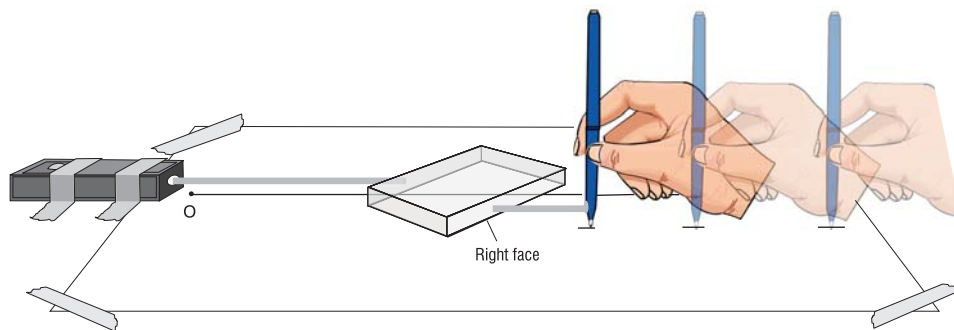
Move the pencil directly left one or two inches and repeat. Continue until you have five or six marks extending left to right across the paper. Draw a “best-fit” straight line through the marks. Label the point at the left end of this line point O. See sketch.



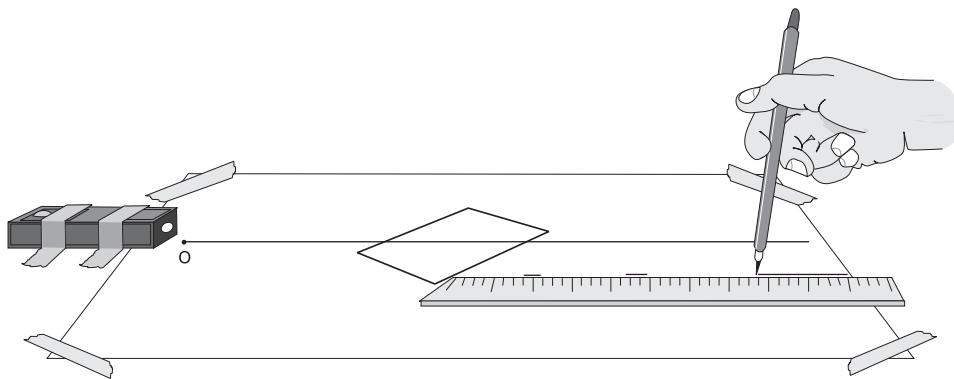
- Place the plastic block with the large surface down. The left face of the block should intersect the line on the paper at a 40- to 50-degree angle, and the laser beam should hit the left face about 1 cm from its lower left corner. Hold the block firmly in place and draw an outline of the block on the paper. See sketch.



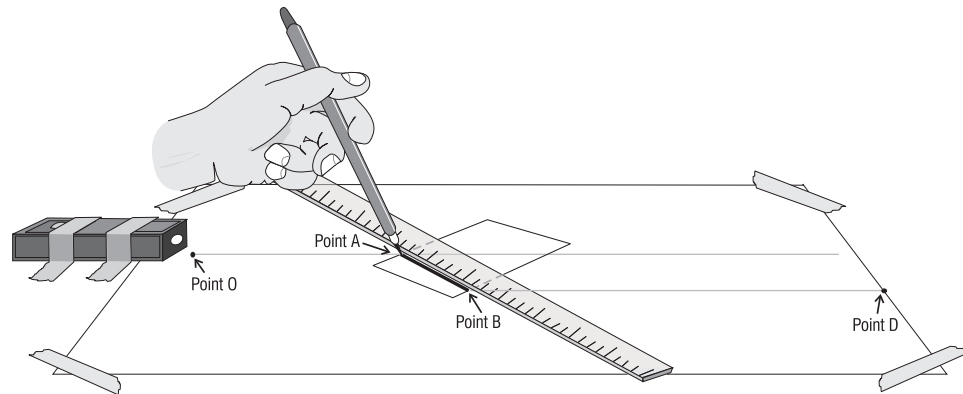
- Turn on the laser. The beam is refracted by the block and should exit the block through the right face. Move your pencil along the right edge of the paper until you find the beam. Mark the location of this exit beam at five or six places on the paper just as you did for the straight beam in step 2. See sketch.



- Remove the plastic block. Draw a best-fit straight line through the marks along the exit beam path. See sketch.

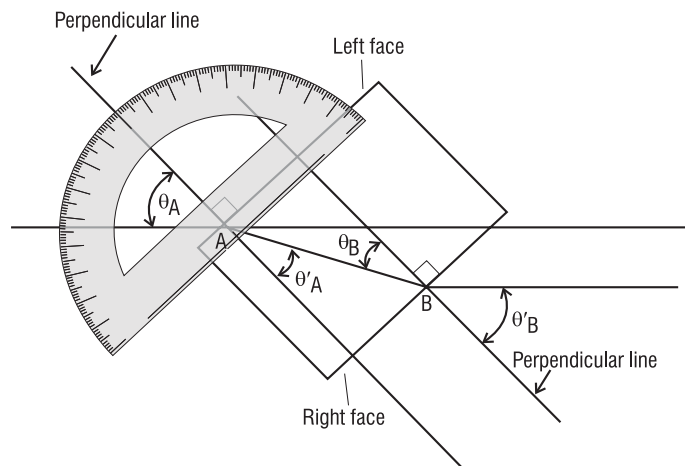


Mark the right end of the line as point D . Extend this line to the left until it intersects the line marking the lower edge of the plastic block. This is the point where the beam left the block. Mark this point as point B . Mark the point where the incident beam hit the left face of the block as point A . Connect points A and B with the line segment AB . See sketch.



Note: The line segment OA describes the path of the laser beam that is incident on the block. The line segment AB describes the path of the refracted beam through the block. The segment BD describes the beam's path after it exits the block.

6. Use the protractor to draw a line through point A that is perpendicular to the left face of the block. Draw another line through point B that is perpendicular to the right face of the block. See sketch.



7. Measure and record the angle θ_A between the incident beam and the normal line at the left face of the block (point A). Do the same for the refracted beam at this face. This is angle θ'_A . See sketch.
8. Measure the angle between the normal line and the incident and refracted (exit) beam at the right face of the block (point B). Label these as θ_B and θ'_B . See sketch.
9. Using θ_A and θ'_A , use Snell's law to find the ratio $\frac{n_{air}}{n_{plastic}}$ at the left interface between air and plastic.

10. Using θ_B and θ_B' , use Snell's law to find the ratio $\frac{n_{\text{plastic}}}{n_{\text{air}}}$ at the right interface between plastic and air.
11. The index of refraction of air, n_{air} , has a value of 1.00000. Use this and the results of steps 9 and 10 to find the numerical value of n_{plastic} from your measurements at A and at B. Average the two values to determine your final estimate of n_{plastic} .
12. Use the definition of index of refraction and the known value of speed c in a vacuum to calculate the speed of light in the plastic block.

Laboratory 1.1B—*Determining the Wavelength of Red Light*

Theory

When a beam of light is incident on a diffraction grating, part of the light will pass straight through. Part of the light is diffracted to paths that diverge at different angles on both sides of the original path. The angle θ at which the light diverges is related to the wavelength and spacing of the lines on the grating. The relationship is described by

$$m\lambda = d \sin\theta_m \quad \text{where } \lambda \text{ is the wavelength of the incident light in meters, } d \text{ is the spacing between lines on the grating in meters, } m \text{ is an integer that takes on the values } 0, 1, 2, \dots, \text{ and } \theta_m \text{ is the diffraction angle for a particular diffraction order } m.$$

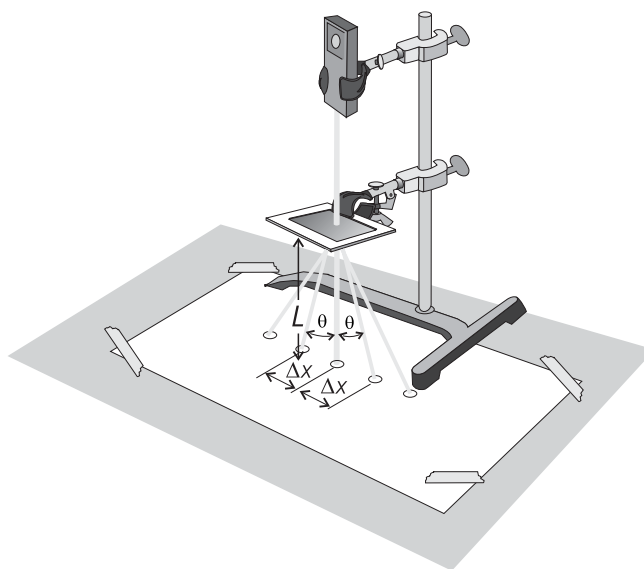
If the diffraction angle θ_m can be measured for a particular order m and the grating spacing d is known, the wavelength of the light can be calculated.

Equipment

1. Laser, either a red laser pointer or a Class I or II HeNe laser
2. Laboratory stand and clamps to hold laser and gratings
3. Transmission grating with 300 to 800 lines/mm
4. $8\frac{1}{2}" \times 14"$ white paper
5. Masking tape
6. Meterstick

Procedure

1. Position the laser so the beam goes straight down through the grating—perpendicular to the grating surface—and onto the white paper. There it produces a center spot with diffracted spots on both sides as shown in the sketch.



2. Measure the vertical distance from the grating to the paper. Record this as L .
3. Measure the distances from the center spot to the first diffracted spots on both sides. Average these two distances and record the average as Δx . For these nearest diffracted spots, $m = 1$.
4. Calculate the diffraction angle using $\theta_1 = \tan^{-1}\left(\frac{\Delta x}{L}\right)$
5. Calculate the wavelength of the red laser light using this first-order diffraction angle, θ_1 where $m = 1$. The equation is then

$$\lambda = \frac{d \sin \theta_1}{1}$$

Laboratory 1.1C—*The Spectrum of Colored Light*

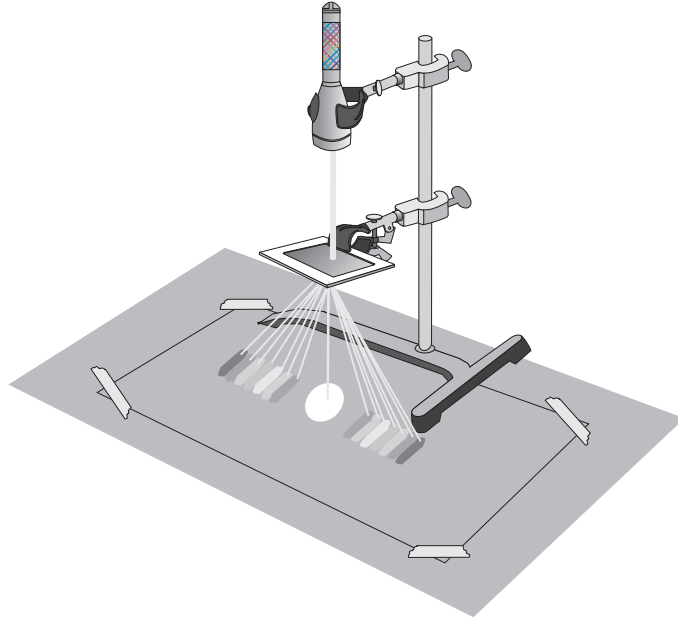
Equipment

1. Flashlight with focusing capability, similar to the Mini-Maglight series
2. Laboratory stand and clamps to hold laser or flashlight and grating
3. Transmission grating with 300 to 800 lines/mm
4. Red, green, blue, purple, yellow, and orange filters
5. 8½" × 14" white paper
6. Masking tape

Procedure

Part 1. The Spectrum of White Light

1. Mount the flashlight in a clamp on the stand with the beam projected straight down onto a sheet of white paper.
2. Mount the diffraction grating in another clamp and position it four or five inches below the flashlight. Focus the light perpendicularly onto the grating surface. See sketch.
3. Move the grating, and flashlight if necessary, up or down until you see clearly both the light transmitted straight down through the grating, forming a white spot on the paper, and the first order spectrum of colors. See sketch.



4. Draw lines through each color you can identify in the spectrum, and label each line with its color.

Part 2. The Components of Different Colors of Light

5. Hold a red transmission filter between the flashlight and the grating. What is the color of the center spot where light is transmitted straight through the grating? List all the colors that you can clearly identify in the diffracted spectrum on either side of the center spot.
6. Replace the red filter with the other filters in this order: green, blue, purple, yellow, and orange. For each filter list all the colors you can identify in the diffracted spectrum of the light formed on either side of the center spot.
7. Answer the following questions with complete sentences.
 - (a) Why are red, green, and blue primary colors?
 - (b) What colors of light must be combined to make purple light?
 - (c) What colors of light must be combined to make yellow light?
 - (d) How can a color TV produce any color it needs when it has only red, green, and blue color guns?

Laboratory 1.1D—*The Polarization of Light*

Equipment

1. Flashlight with focusing capability, similar to the Mini-Maglight series
2. Laboratory stand and clamps to hold laser or flashlight and grating
3. Two polarizing filters
4. Microscope slide or similar thin, flat glass plate
5. $8\frac{1}{2}'' \times 14''$ white paper
6. Masking tape

Procedure

Part 1. Polarizers and Analyzers

1. Clamp a flashlight with the beam projected horizontally about five feet above the floor. It should be arranged so that it is easy to look directly into the light when you are five or six feet from the flashlight.

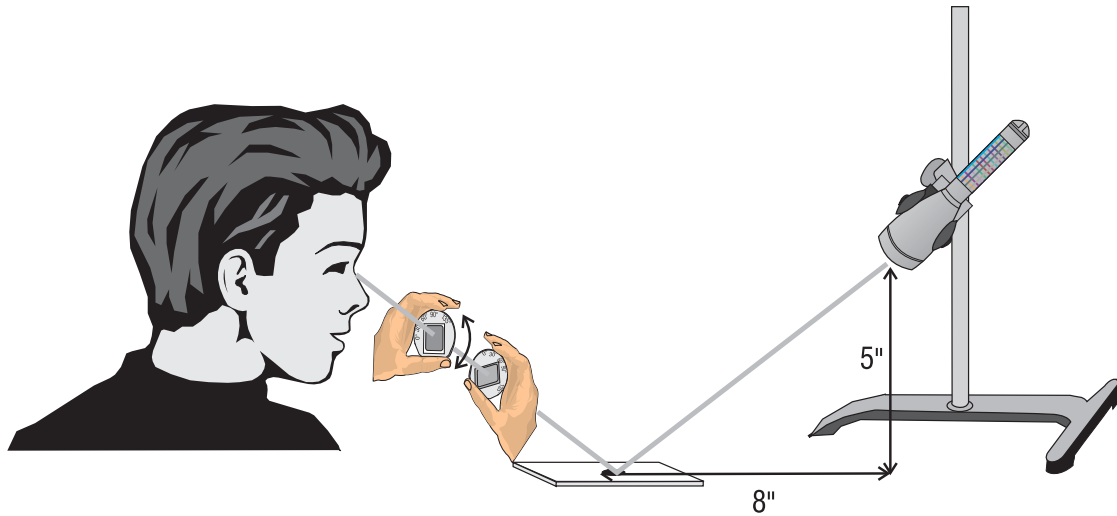
Hold one polarizer at arm's length in front of you and look at the light through the polarizer. The light you see is now polarized in the preferred direction of the filter.

2. Hold a second polarizing filter (analyzer) with your other hand. Place it between you and the first filter. Rotate the second filter about the axis of the light beam. Notice the change in brightness of the light passing through both filters and reaching you.
3. What can you say about the relation between the polarizing direction of the two filters when the light transmitted has its maximum brightness?
4. What can you say about the relation between the polarizing direction of the two filters when the light transmitted has its minimum brightness?

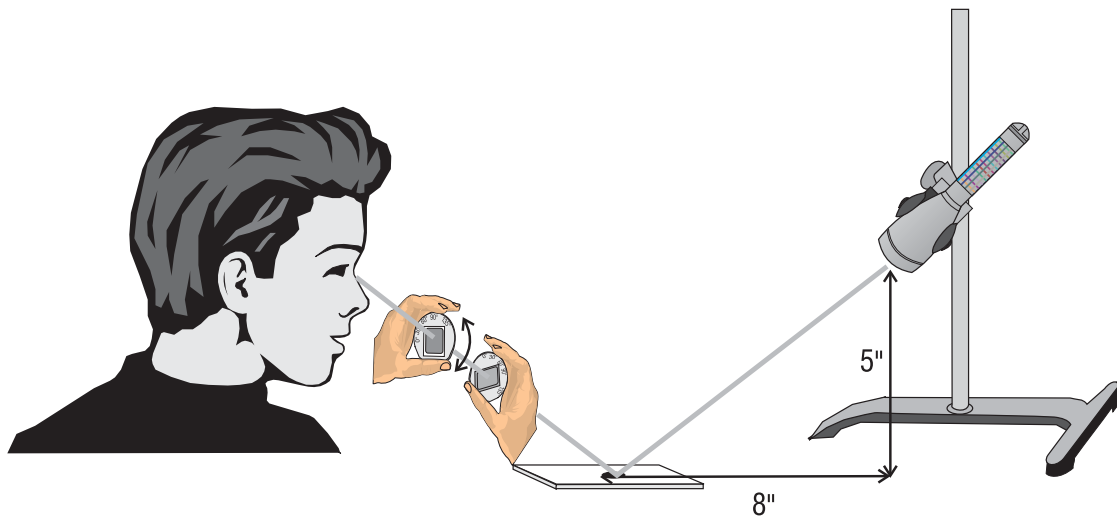
Part 2. Polarization by Reflection

5. Clamp the flashlight so the center of the lens is 5 inches above the table and the light beam is focused on a spot eight inches horizontally from the flashlight. Place a microscope slide on the table at the position of the focused spot. See sketch.

6. Position yourself in line with the microscope slide and the flashlight. Move until you can see the reflection of the flashlight from the slide. See sketch.



7. Hold a polarizing filter so you can see the reflection through the filter. Rotate the filter about the axis of the reflected beam. See sketch. What do you observe about the brightness of the reflection as you rotate the filter?



8. How does the reflection from the glass affect the properties of the reflected light?

Problem Exercises/Questions

1. Discuss some of opportunities for technicians in the photonics field.
2. Define the following properties of light:
 - a. Speed
 - b. Frequency
 - c. Wavelength
 - d. Energy
3. Discuss the dual nature of light wave versus photon and give examples of each.
4. Describe in a short paragraph the electromagnetic spectrum with a diagram of the wavelength regions and typical applications in those regions.
5. An electron in a hydrogen atom is designated to have energy (relative to infinity) with
 - a. Any value
 - b. Any positive value
 - c. Any negative value
 - d. Only certain isolated values
6. Give the primary equations describing the relationships between temperature, wavelength, and energy emitted by a blackbody.
7. Address the mechanisms that affect light propagating in a medium and its transmission.
8. Which of the following light sources emits a continuous spectrum?
 - a. A neon light
 - b. A glowing coal
 - c. A mercury vapor lamp
 - d. Hot, thin interstellar gas
9. Which of the following colors corresponds to the longest wavelength?
 - a. Blue
 - b. Violet
 - c. Red
 - d. Green
10. List the six properties of wave motion.

Student Project

Create a presentation that educates others at the elementary school level about photonics and the uses of light.

Bibliography and Resources

Accetta/Schunder. *IR/EO System Handbook*. ERIM and SPIE Press.

Cobb, Vickie, Joshua Cobb, and Theo Cobb. *Light Action—Amazing Experiments with Optics*. Harper Collins Children's Books, 1993.

Ford, Kenneth. *Basic Physics*. Walton, Massachusetts: Blaisdell Publishing Co., 1968.

Hecht, Jeff. *Optics Light for New Age*.

Jenkins and White. *Fundamentals of Optics*. New York: McGraw-Hill, 1976.

National Photonics Skills Standard for Technicians. Pittsfield, Massachusetts: Laurin Publishing Company, Inc., 1995.

Seyrafi, Khalil. *Electro-Optical Systems Analysis*. Los Angeles: Electro-Optical Research Company, 1985.

Laser As a Tool (Video) and *Career Encounters: Optics and Photonics* (Video). Washington, D.C.: Optical Society of America, Ph 202/223-8130, Fax 202/223-1096, E-mail <postmaster@osa.org>.

The Photonics Dictionary. www.laurin.com/DataCenter/Dictionary//CD/index.htm

Metric Prefixes

Prefix	Abbreviation	Power of Ten	Value
tera	T	10^{12}	thousand billion
giga	G	10^9	billion
mega	M	10^6	million
kilo	k	10^3	thousand
centi	c	10^{-2}	hundredth
milli	m	10^{-3}	thousandths
micro	μ	10^{-6}	millionth
nano	n	10^{-9}	billionth
pico	p	10^{-12}	thousand billionths

Light Sources and Laser Safety

Fred Seeber
Camden County College
Blackwood, New Jersey

This module addresses various nonlaser light sources, such as incandescent lamps, fluorescent discharge lamps, high-intensity discharge lamps, flashlamps, arc lamps, and LEDs. It will also discuss in detail laser safety including the human eye, laser hazards, laser safety controls, maximum permissible exposure, laser safety glasses, and laser accidents. The laboratory experience at the end of the module will enhance your ability to understand the dangers associated with lasers, with small and large beam diameters, and with constant power levels.

Prerequisites

Before you work through this module, you should have completed Module 1-1, *Nature and Properties of Light*. You should also have a working knowledge of intermediate algebra.

Objectives

Upon completion of this module, you should be able to:

1. Define the following properties of laser light:
 - a. Monochromaticity
 - b. Directionality
 - c. Coherence
2. Distinguish between the different types of nonlaser light sources and identify their characteristics.
3. Recognize and avoid various nonbeam hazards, such as electrical and chemical hazards.
4. Label a diagram of the human eye. State the function of each part shown.
5. Match the parts of the eye with the spectral regions to which they are most vulnerable.
6. Given the basic information required, calculate retinal spot size and retinal irradiance.
7. Describe the following laser classifications, based upon the potential hazards:
 - a. Class 1 laser
 - b. Class 2 laser
 - c. Class 3a laser
 - d. Class 3b laser
 - e. Class 4 laser
8. Define maximum permissible exposure. State why it's important.
9. Describe three general types of laser hazard controls.
10. Describe laser-hazard warning signs.
11. Describe the most frequent causes of laser accidents.
12. List five laser safety precautions applicable to all types of lasers.

Scenario—Laser Safety in a Medical Environment

Joan recently graduated from a two-year photonics program at her local community college. She accepted a job at City Hospital in Philadelphia as a laser technician.

City Hospital uses CO₂, argon, and Nd:YAG lasers for its surgical procedures. Joan soon discovered that her responsibilities would also include being laser safety officer (LSO). Joan recalled the discussions she had in school about laser safety and the laser safety concepts she learned. Because being a laser safety officer is a major responsibility and critical to the well-being of both the staff and the patients, the hospital and Joan realized more training would be necessary. The hospital sponsored Joan's attendance at a five-day course on medical laser safety given by the Laser Institute of America in Orlando, Florida. At this course the duties of a laser safety officer were discussed, including such topics as safety education, laser hazards and controls, personal protection, medical surveillance, and more. As Joan begins her duties as LSO, she realizes how important laser safety is for all in the medical working environment. Good luck, Joan, as a photonics technician and laser safety officer.

Opening Demonstration

Note: The following demonstration will provide a view of some of the concepts you will encounter in this module.

Factors that result in the absorption of laser light. In a room or laboratory with reduced ambient light (light already in room), set up a demonstration as seen in Figure D-1 using a laser pointer (laser diode), red and blue plastic or glass filters, and a piece of white cardboard or paper.

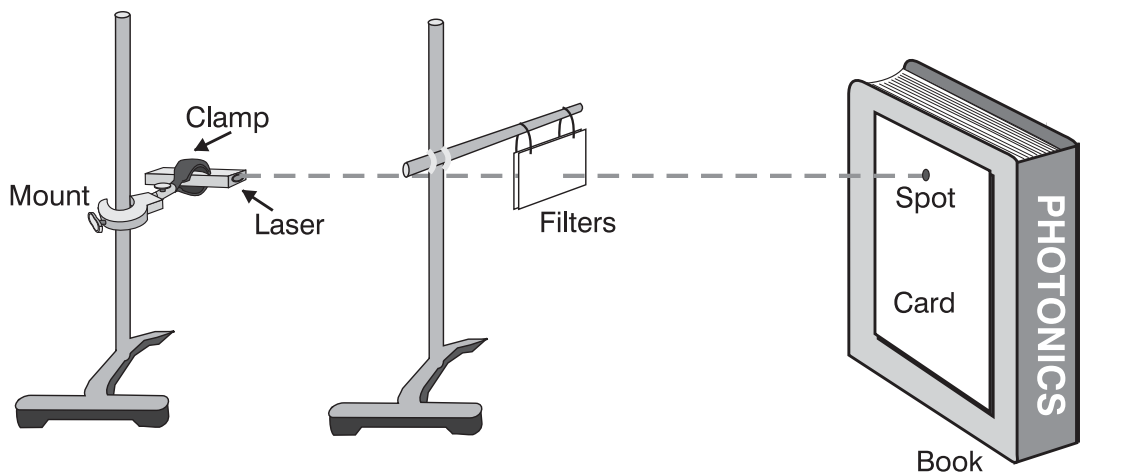


Figure D-1

1. Laser Beam Absorption as a Function of Filter Thickness. Attach a laser pointer to an ordinary test-tube holder and lab stand similar to those found in most general science laboratories. (Note: A class 2 HeNe laser can be used instead of a laser pointer. Modify mounting of laser accordingly.)

Point the laser beam at a piece of white paper or cardboard taped to a wall or a textbook to be used as a stand. Turn on the laser pointer and visually notice the intensity of light on the white paper. Starting with the red filters, hold one thickness in front of the beam exit. Observe whether any decrease in intensity occurs. Repeat with two, three, and up to six equal thicknesses of red filter material and observe light intensity on the white paper as filter thickness increases. What do you notice? Does the intensity with each added thickness decrease in a linear fashion? More slowly than linear? More rapidly than linear?

2. Laser Beam Absorption as a Function of Filter Color. Substitute blue filters in this section of the demonstration for the red filters used in step 1. Start with one blue thickness and increase to two and three equal thicknesses. If you find complete extinction after three blue filters, what can you conclude about the role filter color plays in beam absorption?

As you proceed through this module—and also later modules—you will discover that absorption of a laser beam depends on a number of different factors. These are especially important in the design of laser safety goggles.

Basic Concepts

1. NONLASER LIGHT SOURCES

As discussed in Module 1-1, *Nature and Properties of Light*, light can be understood as consisting of waves traveling through space. Light occurs in different wavelengths, just as waves on the ocean vary in length. Wavelength is the distance between peaks on a light wave, as represented in Figure 2-1. The color of light depends on its wavelength. Violet light has the shortest wavelength of all visible colors, and red light has the longest, as seen in Figure 2-2a. White light is a combination of all visible colors or wavelengths.

Laser light, unlike ordinary light, is unique in that it usually consists of only a single color of light. The wavelengths of this single color, while not perfectly uniform in length, occur within a very narrow range. **Monochromaticity** is the laser light's property of containing only one color.

The light from an ordinary source like a light bulb radiates away from the source in all directions, as shown in Figure 2-2b. This spreading or divergence of ordinary light is what makes it so useful for lighting homes and workplaces.

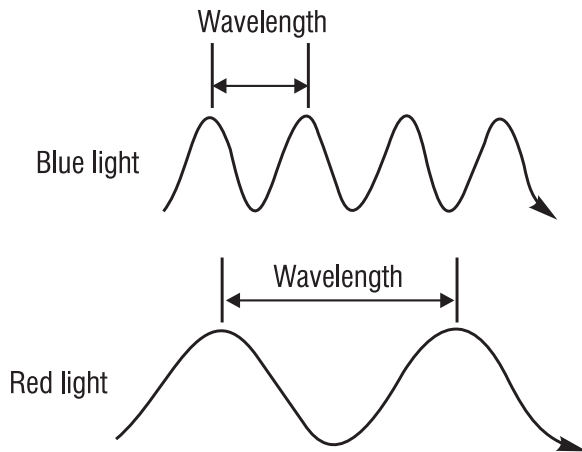


Figure 2-1 *Light waves*

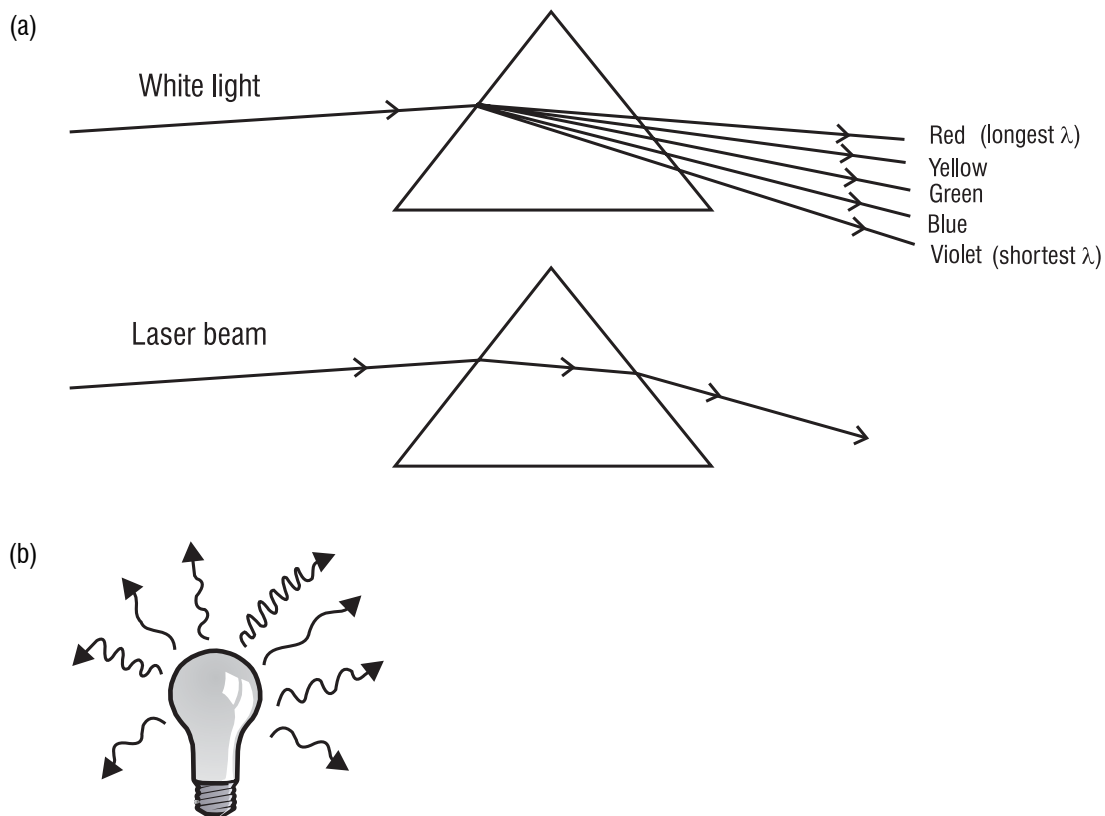


Figure 2-2 *White light, laser light, and light emitted by an incandescent bulb*

Laser light, however, is very directional. By comparison, the light emitted by a laser diverges very little, as shown in Figure 2-3.

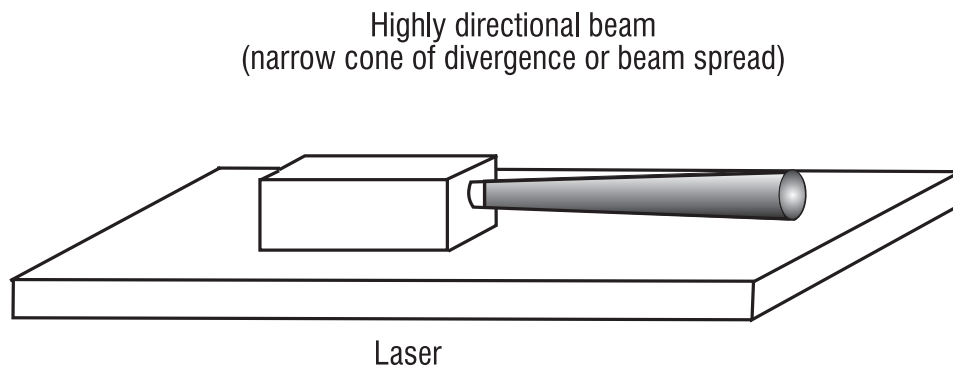


Figure 2-3 Concentrated directionality of laser light

Noncoherent light waves produced by ordinary sources don't form an orderly pattern. They combine in a random fashion and don't produce a wave larger than any of the single various waves, as seen in Figure 2-4a. Such light is said to be *incoherent*. By contrast, the waves produced by a laser travel through space in step with one another. They're said to be "in phase" or *coherent*. When all the separate waves in the beam remain in step with one another, as in Figure 2-4b, the result is a wave much stronger than that of any single wave. A very intense, coherent beam is generated.

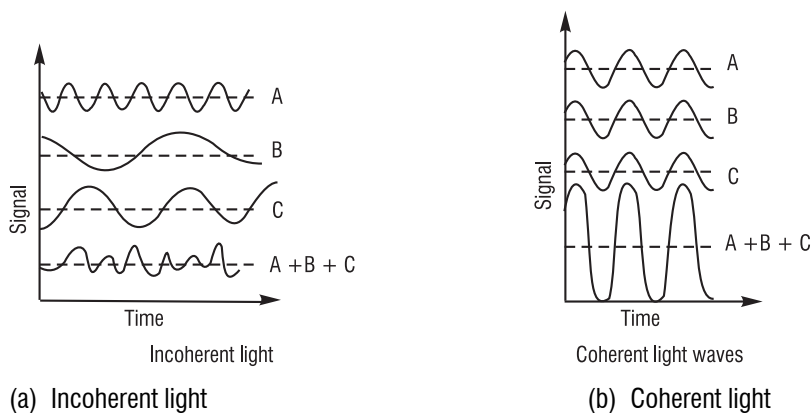


Figure 2-4 Adding coherent and incoherent light waves

The following nonlaser light sources will be discussed:

1. Incandescent sources
2. Fluorescent-low pressure discharge lamps
3. High-intensity and -pressure discharge lamps (HID)
4. Flashlamps and arc lamps
5. Light-emitting diodes (LED)

A. Incandescent Sources

Incandescent sources are similar to but not exactly as intense as blackbody radiators.

A blackbody radiator is considered to be an almost perfect emitter. The first incandescent lamps in the nineteenth century were carbon-, iron-, osmium-, or tantalum-filament lamps. However, in spite of its low ductility tungsten has replaced all these filament materials because of its low vapor pressure, high melting point (3655 K), and strength. Today tungsten is typically alloyed with metals such as thorium and rhenium.

Most modern incandescent lamps are filled with gas to increase the lifetime of the filament. These fill gases are generally mixtures of argon and nitrogen, with high percentages of argon for low-voltage lamps and very high percentages of nitrogen for high-voltage projection lamps. Occasionally krypton is added for still greater lifetime. None of these gases appreciably influences the spectral quality of the incandescent source. However, the tungsten-halogen-type lamp has become increasingly common. Tungsten-halogen lamps contain a small amount of a halogen, such as bromine, chlorine, or iodine. The halogen teams with the tungsten to create a regenerative cycle—particles of tungsten thrown off by the filament combine with the halogen to form a gas that is attracted to the hot filament and attaches to the filament. However, the lost particles of tungsten do not redeposit in exactly the same place, so the filament is modified, spotting and eventually failing as before. The important feature of the tungsten-halogen lamps is that the particles of tungsten collected by the filament are prevented from depositing on the glass. Thus, the lamps do not form a black coating on the inner surface.

Halogen lamps are hot, and they must run hot to keep the regenerating cycle going—nothing less than 500 degrees F will do. For comparison, a 100-watt household lamp is never hotter than about 450 degrees, meaning that this temperature is too low for a halogen lamp. Ordinary glass does not stand the higher temperature needed, so all halogens are made of special heat-resistant glass or of quartz. In fact, the halogen lamp is now perhaps better known as a quartz lamp.

Lamp filaments made of tungsten wire are sometimes used in a straight length but more often formed into a coil, or a coiled coil, Figure 2-5. This drawing shows the various common filament styles. The better (and more expensive) filaments have close-spaced coils to secure a bright, uniform light. Less expensive are the wide-spaced filaments, and these are entirely practical for many applications, such as slide projectors.

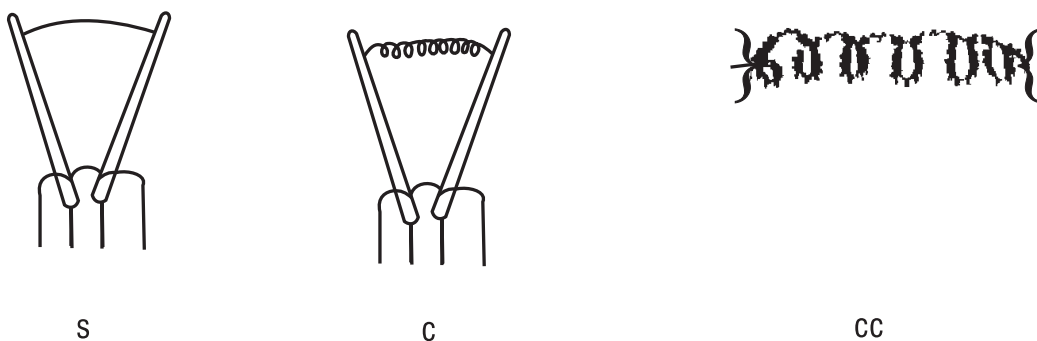


Figure 2-5 Types of incandescent filaments. Filaments are arranged upon support structures within the bulb in a variety of ways. The filament itself may be either straight (s), coiled (c), or coiled coil (cc).

A high-intensity lamp uses a low-voltage automotive 6- or 12-volt bulb. Low-voltage lamps can be used on standard household current by using a small and inexpensive transformer to reduce the usual 110 volts to the required level. With the transformer concealed, the end product is useful for spotlighting.

B. Fluorescent Light Sources

Fluorescent light sources are low-pressure discharge lamps with a fluorescent phosphor. Most fluorescent lights consist of mercury discharge lamps that emit 90% of their energy at 253.7-nm wavelength. These ultraviolet photons can excite a number of phosphors, producing a range of wavelengths from infrared to ultraviolet. Visible wavelengths are characterized as white, warm white, cool white, etc. Fluorescent lamps can have either cold or hot cathode electrodes. Cold electrodes are used when a rapid start is necessary. Hot cathode electrodes give greater luminous efficiency and because of this most lamps today use hot cathode electrodes. The luminous efficiency of a fluorescent lamp increases with its length. For example, an 80-inch lamp is 40% more efficient than a 15-inch lamp. The output of a fluorescent lamp is also a function of the type of ballast used. The two common types of ballast are the rapid-start and the preheat ballast. Most of today's lamps operate with rapid-start ballasts. However, research has suggested that the preheat ballast may yield up to 20% higher light output as compared to the rapid-start type. Lamp life could also be significantly affected by the choice of ballast.

C. High-Intensity Discharge Lamps (HID)

High-intensity discharge (HID) lamps can be made of mercury, sodium, or metal halides. The gas pressures inside the lamps are usually 2–4 atmospheres.

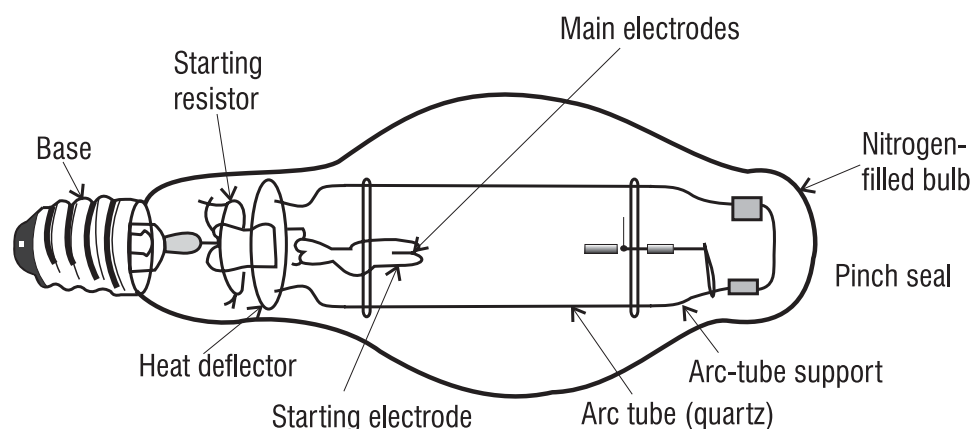


Figure 2-6 Mercury HID lamp. This cutaway drawing shows the construction of an Hg-HID lamp. The inner envelope is of quartz; the outer envelope of borosilicate glass (adapted from Kaufman and Christensen, 1972).

HID lamps—specifically those containing mercury—have two envelopes: the inner quartz discharge tube and an outer glass jacket. The outer jacket absorbs the UV radiation generated by the internal operation of the bulb. In places like gymnasiums, high-bay industrial areas, and public buildings, these lamps are often not enclosed in protective fixtures. The outer jacket can be broken in some instances by flying projectiles without damaging the inner discharge tube. If

the lamp does not extinguish, substantial levels of UV radiation will be emitted to the outside. An HID lamp can cause serious skin burn and eye inflammation from shortwave ultraviolet radiation if the outer envelope of the lamp is broken or punctured. Lamps that will automatically extinguish when the outer envelopes are broken or punctured are commercially available.

D. Flashlamps and Arc Lamps

Flashlamp and arc lamps are high-intensity discharge devices commonly used in photography and laser technology and usually contain gases such as xenon and krypton. The flash or arc is initiated by a high voltage across the discharge tube, which in turn ionizes the gas and produces a high-intensity light with output peaks in both the visible and infrared regions of the electromagnetic spectrum. See Figure 2-7.

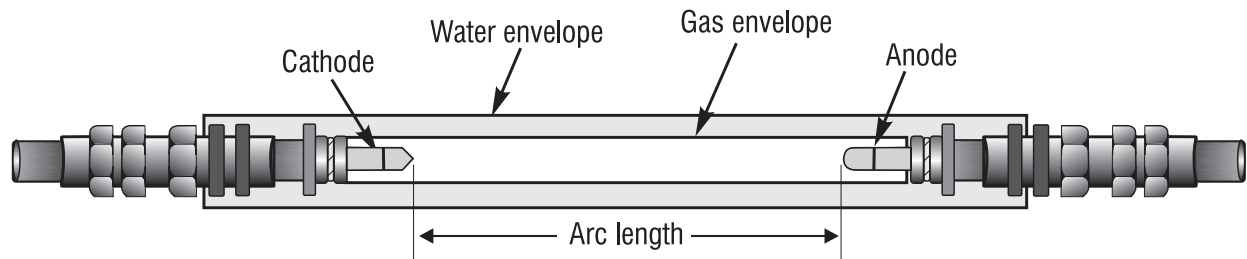


Figure 2-7 Diagram of a gas arc lamp with water jacket for cooling

The quartz tube is usually 1.5 mm in thickness with a metal electrode at each end. Water jackets are provided for cooling purposes.

E. Light-Emitting Diodes (LED)

Light-emitting diodes are semiconductor devices that are directly modulated by varying input current. They are usually made of aluminum-gallium-arsenide (AlGaAs). However, other dopants may be added to vary wavelength. LEDs are common in fiber-optic communication, pocket calculators, and other visual displays. These devices can emit light in both the visible and infrared regions of the spectrum. Unlike a semiconductor laser (a laser pointer for example), a light-emitting diode spews light in all directions and has low irradiance. Power from LEDs generally is in the microwatt range up to maybe a few milliwatts. LEDs are small in size, low temperature, rugged, and inexpensive devices. They operate with *pn* junctions (see Figures 2-8a and 2-8b). In a *pn* junction, two slabs are put together with the *n* material having an excess of electrons and the *p* material having a deficiency of electrons or an excess of holes. Each time an electron falls into a hole (recombination), a photon of light is emitted (see Figure 2-8c). The emitted photons produce an incoherent beam of light. When current flows across a *pn* junction, free electrons from the *n*-type material are forced to combine with holes in the *p*-type material and release energy. The entire device is no larger than a grain of sand.

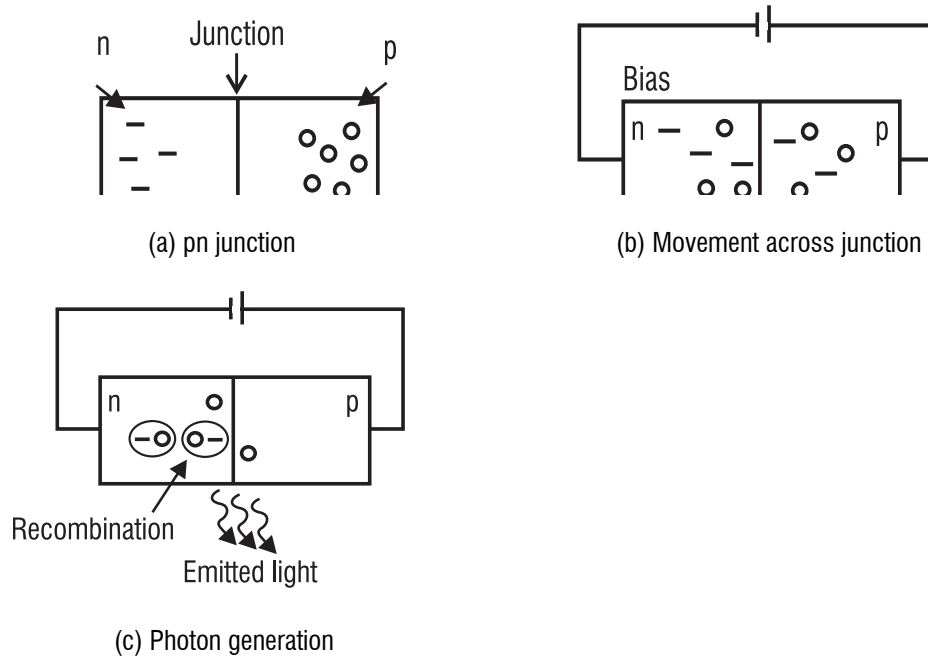


Figure 2-8 Simplified theory of LED operation

II. CONCEPTS OF LASER SAFETY

The use of lasers by industry and by the academic community continues to increase. Many educational institutions are using a wide variety of lasers in many different ways. Traditional disciplines in universities, colleges and high schools, such as biology, chemistry, and physics, now recognize the laser as an essential teaching element. Unique environments associated with educational institutions (such as civil engineering, earth and planetary sciences, and biomedical research) have also incorporated lasers into their educational processes. The objective of this section is to provide reasonable and adequate guidance for the safe use of lasers by evaluating and minimizing hazards associated with laser radiation in educational and workplace environments. The hazard evaluation procedure used is based on the ability of the laser beam to cause biological damage to the eye or skin during intended use, and is related to the classification of the laser or laser system from Class 1, considered to be nonhazardous, to Class 4, very hazardous. Lasers or laser systems are certified by the manufacturer for the specific hazard class in accordance with the Federal Laser Product Performance Standard.

A. Eye Hazards

The eye is the part of the body most vulnerable to laser hazards. Changes to the eye can occur at much lower laser power levels than changes to the skin. And, eye injuries are generally far more serious than injuries to the skin.

The human eye is a complex optical system. It is designed to transmit, focus, and detect light. The normal human eye is roughly the shape of a slightly elongated sphere. It's about the size of a quarter (25 mm, or about 1 inch) in diameter. The outer white layer of the eye is called the

sclera. The sclera—with the aid of the internal fluids (vitreous humor and aqueous humor)—helps to maintain the shape of the eye.

Light passes into the front portion of the eye through the **cornea**. The light that enters is focused to a spot on the back of the eye, the **retina**. There it forms an image on cells especially designed for light detection. Sensitive nerve cells relay the optical image—in the form of electrical signals—to the brain for interpretation. Figure 2-9 shows the essential parts of the human eye. The light irradiance of the image formed on the retina is 100,000 times greater than the light irradiance at the front of the eye. It is this considerable *optical gain* that creates an eye hazard when stray laser beams enter the eye.

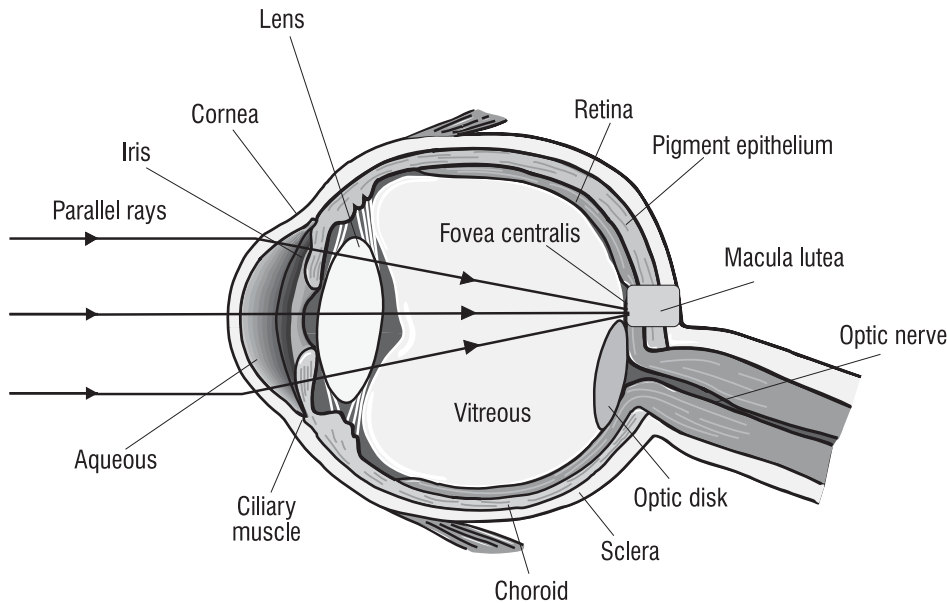


Figure 2-9 Schematic diagram of the eye

The **cornea** is the outermost, transparent layer. It covers the front of the eye. The cornea can withstand dust, sand, and other assaults from the environment. That's partly because corneal cells replace themselves in about 48 hours. Thus, mild injuries to the cornea are healed quickly.

The **aqueous humor** is a liquid (mostly water) between the cornea and the lens. The water in the aqueous humor absorbs heat, so it protects the internal portion of the eye from thermal (heat) radiation. The index of refraction is approximately 1.33, same as water.

The **lens** of the eye is a flexible tissue that changes shape. In conjunction with the cornea, the lens focuses light on the back of the eye. When the lens changes shape, its focal length changes. This lets the eye focus on both near and far objects.

The **iris** controls the amount of light that enters the eye. The iris is the pigmented or colored part of the eye. It responds to light intensity by adjusting its size. The change in iris size adjusts pupil size and controls the amount of light admitted to the eye.

The **pupil** is the opening in the center of the iris through which light passes. The size of a pupil changes from about 2 mm to 7 mm, according to the brightness of light in the environment. The

darker the environment, the larger the pupil. A fully dilated pupil (expanded to admit the greatest amount of light) is considered to be about 7 mm.

The **vitreous humor** is a colorless gel that fills the large area at the center of the eyeball. The vitreous humor helps to maintain the shape of the eye.

The **retina** is the light-sensitive layer located at the back of the eye. Think of it as a sort of viewing screen on which the lens of the eye focuses an image. The retina contains two types of photoreceptor (light-receiving) cells: rods and cones. These cells convert the optical image produced by the lens into electrical signals. The signals then are transmitted to the brain. The **fovea** is the most sensitive, central part of the retina. It's the area responsible for the most detailed vision. A foveal lesion caused by laser radiation is a worst-case scenario for vision. The **optic nerve** carries electrical impulses from the retina to the brain.

B. Absorption of Radiation by the Eye

Certain areas of the eye absorb more light in one spectral region than in other regions. Absorption of laser radiation above a certain level leads to tissue injury. Figure 2-10 shows some absorption characteristics of the eye for different electromagnetic radiation wavelengths. Shorter wavelengths of ultraviolet (UV-C and most UV-B), which range from 315 nm to 280 nm for UV-B and 280 nm to 100 nm for UV-C, are absorbed primarily in the cornea (see Figure 2-10a). Longer wavelengths of ultraviolet light (UV-A), which range from 400 nm to 315 nm, are absorbed principally in the lens of the eyeball (see Figure 2-10b).

Radiation in the visible and IR-A (400 nm to 1400 nm—Figure 2-10c) is the most hazardous and is transmitted by the optical components of the eye. It eventually reaches the retina, where most of the radiation is absorbed in the retinal pigment epithelium and in the choroid, which is a dark brown layer with exceptionally large blood vessels and high blood flow rate. Some infrared radiation in the IR-A, which ranges from 700 nm to 1400 nm, and IR-B, which ranges from 1400 nm to 3000 nm, is absorbed by the lens. Far-infrared radiation, which ranges from 3000 nm to 1 mm, is absorbed primarily by the cornea. Table 2-1 identifies different wavelength regions coded by symbols from IR to UV.

Table 2-1

Laser Wavelength Regions
IR-C = 1 mm to 1400 nm
IR-B = 3000 nm to 1400 nm
IR-A = 1400 nm to 700 nm
Visible light = 700 nm to 400 nm
UV-A = 400 nm to 315 nm
UV-B = 315 nm to 280 nm
UV-C = 280 nm to 100 nm

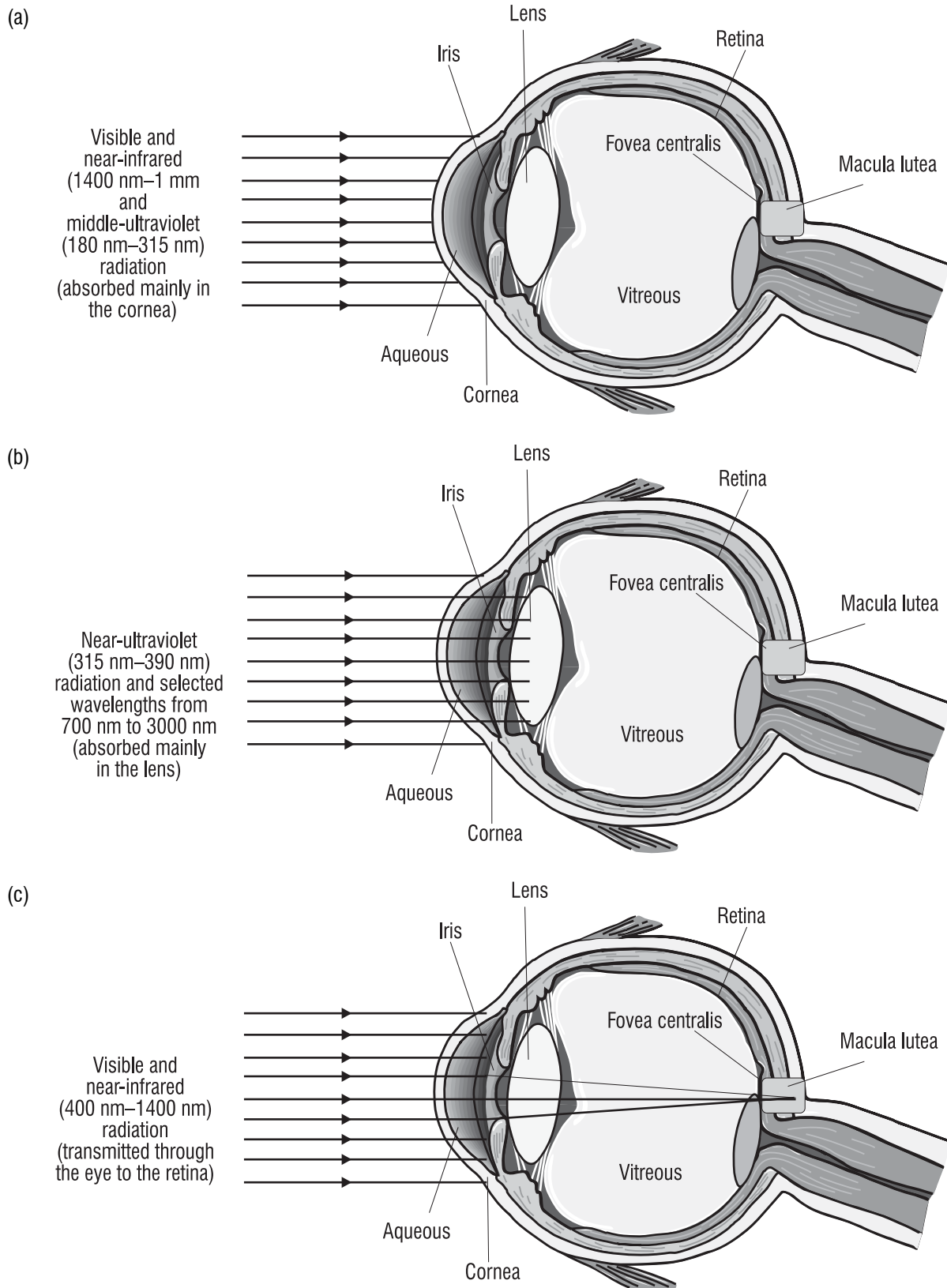


Figure 2-10 Absorption characteristics of the eye for different wavelengths

Figure 2-11 gives a representation of the electromagnetic spectrum and indicates the position of IR, visible, and UV as related to other wavelengths.

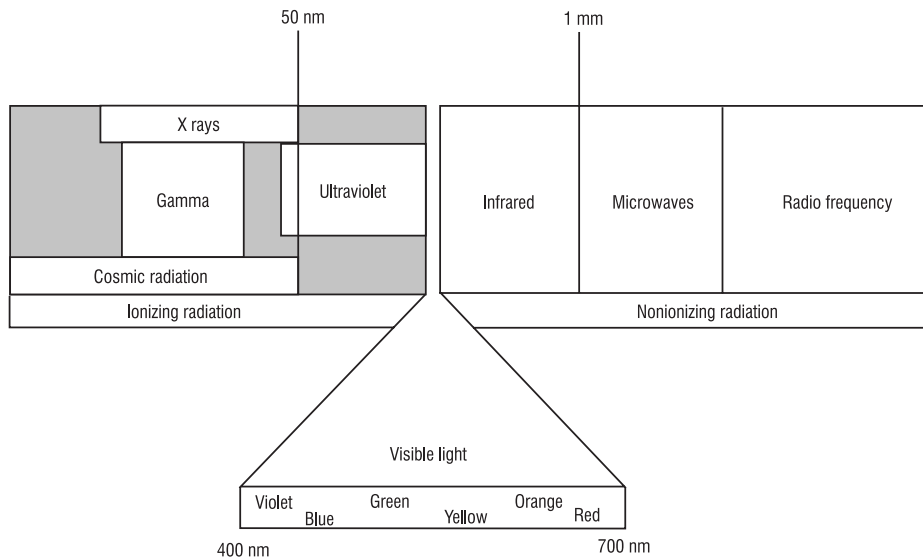


Figure 2-11 Important regions of the electromagnetic spectrum

An important quantity of measure related to laser damage to the eye is retinal irradiance (E), defined as power per unit area, usually expressed in W/cm^2 or mW/cm^2 .

$$E = \frac{\text{power (W)}}{\text{area (cm}^2\text{)}} \text{ or } E = \frac{\text{power (mW)}}{\text{area (cm}^2\text{)}} \quad 2-1$$

Irradiance is much higher at the retina than at the cornea or lens. That's because the laser light is focused by a strong converging lens to a small area on the retina, as shown in Figure 2-12. This accounts for the optical gain discussed earlier.

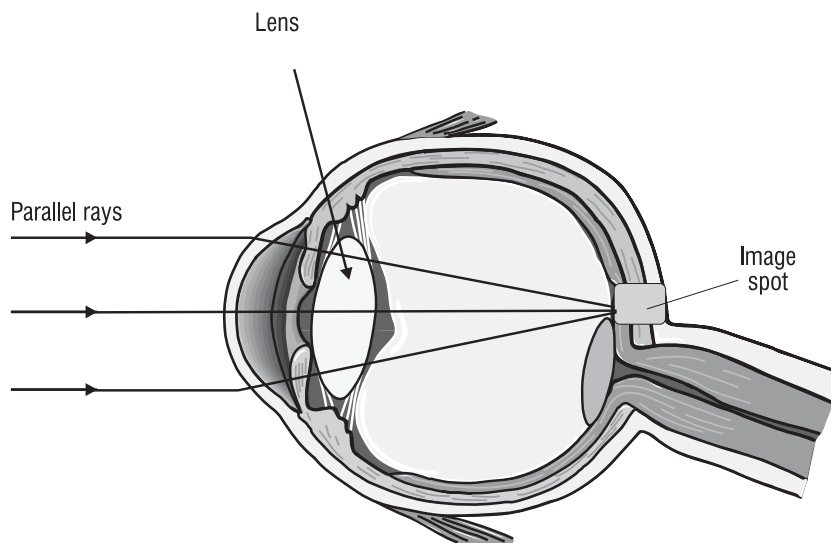


Figure 2-12 Focusing of parallel rays to a spot on the retina

The reduced size of the irradiated area means higher values of irradiance. In the condition called **intrabeam viewing**, shown in Figure 2-13, the focusing of the incident beam causes a tremendous increase of irradiance. Example 1 takes you through the calculation of the irradiance at the retina.

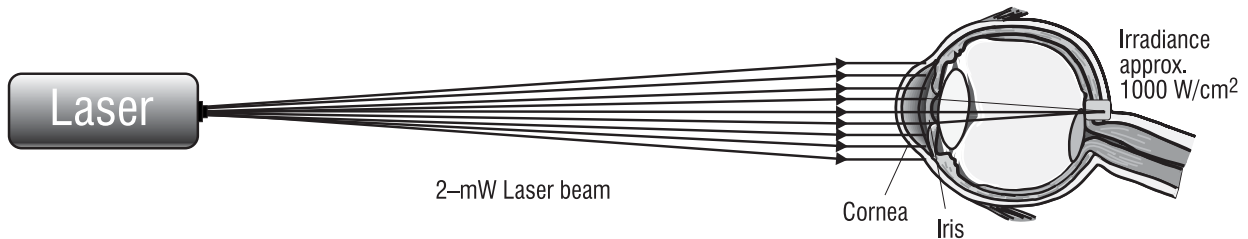


Figure 2-13 Intrabeam viewing conditions

Example 1

Irradiance of a Focused Laser Pointer Beam on the Retina

A laser pointer produces a 2-mW beam. The beam enters the eye and is focused by the cornea and lens to a spot on the retina 16 μm in diameter.

Given: $P = 2 \text{ mW}$

Retinal spot diameter = 16 μm

Find: The irradiance on the retina, assuming that all of the 2 mW of power is focused on the retina.

Solution: Area of spot:

$$A = \frac{\pi d^2}{4}$$

$$A = \frac{(3.14)(1.6 \times 10^{-3} \text{ cm})^2}{4}$$

$$A = 2 \times 10^{-6} \text{ cm}^2$$

Irradiance:

$$E = \frac{P}{A}$$

$$E = \frac{2 \times 10^{-3} \text{ W}}{2 \times 10^{-6} \text{ cm}^2}$$

$$E = 1000 \text{ W/cm}^2$$

Example 2

A 1-mW HeNe laser beam with a divergence of 0.5 mrad enters the eye. Find the irradiance on the retina if the focal length of the eye, from cornea to retina, is equal to $f = 1.7$ cm.

Given: $P = 1$ mW

$\theta = 0.5$ mrad

$f =$ focal length of eye focusing system (1.7 cm)

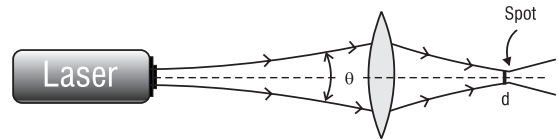
Solution: The solution of this problem requires three steps:

1. Calculate focal spot diameter. (Note: Laser propagation theory shows that, when a laser beam of divergence θ is focused by a lens of focal length f to a spot of diameter d , the spot diameter d is given by $d = f\theta$.)

$$d = f\theta$$

$$d = (1.7 \text{ cm}) (0.5 \times 10^{-3} \text{ rad})$$

$$d = 8.5 \times 10^{-4} \text{ cm}$$



2. Calculate the area of the spot.

$$A = \frac{\pi d^2}{4}$$

$$A = \frac{\pi(8.5 \times 10^{-4} \text{ cm})^2}{4}$$

$$A = 5.7 \times 10^{-7} \text{ cm}^2$$

3. Calculate the retinal irradiance.

$$E = \frac{\text{Power}}{\text{Area}}$$

$$E = \frac{1 \times 10^{-3} \text{ W}}{5.7 \times 10^{-7} \text{ m}^2}$$

$$E = 1.75 \times 10^3 \text{ W/cm}^2$$

C. The Skin

The risk of skin injury by lasers might seem to be greater than the risk of eye damage because we have so much more skin tissue than eye tissue. In fact, potential risks to the skin are considered secondary to risks to the eyes. This is because skin injuries don't often have the dire consequences (such as partial loss of vision) that eye injuries do. Skin injuries may affect only the external dead layer of the skin cells; and even more penetrating damage usually will heal eventually. Large-area exposures that can lead to serious skin infections and other complications are not commonly associated with laser use because the beam is relatively small.

Laser radiation can affect the skin thermally or photochemically. To understand the effects of optical radiation on the skin, some knowledge of the anatomy of the skin is helpful.

Skin consists of two main layers: the **epidermis** (surface layer) and the **dermis** (the underlying layer). The layers are shown in the greatly enlarged diagram in Figure 2-14.

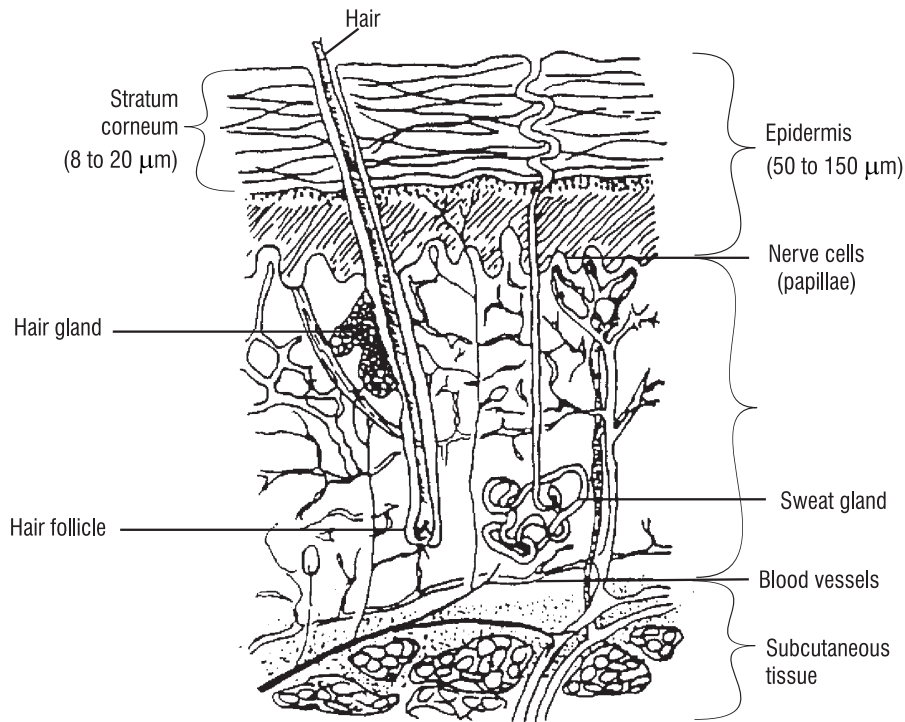


Figure 2-14 *Anatomy of the skin*

The outermost layer of the epidermis is the **stratum corneum**. It consists mainly of dead cells. The stratum corneum gives protection against water loss, abrasion, dust, air, and radiant energy. This layer of dead cells is about 8 μm to 20 μm thick.

In the layer of the epidermis just below the stratum corneum, there are specialized cells that produce melanin pigment granules. Once produced, these granules migrate throughout the epidermis. They help to protect the dermis against harmful ultraviolet radiation. As they absorb radiation, they darken. So they're responsible for the skin's color change during suntanning. The entire epidermis averages about 0.1 mm in thickness.

The **dermis** is the site of many specialized cells and glands. It contains mostly connective tissue that gives elasticity and support to the skin. The dermis also has numerous blood vessels, nerve cells, sweat glands, and hair follicles. Sweat glands have an essential role in regulating body temperature. Sweat allows for evaporative cooling of the body and, in this way, gets rid of excess heat. Nerve cells include heat sensors, cold sensors, pain sensors, and tactile (touch) sensors. Blood vessels allow for the maintenance of healthy tissue and play an important role in heat regulation.

The skin reflects most visible and IR-A (near-infrared) radiation. The epidermis is highly absorbing at UV-B and UV-C wavelengths and at IR-B and IR-C wavelengths. The skin is less

sensitive to UV-A than to UV-B or UV-C. The melanin granules of the epidermis absorb much of the ultraviolet radiation incident upon the skin and, as mentioned, thus protect the dermis from harmful UV. However, with enough power and duration, the incident radiation of any wavelength in the optical spectrum can penetrate the protective filter of the epidermis and cause deep internal injury.

Ultraviolet radiation in the UV-B and UV-C regions causes reddening (sunburn) and eventual tanning—known as erythema—of the skin. Erythema is a photochemical occurrence. Chronic exposures to UV are known to accelerate aging of the skin. They increase the risk of skin cancer. We don't know the precise wavelength range that may lead to the occurrence of cancer, but UV-B has been implicated by some studies.

Laser-induced thermal change to the skin is most pronounced at far-infrared wavelengths such as are produced by CO₂ lasers. Thermal damage also can be caused by visible and near-infrared wavelengths, but at higher irradiance values than for far-infrared laser beams.

The pain from thermal injury to the skin by most lasers is enough to alert the user to move out of the beam path. However, a number of high-power visible and IR lasers now are used. They are capable of producing significant burns to the skin in much less than one second. And users are unable to move out of the beam path before serious injury occurs.

D. Laser Safety Standards and Safety Classifications

The primary laser safety standard in use today is the ANSI (American National Standards Institute) Z-136.1. A new laser safety standard for educational institutions ANSI A-136.5 should be available in fall 1999. Both standards can be obtained from the Laser Institute of America in Orlando, Florida. ANSI is a nonprofit organization for which expert volunteers participate on committees to set standards in various fields. In 1969, ANSI addressed the issue of laser safety at the request of the U.S. Department of Labor. The resulting standard (ANSI Z-136), issued in October 1973, was directed primarily to users: operators, technicians, laser safety officers safety engineers, and research personnel. The standard was rewritten in 1980 and 1986 and revised in 1993. It provides guidance for the safe use of lasers and laser systems by defining control measures for each of the four broad laser classifications. However, technical information on measurements, calculations, and biological effects also are provided for those who want to carry out their own measurements and calculations.

LASER HAZARD CLASSIFICATIONS

The most important criterion you will use in applying laser safety control measures is the hazard classification designated by manufacturers on the equipment labels. Certain controls are required for each class (except Class 1) listed below:

Class 1 (exempt lasers) cannot, under normal operating conditions, emit a hazardous level of optical radiation. Included in this category is laboratory equipment using lasers with all beam paths and reflections enclosed.

Class 2, or low-power visible laser device of 1 milliwatt, does not have enough output power to injure a person accidentally, but may injure the eye when stared at for a long period. A “caution” label must be on the device. Many HeNe lasers are Class 2 but not all. These lasers are used for alignment procedures and optical experiments.

Class 3a lasers—rated in power from 1 milliwatt to 5 milliwatts—cannot injure a normal person when viewed with the unaided eye but may cause injury when the energy is collected and put into the eye as with binoculars. Most laser pointers fall into this category. A danger or caution sign must label the device, depending on its irradiance.

Class 3b lasers from 5 milliwatts to 500 milliwatts can produce eye injury when viewed without eye protection. This class of laser requires a danger label and could have dangerous specular reflections. Eye protection is required.

Class 4 lasers above 500 milliwatts in power can injure you if viewed directly or by viewing both the specular and diffuse reflections of the beam. A danger sign will label this laser. These lasers can also present a fire hazard. Eye and skin protection is required.

TYPES OF CONTROLS: ENGINEERING, ADMINISTRATIVE, PERSONAL PROTECTIVE

The hazards posed by Class 3b and Class 4 lasers require systematically applied controls to prevent skin and eye injuries. First, we will examine the types of controls available. Control measures may be broken down into three main types: engineering controls, administrative controls, and personal protective equipment.

Engineering controls involve design features or devices applied to the laser, laser beam, or laser environment that restrict exposure or reduce irradiance. Such controls include beam shutters, beam attenuators, remote firing and monitoring systems, and the protective housing placed entirely around some laser systems.

Administrative control measures involve procedures and information rather than devices or mechanical systems. Some important administrative controls are posting of warning signs and labels, establishment of standard operating procedures (SOP), and safety training.

Personal protective equipment is worn by personnel using the laser or in the vicinity of the laser. It includes protective eyewear, gloves, and special clothing. Table 2-2 lists some of the more commonly used control measures for Class 3b and Class 4 lasers and laser systems.

Table 2-2. Control Measures for Class 3b and Class 4 Lasers

Engineering Controls	Administrative and Procedural Controls	Personal Protective Equipment
Protective housing and service panel Interlocks on the protective housing Door interlocks and remote-control connector Beam attenuators and beam shutters Key switch or padlock Filtered viewing optics and windows Emission delay Warning lights, emission indicators Beam enclosure Controlled beam path Laser controlled area Beamstops Remote firing and/or monitoring	Laser safety officer Standard operating procedures Limitations on use by class Entry limitations for visitors Education and training Maintenance and service manuals Marking of protective devices Warning signs and labels	Eyewear Clothing Gloves

* David Sliney and Myron Wolbarsht, *Safety with Lasers and Other Optical Sources: A Comprehensive Handbook* (New York: Plenum Press, 1980), p. 512.

MAXIMUM PERMISSIBLE EXPOSURE (MPE)

How much exposure to laser light is hazardous? To answer this question, you have to take into account the output characteristics of the laser. Those characteristics include wavelength, output energy and power, size of the irradiated area, and duration of exposure. If you're using a pulsed laser, you also must consider the pulse repetition rate.

Sensitivity to a given wavelength varies significantly from person to person. **Maximum permissible exposure (MPE)** limits indicate the greatest exposure that most individuals can tolerate without sustaining injury.

An MPE is usually expressed in terms of the allowable exposure time (in seconds) for a given irradiance (in watts/cm^2) at a particular wavelength. Table 2-3 gives the maximum permissible exposure for the eye for a variety of lasers operating at different irradiance levels. For example, Table 2-3 indicates that, for an argon laser operating at $0.514 \mu\text{m}$, the MPE is $2.5 \times 10^{-3} \text{ W/cm}^2$ for an exposure time of 0.25 second, $16.7 \times 10^{-6} \text{ W/cm}^2$ for an exposure time of 600 seconds, and $1 \times 10^{-6} \text{ W/cm}^2$ for an exposure time of 3×10^4 seconds. MPEs are useful for determining optical densities for eyewear or windows.

Table 2-3. MPE for Selected Lasers and Exposure Times
(Reference: ANSI Z136.1-1993)

Laser Type	Wavelength (μm)	MPE (average power density—watts/cm ²)			
		Exposure time in seconds			
		0.25 s	10 s	600 s	3×10^4 s
CO ₂	10.6	—	0.1 W/cm ²	—	0.1 W/cm ²
Nd:YAG (cw) ^a	1.33	—	5.1×10^{-3}	—	1.6×10^{-3}
Nd:YAG (cw)	1.064	—	5.1×10^{-3}	—	1.6×10^{-3}
Nd:YAG Q-switched ^b	1.064	—	17×10^{-3}	—	2.3×10^{-6}
GaAs (diode)	0.840	—	1.9×10^{-3}	—	610×10^{-6}
InGdAlP (diode)	0.670	2.5×10^{-3}	—	—	—
HeNe	0.633	2.5×10^{-3}	—	293×10^{-6}	17.6×10^{-6}
Krypton	0.647	2.5×10^{-3}	—	364×10^{-6}	28.5×10^{-6}
	0.568	2.5×10^{-3}	—	31×10^{-6}	18.6×10^{-6}
	0.530	2.5×10^{-3}	—	16.7×10^{-6}	1.0×10^{-6}
Argon	0.514	2.5×10^{-3}	—	16.7×10^{-6}	1.0×10^{-6}
XeFl ^c	0.351	—	—	—	33.3×10^{-6}
XeCl ^c	0.308	—	—	—	1.3×10^{-6}

^a Operating at less common 1.33 μm

^b Pulsed operation at 11 Hz, 12-ns pulse, 20-mJ/pulse

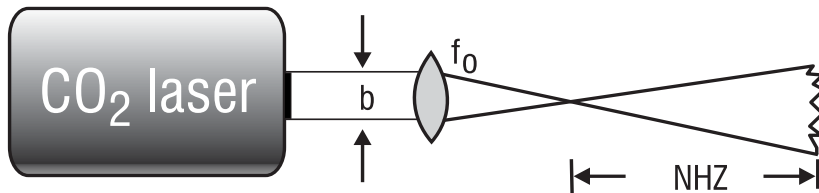
^c When repeated exposure levels are anticipated the MPE level must be reduced by a factor of 2.5.

Another quantity of interest in laser safety is called the **nominal hazard zone (NHZ)**. This zone describes the region within which the level of direct, reflected, or scattered (diffuse) laser radiation is above the allowable MPE. Example 3 shows how to calculate an NHZ for a given laser and given MPE. Table 2-4 lists nominal hazard zones for three common lasers, for a direct laser beam, a laser with a focusing lens attached, and diffusely scattered laser light. The data used in the NHZ calculations for the three lasers listed in Table 2-4 are given in Table 2-5.

Example 3

Nominal Hazard Zone Calculation

A CO₂ laser with a power P of 500 W and exit beam diameter b out of the laser of 3 cm is focused by a convex lens of focal length $f_0 = 20$ cm. Determine the nominal hazard zone (NHZ) from the focal point of lens. (Note: From laser safety theory, it is shown that the NHZ for a laser of given power and exit beam diameter is given by the formula $\text{NHZ} = \frac{f_0}{b} \left(\frac{4P}{\pi(\text{MPE})} \right)^{1/2}$).



Given:

$P = 1000$ W (power)

$f_0 = 20$ cm (focal length of lens)

$b = 3$ cm (beam diameter at exit)

$\text{MPE} = 0.1$ W/cm² (refer to Table 2-3 or Table 2-5)

$$\text{NHZ} = \frac{f_0}{b} \sqrt{\frac{4P}{\pi \text{MPE}}}$$

$$\text{NHZ} = \frac{20}{3} \sqrt{\frac{4 \cdot 500}{3.14 \cdot (0.1)}}$$

$$\text{NHZ} = (6.67) (79.8) = 532 \text{ cm}$$

$$\text{NHZ} = 5.32 \text{ m}$$

Note that this value agrees with the entry in Table 2-4 for a CO₂ laser with an exposure duration of either 8 h or 10 s, with a lens-on-laser.

Table 2-4. Nominal Hazard Zones (NHZ) for Various Lasers

Laser Type	Exposure Duration	Nominal Hazard Distance (m)		
		Direct	Lens-on-Laser	Diffuse
Nd:YAG	8 h	1410	11.4	1.4
	10 s	790	6.4	0.8
CO ₂	8 h	399	5.3	0.4
	10 s	399	5.3	0.4
Argon	8 h	25,200	1700	12.6
	0.25 s	505	33.6	0.25

Table 2-5. Laser Criteria Used for NHZ Calculations

Laser Parameter	Nd:YAG	CO ₂	Argon
Wavelength, λ (μm)	1.064	10.6	0.488, 0.514
Beam power, P (W)	100	500	5.0
Beam divergence, ϕ (mrad)	2.0	2.0	1.0
beam size at aperture, a (mm)	2.0	20.0	2.0
Beam size at lens, b (mm)	6.3	30.0	3.0
Lens focal length, f_o (mm)	25.4	200	200
MPE, 8 hours ($\mu\text{W}/\text{cm}^2$)	1.6×10^3	1.0×10^5	1.0
MPE, 10 s ($\mu\text{W}/\text{cm}^2$)	5.1×10^3	1.0×10^5	—
MPE, 0.25 s ($\mu\text{W}/\text{cm}^2$)	—	—	2.5×10^3

ABSORPTION, LAMBERT'S LAW, AND OPTICAL DENSITY

In general, when a light beam is transmitted through a transparent material, some of the light energy is *absorbed* and the beam intensity decreases continuously. The absorption may be minimal, as in air, or considerable, as in a dark piece of glass. In each case, the absorption process removes energy from the beam and converts it to thermal energy of the transparent medium—by increasing the kinetic energy of the atoms and molecules.

If a beam of irradiance E_0 watts/cm² is incident on a slab of material medium as shown in Figure 2-15, the beam irradiance decreases to a value E_x watts/cm² after passing through a thickness x of the transparent medium. The value of irradiance E_x is given by *Lambert's law*.

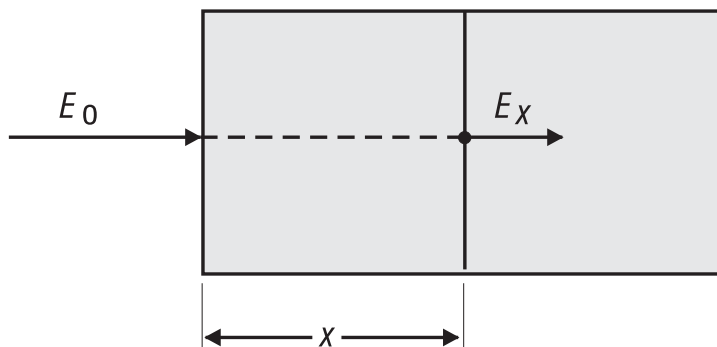


Figure 2-15 Absorption of light in transparent materials

$$E_x = E_0 e^{-\sigma_\lambda x}$$

2-2

where E_x is the irradiance after traveling through a thickness x

E_0 is the irradiance incident on the absorber

σ_λ is the absorption coefficient

x is the absorbing thickness of the absorbing medium

e is the Naperian logarithmic base

The absorption coefficient σ_λ is a property of the absorbing medium and the wavelength of the light. For eye filters (used in laser safety goggles) where the absorption coefficient for the filter is known and the filter thickness is predetermined, the product $\sigma_\lambda x$ is replaced by a quantity called the *optical density*, described by the symbol OD. In that case, with the base 10 replacing the base e , Lambert's law takes the form

$$E_x = E_0 10^{-OD} \quad 2-3$$

And since the transmission of light through an absorber is defined as the ratio of E_x/E_0 , we can rewrite Equation 2-3 in a form used commonly with optical filters

$$T = 10^{-OD} \quad 2-4$$

where OD = optical density

T = transmission

Note that, since $OD = \sigma_\lambda x$, the optical density itself is a wavelength-dependent quantity. Optical density is usually labeled on the temple of the goggles or on the filter itself. Thus, if the OD of a filter is known, the transmission through the filter can be calculated from Equation 2-4. For example, laser safety eyewear with an $OD = 3$ has a transmission T of 10^{-3} or 0.1%. Table 2-6 below gives transmission values for a range of OD values.

Table 2-6. Transmission at Various OD Values

OD	Transmission		%Transmission
0	1	1.0	100%
1	10^{-1}	0.1	10%
2	10^{-2}	0.01	1%
3	10^{-3}	0.001	0.1%
4	10^{-4}	0.0001	0.01%
5	10^{-5}	0.00001	0.001%
6	10^{-6}	0.000001	0.0001%
7	10^{-7}	0.0000001	0.00001%

Suppose the maximum permissible exposure (MPE) for a given laser beam of exit irradiance E_0 is known. Then a filter must be used to reduce the incident irradiance (in W/cm^2) to the permissible irradiance (in W/cm^2). See Figure 2-16.

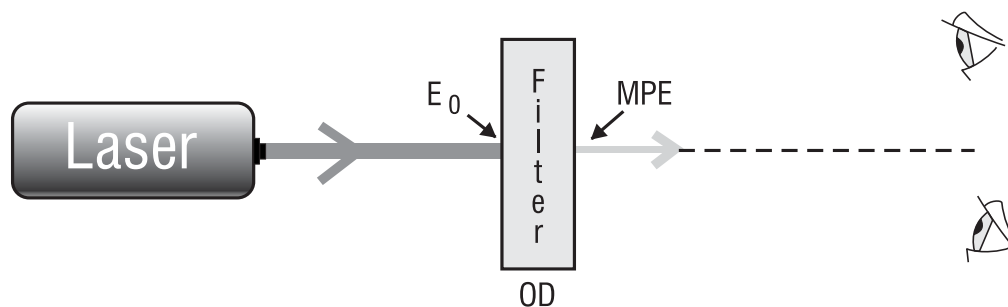


Figure 2-16 Using an optical filter to reduce beam irradiance

If we replace E_x in Equation 2-3 with MPE, we get $MPE = E_0 10^{-OD}$ or, equivalently, $10^{+OD} = E_0/MPE$. Solving the last expression for OD gives another useful equation.

$$OD = \log_{10} \left(\frac{E_0}{MPE} \right) \quad 2-5$$

Example 4 involves a sample calculation for the minimum optical density (OD) required for a specified laser power and recommended MPE.

Example 4

An Nd:YAG laser with a power of 40 watts is projected onto a fully dilated human eye pupil of 7-mm diameter. The eye is exposed for a duration of 10 seconds. Calculate the minimum optical density OD of a laser safety goggle needed to protect the eye from damage. Refer to Table 2-3 for the appropriate maximum permissible exposure (MPE).

Given:

- Nd:YAG
- $\lambda = 1064 \text{ nm}$
- $t = 10 \text{ s}$
- Eye diameter = 7 mm
- $MPE = 0.0051 \text{ W/cm}^2$ (See Table 2-3 for Nd:YAG, 1.064 μm , 10-second exposure.)
- $P = 40 \text{ W}$
- Area of pupil = $\frac{\pi D^2}{4} = \frac{(3.14)(0.7 \text{ cm})^2}{4} = 0.38 \text{ cm}^2$

Solution:

Using Eq 2-5, where $OD = \log_{10} \left[\frac{E_0}{MPE} \right]$ and $E_0 = P/A$, we get:

$$OD = \log_{10} \left[\frac{40 \text{ W} / 0.38 \text{ cm}^2}{0.0051 \text{ W} / \text{cm}^2} \right]$$

$$OD = \log_{10} [20,640]$$

$$OD = 4.3$$

The required optical density for the laser safety goggles would be 4.3 or larger.

Laser safety eyewear is required for Class 3b and Class 4 lasers and laser systems. Eyewear should be comfortable and in good condition, not scratched or discolored. Some filters may be clear and others tinted, depending on the laser wavelength.

LASER SAFETY SIGNS

The most common signs used for lasers and laser systems are the DANGER, CAUTION, and NOTICE signs. See Figure 2-17 below.

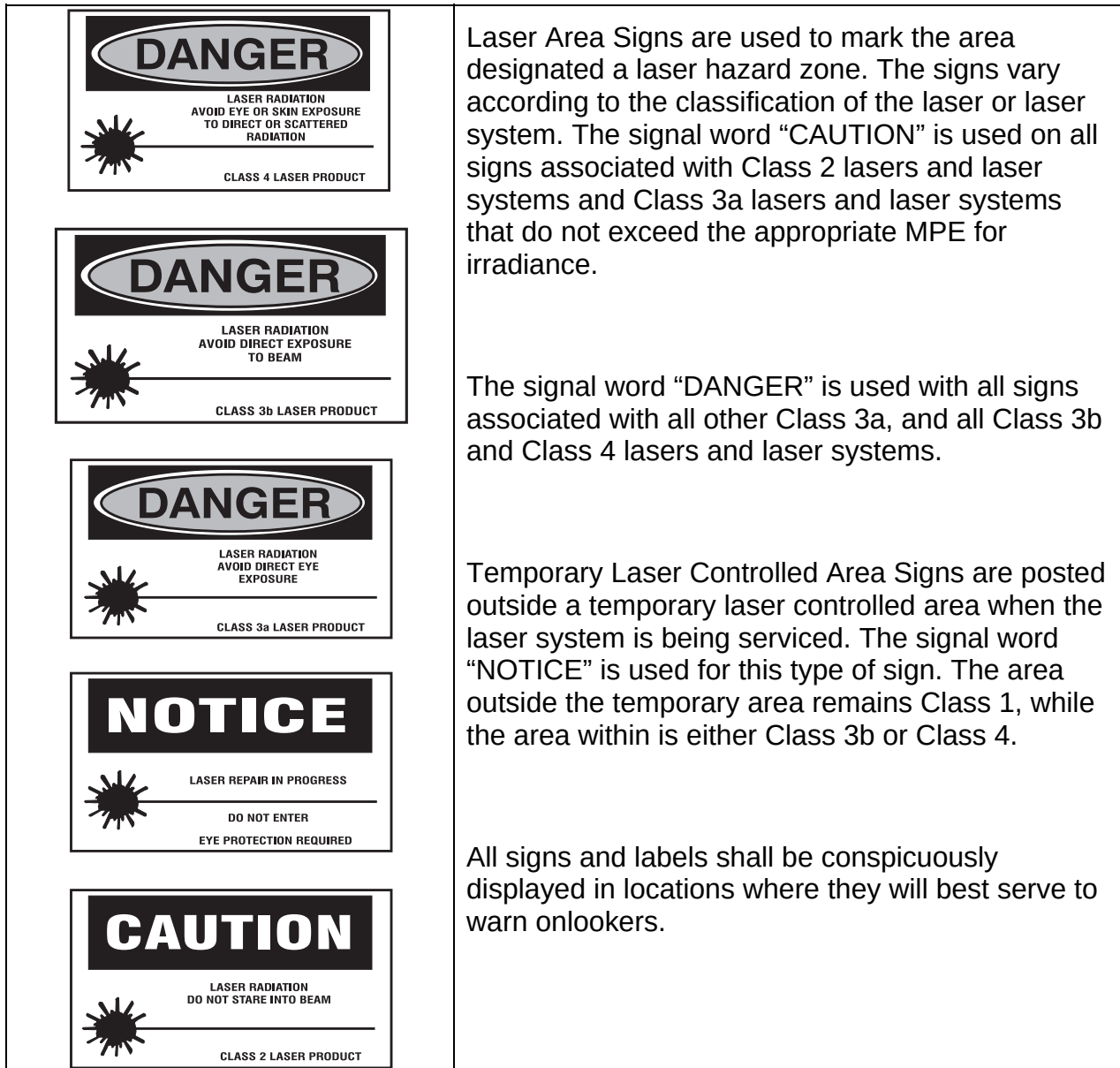


Figure 2-17 *Laser area warning signs for lasers*

Caution signs are used for a Class 2 laser and some Class 3a lasers, depending on their irradiance. Danger signs are used on some Class 3a, Class 3b, and Class 4 lasers and laser systems. A notice sign is used for a temporary laser controlled area while service or repair is being done.

LASER ACCIDENTS

Most laser accidents occur because adequate control measures are not in place: for example, doing alignment procedures without laser safety eyewear or wearing the wrong eyewear for the laser used. Below is a chart that summarizes the causes of most accidents using lasers. See Figure 2-18.

Causes of Laser Accidents Percentage of Occurrence

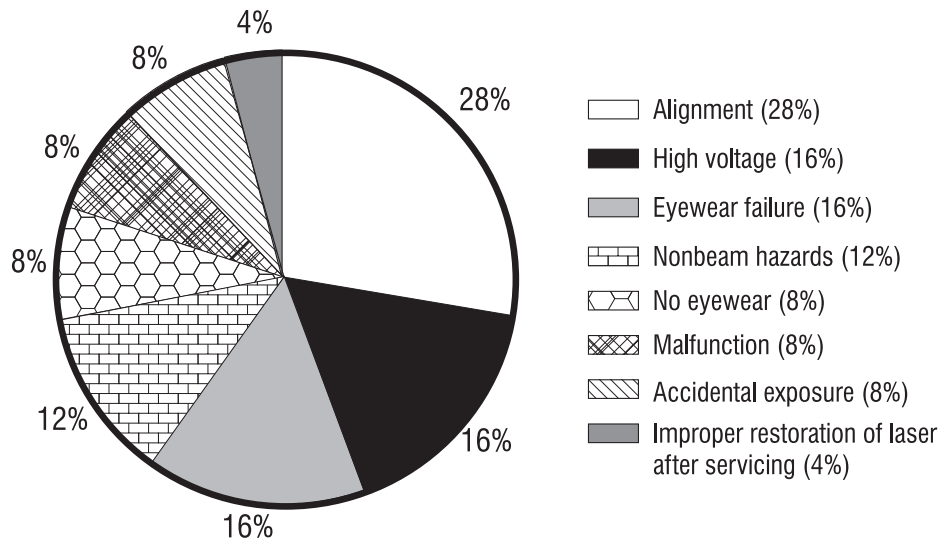


Figure 2-18 *Causes of most laser accidents*

As in all aspects of laboratory, field, classroom, or industrial safety, the best measures are a positive attitude and common sense. The following list of precautions is not all-inclusive for every application, although it does represent a comprehensive collection of thoughts from a number of people who are experienced in the safe use of lasers.

SAFETY RULES FOR ALL LASERS, REGARDLESS OF OUTPUT POWER LEVEL

1. Avoid looking directly into any laser beam or at its reflection.
2. Remove all unnecessary specular (shiny) reflecting surfaces from the work area.
3. Operate lasers in well-defined areas to which access can be controlled. The area should be posted with appropriate signs to alert persons passing by the area that a potential hazard exists.
4. The laser system should be operated only by or under the direct supervision of a person knowledgeable of the hazards and control methods for both beam and nonbeam conditions. This individual is usually the laser safety officer (LSO) who is designated by the administration of the company, hospital, or educational institution. The LSO shall have the authority and the responsibility to effect monitoring and enforce the control of laser hazards and to achieve the knowledgeable control of laser hazards.
5. Any accident should immediately be reported to the responsible medical authority. If there is an accidental exposure to the eye, the services of an ophthalmologist should be sought.

NONBEAM HAZARDS

There are four well-known nonbeam potential hazards associated with the use of lasers and laser systems.

1. Fire hazard
2. Explosion hazard
3. Electrical hazard
4. Chemical hazard

FIRE HAZARD

Class 4 laser systems represent a fire hazard. Enclosure of Class 4 laser beams can result in potential fire hazards if enclosure materials are likely to be exposed to irradiances exceeding 10 W/cm^2 or beam powers exceeding 0.5 W. The use of flame-retardant materials is advisable and necessary.

EXPLOSION HAZARD

High-pressure arc lamps, filament lamps, and capacitor banks in laser equipment shall be enclosed in housings that can withstand the maximum explosive pressure resulting from component disintegration. The laser target and elements of the optical train that may shatter during laser operation shall also be enclosed or equivalently protected to prevent injury to operators and observers. Explosive reactions of chemical laser reactants or other laser gases may be a concern in some cases.

ELECTRICAL HAZARD

The use of lasers or laser systems can present an electric shock hazard. This may occur from contact with exposed utility power use, device control, and power-supply conductors operating at potentials of 50 volts and above. These exposures can occur during laser setup or installation, maintenance, and service, where equipment protective covers are often removed to allow access to active components as required for those activities. Those exposed can be equipment installers, users, technicians, and uninformed members of the public, such as passersby.

The following potential problems have frequently been identified during laser facility audits.

1. Uncovered electrical terminals
2. Improperly insulated electrical terminals
3. Hidden “power-up” warning lights
4. Lack of personnel trained in current cardiopulmonary resuscitation practices, or lack of refresher training
5. “Buddy system” not being practiced during maintenance and service
6. Non earth-grounded or improperly grounded laser equipment
7. Non-adherence to the OSHA lock-out standard (29 CFR 1910.147)
8. Excessive wires and cables on floor that create fall or slip hazards

CHEMICAL HAZARDS

Dye lasers use a complex fluorescent organic compound that, when in solution with certain solvents, forms a lasing medium for dye lasers. Certain dyes are highly toxic or carcinogenic. Since these dyes frequently have to be changed, special care must be taken when handling, preparing solutions, and operating dye lasers.

The use of dimethylsulfoxide (DMSO) as a solvent for cyanine dyes in dye lasers should be discontinued if possible. DMSO aids in the transport of dyes into the skin. If another solvent cannot be found, low-permeability gloves should be worn by personnel any time a situation arises where contact with the solvent may occur.

Laboratory

Irradiance

As discussed in the text, irradiance is defined as power per unit area and one of the most important concepts in regard to laser safety.

Equipment List

The following equipment is needed to complete this laboratory:

HeNe laser (Class 2 or 1 mW)

Photoelectric power meter

Diverging lens

Neutral-density filter, OD = 2.0 (1% transmission)

Sheet of graph paper with millimeter divisions or a metric ruler. A calibrated aperture can also be used to determine beam diameter. This could be a simple hole in an aluminum sheet in the range of 2-6 mm.

Procedure

In this laboratory exercise, you will use an optical power meter to determine the power and irradiance of laser light. The power and area of a HeNe laser beam will be expanded with a diverging lens, and the irradiance will be determined at several distances. You will prepare a graph of irradiance-versus-distance from the lens.

A. Power and Irradiance of a HeNe Laser Beam

1. Observing laser safety procedures, turn on HeNe Class 2 laser.
2. Position the sensor head of the photoelectric power meter about 10 cm from the output aperture of the laser. Turn on the power meter to a full-scale range. Measure the laser beam power and record in Lab Table 1.

- Place a neutral-density (ND) filter in the beam path to reduce its intensity.
Note: Do not handle the ND filter with your fingers since its surface is delicate. Use lens tissue and handle it by the edges only!
- To the right of the ND filter, position a piece of graph paper (10 cm from output aperture of laser) so that the laser beam is incident upon it, or place a nonreflecting meter ruler as vertical as possible. If you use an aperture, you may place that aperture 10–20 cm from the laser.
- Calculate and record the area of the beam and the irradiance of the beam in Lab Table 1.

B. Irradiance of a Diverging Beam

- Observing all safety precautions, turn on the laser.
- Place a diverging lens in the laser beam about 10 cm from the laser output aperture, as illustrated in Figure L-1. Be careful of reflections off the lens when inserting the lens in the beam. Be sure that neither reflections nor the expanded laser beam enters the eye directly.

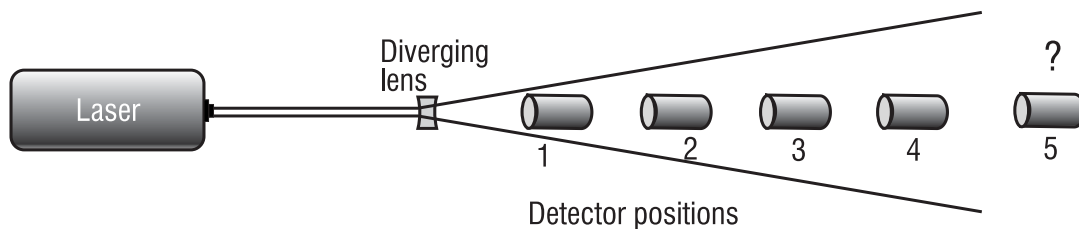


Figure L-1

- Measure the power of the expanded beam at four distances from the diverging lens, as indicated in Figure L-1. The beam should be larger than the detector at the nearest point, and all measurements should be made on the center of the beam. Record distances and corresponding powers in Lab Table 2.
- Calculate the irradiance at each point by dividing the power detected by the receiving area on the detector or the area of the opening of an ambient light shade if one is used.
- Draw a graph of irradiance-versus-distance from the diverging lens.
- Select a fifth distance beyond the fourth measurement. Use your graph to predict the irradiance of the expanded beam distance. Measure the power at the fifth distance, calculate the irradiance, and compare the calculated value with your predicted value.
- Describe the effect a diverging lens has on the irradiance of a laser beam.

Lab Table 1

Power and Irradiance of a HeNe Laser Beam

Laser power using photoelectric power meter:

$P = \underline{\hspace{2cm}}$ mW

Beam diameter: $d = \underline{\hspace{2cm}}$ mm = $\underline{\hspace{2cm}}$ cm

Beam area: $A = \frac{\pi d^2}{4}$

$A = \underline{\hspace{2cm}}$ cm²

Irradiance: $E = \frac{P}{A}$

$E = \underline{\hspace{2cm}}$ mW/cm²

Lab Table 2

Data Table Irradiance of a Diverging Beam

Detector area: $A = \underline{\hspace{2cm}}$

Position	Distance (cm)	Power (mW)	Irradiance (mW/cm ²)
1			
2			
3			
4			

Other Resources

The Laser Institute of America located in Orlando, Florida, provides many resources on laser safety. These resources include laser hazard software packages, laser safety guides for eyewear protection, all of the ANSI (American National Standards Institute) safety standards, and laser safety reference guides for all environments in which lasers are used.

References

1. American National Standard for the Safe Use of Lasers Z-136.1, New York, New York, 1993.
2. Center for Occupational Research and Development. *Introduction to Lasers*, Waco, Texas, 1990.
3. Engineering Technology, Inc. *Laser Safety*, Waco, Texas, 1993.
4. Killen, Harold B. *Fiber-Optics Communications*, Prentice Hall, Englewood Cliffs, New Jersey, 1991.
5. Lacy, Edward A. *Fiber-Optics*, Prentice Hall, Englewood Cliffs, New Jersey, 1986.
6. Popular Optics Library. *Popular Optics*, Edmund Scientific No. 9445, Barrington, New Jersey, April 1997.
7. Rockwell, R. James, Laser Accidents: “Are they all reported and what can be learned from them?” *Journal of Laser Applications*, Oct. 1989.
8. Seeber, F. P., “Laser Practices and Problems in Educational Institutions,” Proceedings, International Laser Safety Conference, Cincinnati, Ohio, 1992.
9. Seeber, F. P., “Now Is the Time for the Safe Use of Lasers in Educational Institutions,” International Laser Safety Conference, Orlando, Florida, 1999.
10. Sliney, D. H., and M. Wolbarsht. *Safety with Lasers and Other Optical Sources*, New York: Plenum Press, 1980.
11. Sliney, D. H., and Stephen L. Trokel., *Medical Lasers and Their Safe Use*, Springer Verlag, 1993.

Problem Exercises

1. Explain the difference between incandescent and fluorescent light sources.
2. Define coherence and monochromaticity.
3. Calculate the irradiance of a laser beam in watts/cm^2 if the power of the laser beam is 500 milliwatts and the diameter of laser beam is 2 mm.
4. Which laser wavelengths are hazards for the following?
 - a. cornea of the eye
 - b. retina of the eye
5. Give the power levels for Classes 2, 3a, 3b, and 4 laser safety classifications.
6. Explain when to use a:
 - a. caution sign
 - b. danger sign
 - c. notice sign

7. Calculate the transmission for the following optical densities:
 - a. OD = 2
 - b. OD = 5
 - c. OD = 6
 - d. OD = 8
8. Explain the difference between engineering controls, administrative controls, and personal protective equipment in regard to control measures.
9. Explain the role of the laser safety officer (LSO) in an industrial environment.
10. Identify the maximum intrabeam permissible exposure (MPE) for the eye:
 - a. CO₂ laser for 10 seconds
 - b. HeNe laser for 0.25 second
 - c. Argon laser for 600 seconds
 - d. XeCl laser for 30,000 seconds
11. Calculate the minimum optical density for protective eyewear using an argon laser of wavelength 0.514 μm , for a given exposure time of 0.25 second, a power of 5 watts, and a worst-case scenario of a dilated eye of 7-mm pupil diameter
12. List the three major causes of laser accidents.
13. Discuss the nonbeam hazards described in this module and suggest procedures and controls to reduce or eliminate them.
14. An argon laser with a power of 15 W and a beam diameter out of the laser of 4 mm is focused for 10 minutes by a converging lens of focal length 15 cm. Determine the NHZ measured from the focal point of the lens.
15. A laser delivers an irradiance of 105 W/cm² in a certain target area. If you are working in this area and the MPE for this laser is given by Table 2-3 to be 0.005 W/cm², what is the minimum OD for the laser goggles you should be wearing?

Basic Geometrical Optics

Leno S. Pedrotti

CORD

Waco, Texas

Optics is the cornerstone of photonics systems and applications. In this module, you will learn about one of the two main divisions of basic optics—*geometrical* (ray) *optics*. In the module to follow, you will learn about the other—*physical* (wave) *optics*. *Geometrical optics* will help you understand the basics of light reflection and refraction and the use of simple optical elements such as mirrors, prisms, lenses, and fibers. *Physical optics* will help you understand the phenomena of light wave interference, diffraction, and polarization; the use of thin film coatings on mirrors to enhance or suppress reflection; and the operation of such devices as gratings and quarter-wave plates.

Prerequisites

Before you work through this module, you should have completed Module 1-1, *Nature and Properties of Light*. In addition, you should be able to manipulate and use algebraic formulas, deal with units, understand the geometry of circles and triangles, and use the basic trigonometric functions (sin, cos, tan) as they apply to the relationships of sides and angles in right triangles.

Objectives

When you finish this module you will be able to:

- Distinguish between light rays and light waves.
- State the *law of reflection* and show with appropriate drawings how it applies to light rays at plane and spherical surfaces.
- State *Snell's law of refraction* and show with appropriate drawings how it applies to light rays at plane and spherical surfaces.
- Define *index of refraction* and give typical values for glass, water, and air.
- Calculate the *critical angle* of incidence for the interface between two optical media and describe the process of *total internal reflection*.
- Describe how total internal reflection can be used to redirect light in prisms and trap light in fibers.
- Describe *dispersion* of light and show how a prism disperses white light.
- Calculate the *minimum angle of deviation* for a prism and show how this angle can be used to determine the refractive index of a prism material.
- Describe what is meant by *Gaussian* or *paraxial* optics.
- Describe the relationship between *collimated light* and the *focal points* of convex and concave *mirrors*.
- Use *ray-tracing techniques* to locate the images formed by plane and spherical mirrors.
- Use the *mirror equations* to determine location, size, orientation, and nature of images formed with spherical mirrors.
- Distinguish between a *thin lens* and a *thick lens*.
- Describe the shapes of three typical *converging (positive)* thin lenses and three typical *diverging (negative)* thin lenses.
- Describe the *f-number* and *numerical aperture* for a lens and explain how they control image brightness.
- Use *ray-tracing techniques* to locate images formed by *thin lenses*.
- Describe the relationship between *collimated light* and the *focal points* of a *thin lens*.
- Use the *lensmaker's equation* to determine the focal length of a thin lens.
- Use the *thin-lens equations* to determine location, size, orientation, and nature of the images formed by simple lenses.

Scenario—Using Geometrical Optics in the Workplace

Manuel Martinez is a photonics technician hired recently to work for a large optical company that manufactures optical components such as mirrors, lenses, prisms, beam splitters, fibers, and Brewster windows—all to customer specifications. While in school Manuel studied light imaging with mirrors and lenses, ray tracing, and calculations with simple formulas. After two months on the job he has discovered that he uses those same ideas day in and day out. To be sure, things are much more “high tech” in his company, for now Manuel has access to powerful computers and computer programs that trace rays through complicated optical systems, often containing elements with nonspherical surfaces, something Manuel never had a chance to do at school. He enjoys the challenge of using state-of-the-art lab equipment he’s never seen before, including autocollimators, spectroreflectometers, and surface profilers. All in all, he’s really satisfied because all of the optics he had in his “Geo” course back at school really prepared him well for his laboratory work here. This month Manuel is learning how to “grind and polish optical surfaces to spec,” and how to apply the principles of geometrical optics to determine when the surfaces are “near tolerance.” Manuel finds his work fascinating and can hardly wait to get to work each morning. “Geo” was never so much fun.

Opening Demonstrations

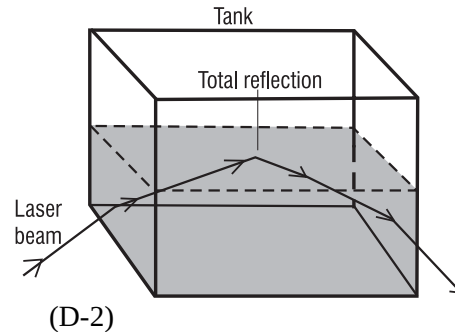
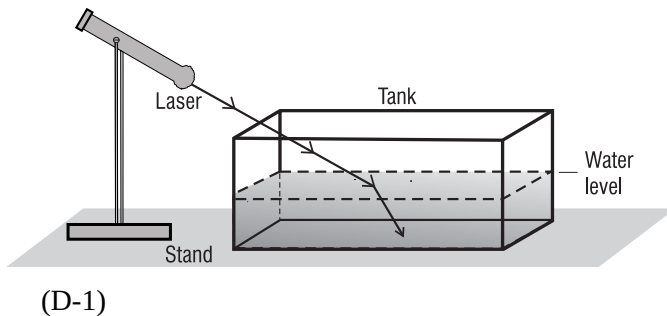
Note: The hands-on exercises that follow are to be used as short introductory laboratory demonstrations. They are intended to provide you with a glimpse of some of the phenomena covered in this module and to stimulate your interest in the study of optics and photonics.

1. Comparing Ordinary Light with Laser Light. In an appropriately darkened room, and with plenty of “chalked-up” erasers, examine the dramatic difference between ordinary “flashlight” light and laser light. Use a focusable mini MAGLITE (MAG Instrument, Ontario, Canada, 909-947-1006) and a well-collimated, ordinary low power (5.0 mW or less) diode laser pointer (Edmund Scientific Company, Barrington, New Jersey, 609-573-6250). Shine each light beam, in turn, from one side of the room to the other. Have participants “pat the erasers” together over the entire path of the light beams. The light beams outline themselves dramatically as they scatter their light energy off the settling chalk particles. Which beam remains well defined along its path? Which beam more closely describes a “ray of light”?

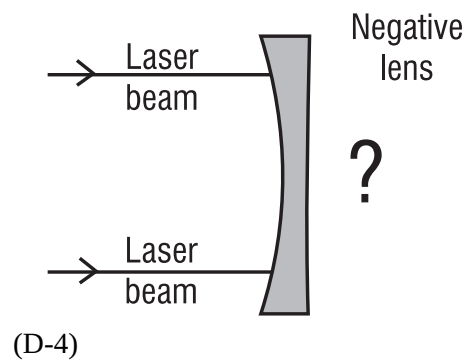
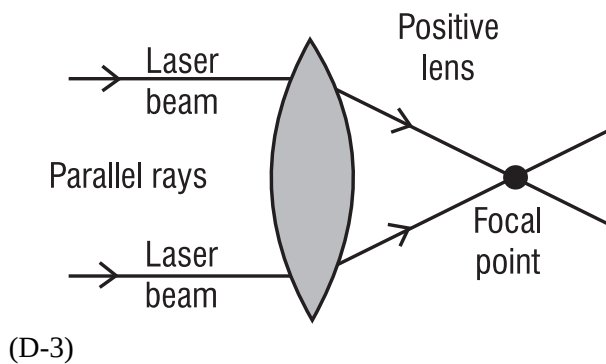
2. Bending Light Rays in a Fish Tank. Fill an ordinary rectangular five-gallon acrylic fish tank half full of water. Use the diode laser pointer to trace a “light ray” through the water in the fish tank.

- a. Attach the laser—generally cylindrical in shape—to a stand, making sure that it can be directed easily in different directions. From above the tank, direct a beam onto the top of the water at an angle of incidence near 50° . (A plane mirror placed under the tank will reflect more light back into the water.) See sketch D-1 below. Use milk or a food coloring (very sparingly—a drop at a time) to illuminate the beam. Experimenting beforehand—with a smaller container—to determine the right amount of coloring will pay big dividends. With the ray visible in the tank, observe the bending of the light beam as it moves from air into water, the phenomenon of *refraction*.

b. Next, direct the diode laser beam through one wall of the tank, up toward the water surface. See sketch D-2. Experiment with the laser beam direction until no light emerges at the water-air surface and the beam is seen to be totally reflected back into the water. The incident angle at the water-air interface is now larger than the critical angle. This phenomenon of total internal reflection is used to trap light in fibers.



3. Focusing Parallel Light Rays with a Thin Lens. Set up a positive thin lens, several inches in diameter and of focal length around 3 inches, on an optical bench. Arrange two diode laser pointers, on stands, so that they send parallel beams onto the front surface of the lens, near its outer edge. See sketch D-3. Lower the room lights and use chalk dust as in Demonstration 1 to illuminate the beams on the imaging side of the lens. The distance from the lens to the point where the beams cross is the *focal length* of the lens. Repeat with a negative lens of the same diameter and focal length, sketch D-4. What do the beams do? Where is the focal point?



Basic Concepts

I. THE LAWS OF REFLECTION AND REFRACTION

We begin our study of *basic geometrical optics* by examining how light reflects and refracts at smooth, plane interfaces. Figure 3-1a shows ordinary reflection of light at a plane surface, and Figure 3-1b shows refraction of light at two successive plane surfaces. In each instance, light is pictured simply in terms of straight lines, which we refer to as *light rays*.

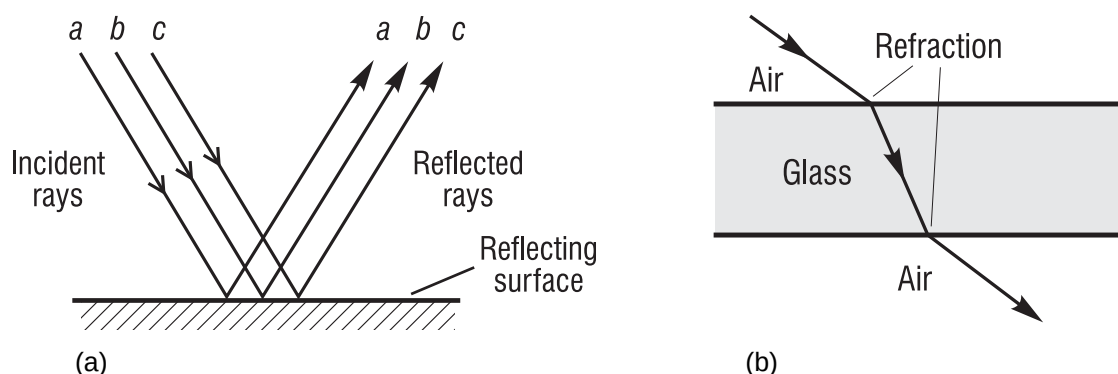


Figure 3-1 Light rays undergoing reflection and refraction at plane surfaces

After a study of how light reflects and refracts at plane surfaces, we extend our analysis to smooth, curved surfaces, thereby setting the stage for light interaction with mirrors and lenses—the basic elements in many optical systems.

In this module, the analysis of how light interacts with plane and curved surfaces is carried out with *light rays*. A light ray is nothing more than an imaginary line directed along the path that the light follows. It is helpful to think of a light ray as a narrow pencil of light, very much like a narrow, well-defined laser beam. For example, earlier in this module, when you observed the passage of a laser beam in a fish tank and visually traced the path of the beam from reflection to reflection inside the tank, you were, in effect, looking at a “light ray” representation of light in the tank.

A. Light rays and light waves

Before we look more closely at the use of *light rays* in geometrical optics, we need to say a brief word about *light waves* and the geometrical connection between light rays and light waves. For most of us, wave motion is easily visualized in terms of water waves—such as those created on a quiet pond by a bobbing cork. See Figure 3-2a. The successive high points (crests) and low points (troughs) occur as a train of circular waves moving radially outward from the bobbing cork. Each of the circular waves represents a *wave front*. A wave front is defined here as a locus of points that connect identical wave displacements—that is, identical positions above or below the normal surface of the quiet pond.

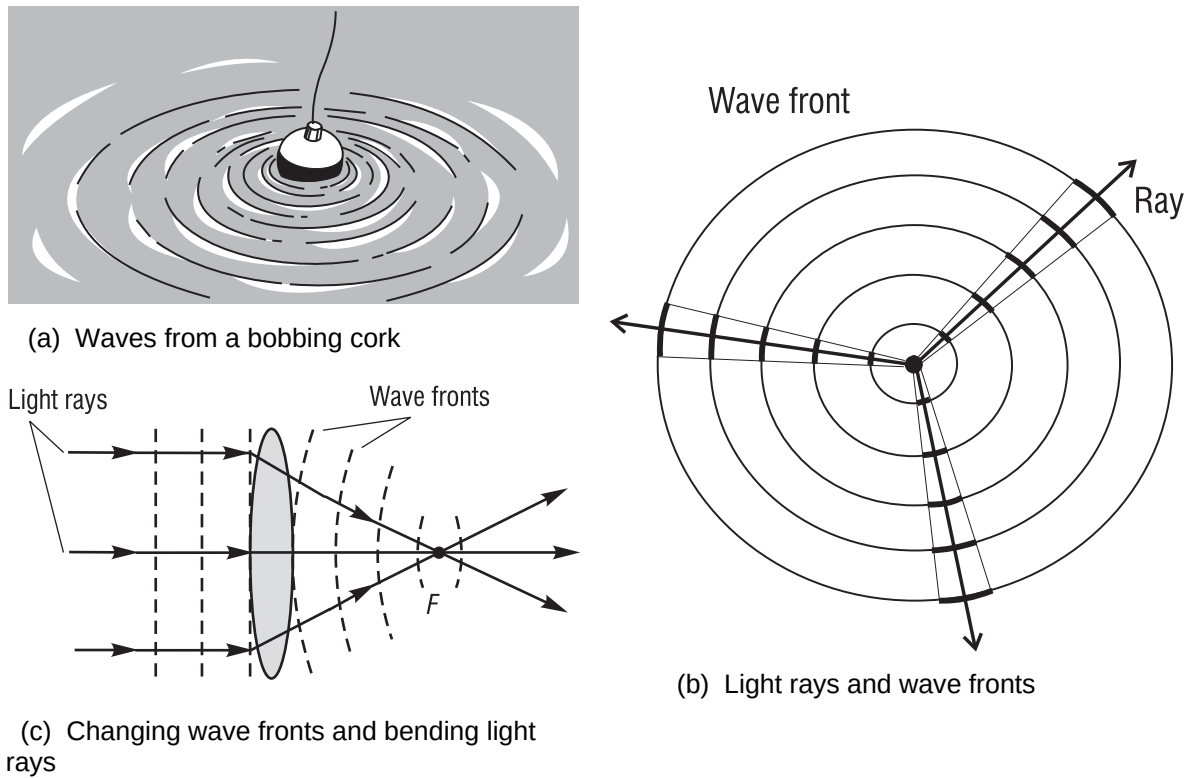


Figure 3-2 Waves and rays

In Figure 3-2b, circular wave fronts are shown with radial lines drawn perpendicular to them along several directions. Each of the rays describes the motion of a *restricted part* of the wave front along a particular direction. Geometrically then, a *ray* is a line perpendicular to a series of successive *wave fronts* specifying the *direction of energy flow* in the wave.

Figure 3-2c shows plane wave fronts of light bent by a lens into circular (spherical in three dimensions) wave fronts that then converge onto a focal point F . The same diagram shows the *light rays* corresponding to these wave fronts, bent by the lens to pass through the same focal point F . Figure 3-2c shows clearly the connection between actual waves and the rays used to represent them. In the study of geometrical optics, we find it acceptable to represent the interaction of light waves with plane and spherical surfaces—with mirrors and lenses—in terms of *light rays*.

With the useful geometric construct of a light ray we can illustrate propagation, reflection, and refraction of light in clear, uncomplicated drawings. For example, in Figure 3-3a, the propagation of light from a “point source” is represented by equally spaced light rays emanating from the source. Each ray indicates the geometrical path along which the light moves as it leaves the source. Figure 3-3b shows the reflection of several light rays at a curved mirror surface, and Figure 3-3c shows the refraction of a single light ray passing through a prism.

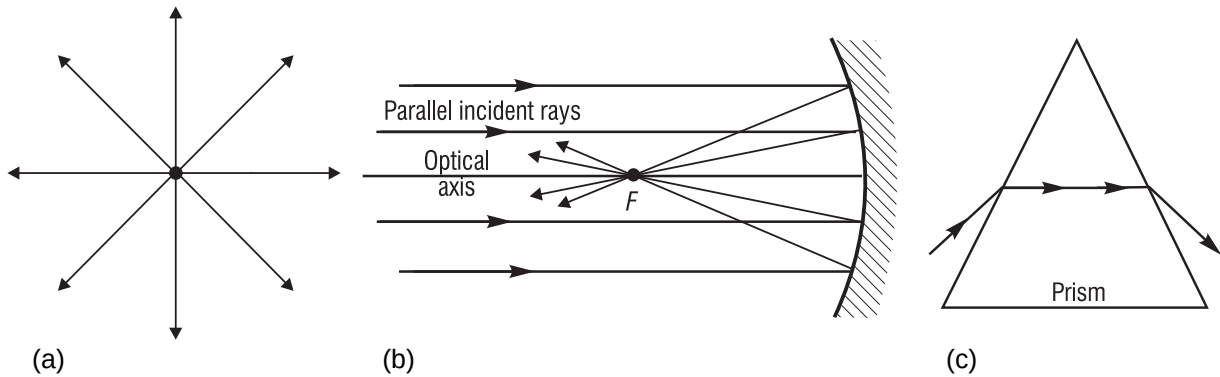


Figure 3-3 Typical light rays in (a) propagation, (b) reflection, and (c) refraction

B. Reflection of light from optical surfaces

When light is incident on an *interface* between two transparent optical media—such as between air and glass or between water and glass—four things can happen to the incident light.

- It can be partly or totally reflected at the interface.
- It can be scattered in random directions at the interface.
- It can be partly transmitted via refraction at the interface and enter the second medium.
- It can be partly absorbed in either medium.

In our introductory study of geometrical optics we shall consider only smooth surfaces that give rise to *specular* (regular, geometric) reflections (Figure 3-4a) and ignore ragged, uneven surfaces that give rise to *diffuse* (irregular) reflections (Figure 3-4b).

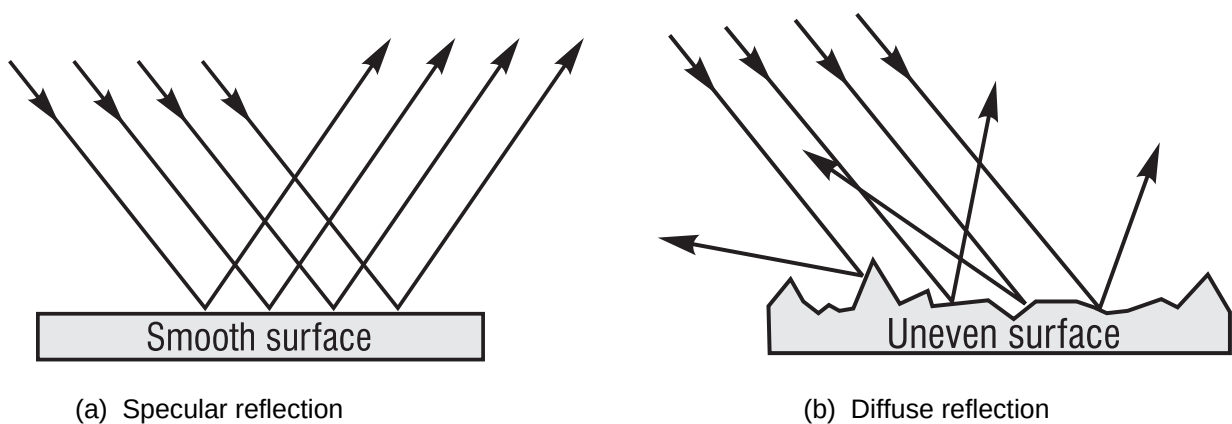


Figure 3-4 Specular and diffuse reflection

In addition, we shall ignore absorption of light energy along the path of travel, even though absorption is an important consideration when percentage of light transmitted from source to receiver is a factor of concern in optical systems.

1. The law of reflection: plane surface. When light reflects from a plane surface as shown in Figure 3-5, the angle that the reflected ray makes with the *normal* (line perpendicular to the surface) at the point of incidence is *always equal* to the angle the incident ray makes with the same normal. Note carefully that the incident ray, reflected ray, and normal always lie in the *same* plane.

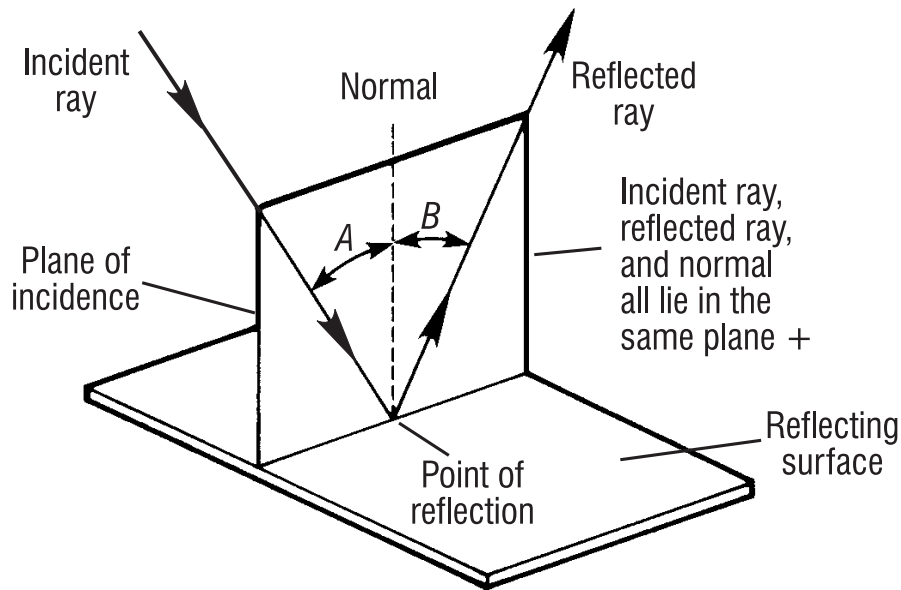


Figure 3-5 Law of reflection: Angle B equals angle A.

The geometry of Figure 3-5 reminds us that reflection of light rays from a plane, smooth surface is like the geometry of pool shots “banked” along the wall of a billiard table.

With the *law of reflection* in mind, we can see that, for the specular reflection shown earlier in Figure 3-4a, each of the incident, parallel rays reflects off the surface at the same angle, thereby remaining parallel in reflection as a group. In Figure 3-4b, where the surface is made up of many *small*, randomly oriented plane surfaces, each ray reflects in a direction different from its neighbor, even though each ray does obey the *law of reflection* at its own small surface segment.

2. Reflection from a curved surface. With spherical mirrors, reflection of light occurs at a curved surface. The *law of reflection* holds, since at each point on the curved surface one can draw a *surface tangent* and erect a *normal* to a point *P* on the surface where the light is incident, as shown in Figure 3-6. One then applies the *law of reflection* at point *P* just as was illustrated in Figure 3-5, with the incident and reflected rays making the same angles (*A* and *B*) with the normal to the surface at *P*. Note that successive surface tangents along the curved surface in Figure 3-6 are *ordered* (not random) sections of “plane mirrors” and serve—when smoothly connected—as a spherical surface mirror, capable of forming distinct images.

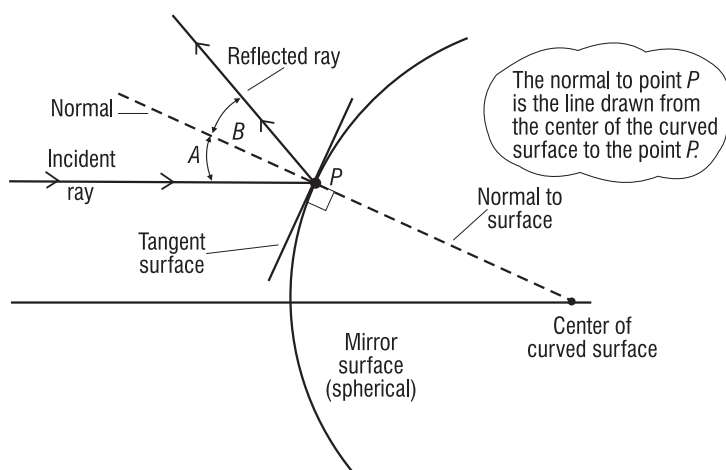
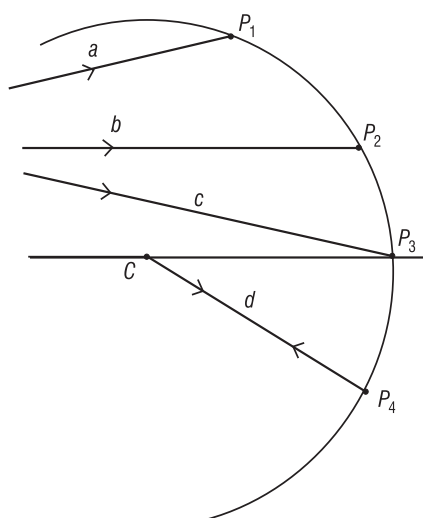


Figure 3-6 Reflection at a curved surface: Angle B equals angle A .

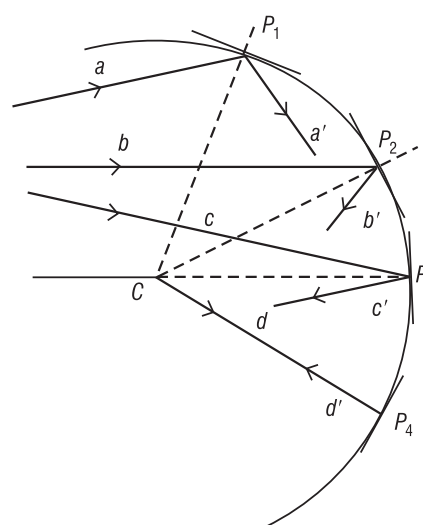
Since point P can be moved anywhere along the curved surface and a normal drawn there, we can always find the direction of the reflected ray by applying the *law of reflection*. We shall apply this technique when studying the way mirrors reflect light to form images.

Example 1

Using the *law of reflection*, complete the ray-trace diagram for the four rays (a, b, c, d) incident on the curved surface shown at the left below, given the center of the curved surface is at point C .



Beginning of ray trace



Completion of ray trace

Solution: Draw a normal (shown dashed) from point C to each of the points P_1, P_2, P_3 , and P_4 , as shown above in the drawing at the right. At each point, draw the appropriate reflected ray (a', b', c', d') so that it makes an angle with its normal equal to the angle made by the incident ray (a, b, c, d) at that point. Note that ray d reflects *back along itself* since it is incident along the line of the *normal* from C to point P_4 .

C. Refraction of light from optical interfaces

When light is incident at an *interface*—the geometrical plane that separates one optical medium from another—it will be partly reflected and partly transmitted. Figure 3-7 shows a three-dimensional view of light incident on a partially reflecting surface (interface), being reflected there (according to the *law of reflection*) and refracted into the second medium. The bending of light rays at an interface between two optical media is called *refraction*. Before we examine in detail the process of refraction, we need to describe optical media in terms of an *index of refraction*.

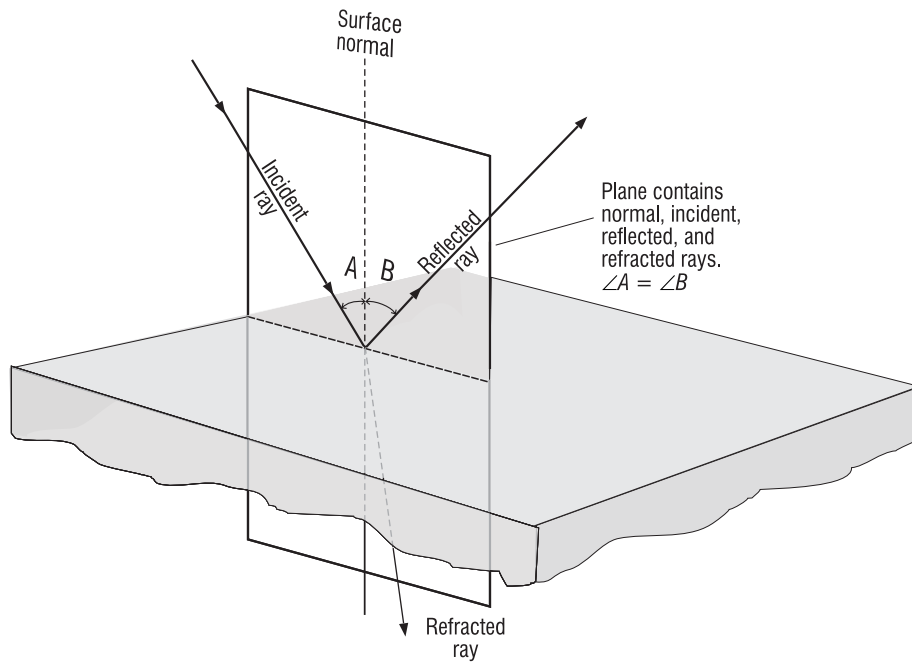


Figure 3-7 Reflection and refraction at an interface

1. Index of refraction. The two transparent optical media that form an interface are distinguished from one another by a constant called the *index of refraction*, generally labeled with the symbol n . The index of refraction for any transparent optical medium is defined as the ratio of the speed of light in a vacuum to the speed of light in the medium, as given in Equation 3-1.

$$n = \frac{c}{v} \quad (3-1)$$

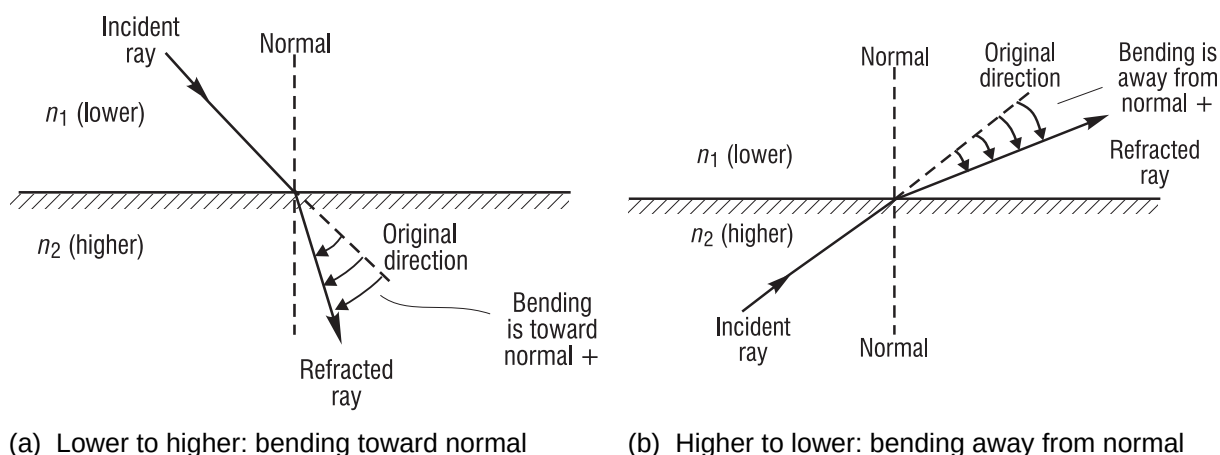
where c = speed of light in free space (vacuum)
 v = speed of light in the medium
 n = index of refraction of the medium

The index of refraction for free space is exactly *one*. For air and most gases it is very nearly one, so in most calculations it is taken to be 1.0. For other materials it has values greater than one. Table 3-1 lists indexes of refraction for common materials.

Table 3-1 Indexes of Refraction for Various Materials at 589 nm

Substance	n	Substance	n
Air	1.0003	Glass (flint)	1.66
Benzene	1.50	Glycerin	1.47
Carbon Disulfide	1.63	Polystyrene	1.49
Corn Syrup	2.21	Quartz (fused)	1.46
Diamond	2.42	Sodium Chloride	1.54
Ethyl Alcohol	1.36	Water	1.33
Gallium Arsenide (semiconductor)	3.40	Ice	1.31
Glass (crown)	1.52	Germanium	4.1
Zircon	1.92	Silicon	3.5

The greater the index of refraction of a medium, the lower the speed of light in that medium and the more light is bent in going from air into the medium. Figure 3-8 shows two general cases, one for light passing from a medium of lower index to higher index, the other from higher index to lower index. Note that in the first case (lower-to-higher) the light ray is *bent toward the normal*. In the second case (higher-to-lower) the light ray is *bent away from the normal*. It is helpful to memorize these effects since they often help one trace light through optical media in a generally correct manner.

**Figure 3-8** Refraction at an interface between media of refractive indexes n_1 and n_2

2. Snell's law. *Snell's law of refraction* relates the sines of the angles of incidence and refraction at an interface between two optical media to the indexes of refraction of the two media. The law is named after a Dutch astronomer, Willebrord Snell, who formulated the law in the 17th century. Snell's law enables us to calculate the direction of the refracted ray if we know the refractive indexes of the two media and the direction of the incident ray. The mathematical expression of Snell's law and an accompanying drawing are given in Figure 3-9.

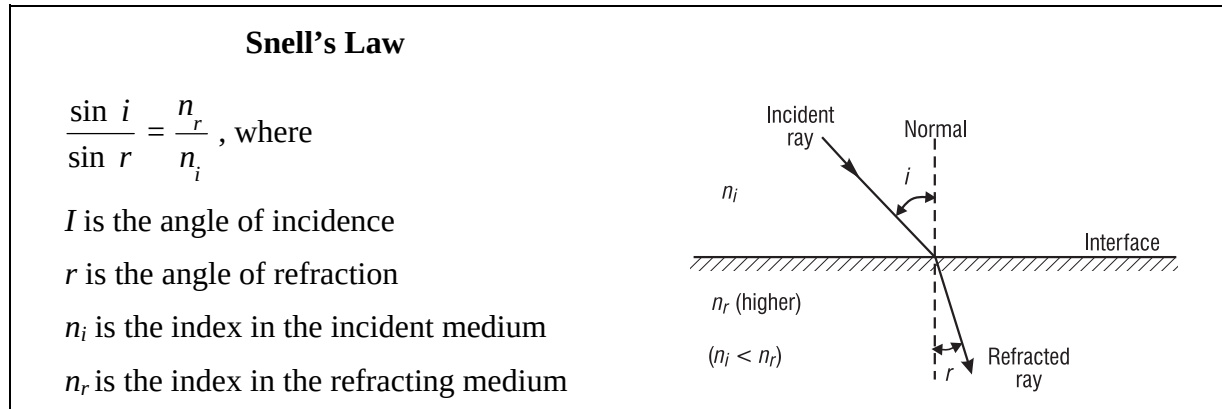


Figure 3-9 Snell's law: formula and geometry

Note carefully that both the angle of incidence (i) and refraction (r) are measured with respect to the surface normal. Note also that the incident ray, normal, and refracted ray all lie in the same geometrical plane.

In practice Snell's law is often written simply as

$$n_i \sin i = n_r \sin r \quad (3-2)$$

Now let's look at an example that make use of *Snell's law*.

Example 2

In a handheld optical instrument used under water, light is incident from water onto the plane surface of flint glass at an angle of incidence of 45° .

- What is the angle of reflection of light off the flint glass?
- Does the refracted ray bend toward or away from the normal?
- What is the angle of refraction in the flint glass?

Solution:

- From the *law of reflection*, the reflected light must head off at an angle of 45° with the normal. (Note: The angle of reflection is **not** dependent on the refractive indexes of the two media.)
- From Table 3-1, the index of refraction is 1.33 for water and 1.63 for flint glass. Thus, light is moving from a lower to a higher index of refraction and will bend **toward** the normal. We know then that the angle of refraction r should be less than 45° .
- From *Snell's law*, Equation 3-2, we have:

$$n_i \sin i = n_r \sin r$$

where $n_i = 1.33$, $i = 45^\circ$, and $n_r = 1.63$

$$\text{Thus, } \sin r = \frac{1.33 \sin 45^\circ}{1.63} = \frac{(1.33)(0.707)}{1.63} = 0.577$$

$$\text{So } r = \sin^{-1}(0.577) = 35.2^\circ$$

The angle of refraction is about 35° , clearly less than 45° , just as was predicted in part (b).

Note: The function $\sin^{-1}$ is of course the *arcsin*. We will use the $\sin^{-1}$ notation since that is what is found on scientific calculators.

3. Critical angle and total internal reflection. When light travels from a medium of higher index to one of lower index, we encounter some interesting results. Refer to Figure 3-10, where we see four rays of light originating from point O in the higher-index medium, each incident on the interface at a different angle of incidence. Ray 1 is incident on the interface at 90° (normal incidence) so there is no bending.

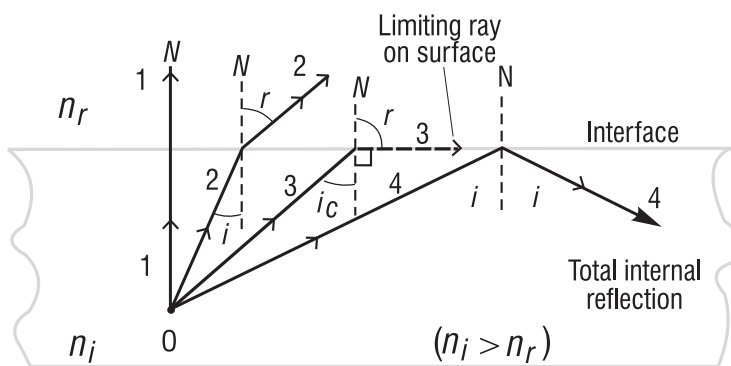


Figure 3-10 Critical angle and total internal reflection

The light in this direction simply speeds up in the second medium (why?) but continues along the same direction. Ray 2 is incident at angle i and refracts (bends away from the normal) at angle r . Ray 3 is incident at the *critical angle* i_c , large enough to cause the refracted ray bending away from the normal (N) to bend by 90° , thereby traveling along the interface between the two media. (This ray is trapped in the interface.) Ray 4 is incident on the interface at an angle *greater than* the critical angle, and is *totally reflected* into the same medium from which it came. Ray 4 obeys the *law of reflection* so that its angle of reflection is exactly equal to its angle of incidence. We exploit the phenomenon of total internal reflection when designing light propagation in fibers by trapping the light in the fiber through successive internal reflections along the fiber. We do this also when designing “retroreflecting” prisms. Compared with ordinary reflection from mirrors, the sharpness and brightness of totally internally reflected light beams is enhanced considerably.

The calculation of the critical angle of incidence for any two optical media—whenever light is incident from the medium of higher index—is accomplished with *Snell’s law*. Referring to Ray 3 in Figure 3-10 and using Snell’s law in Equation 3-2 appropriately, we have

$$n_i \sin i_c = n_r \sin 90^\circ$$

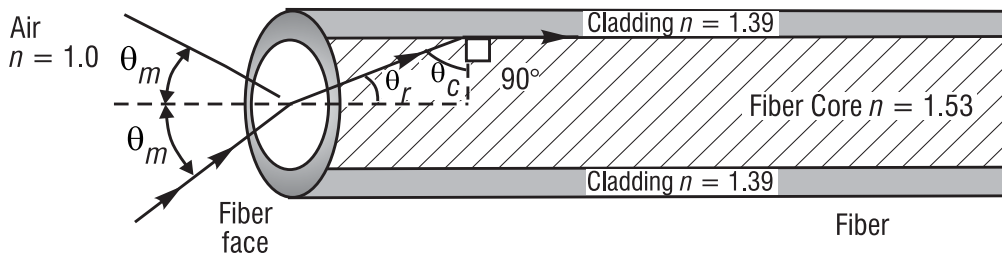
where n_i is the index for the incident medium, i_c is the critical angle of incidence, n_r is the index for the medium of lower index, and $r = 90^\circ$ is the angle of refraction at the critical angle. Then, since $\sin 90^\circ = 1$, we obtain for the critical angle,

$$i_c = \sin^{-1} \left(\frac{n_r}{n_i} \right) \quad (3-3)$$

Let's use this result and Snell's law to determine the entrance cone for light rays incident on the face of a clad fiber if the light is to be trapped by total internal reflection at the core-cladding interface in the fiber.

Example 3

A step-index fiber 0.0025 inch in diameter has a core index of 1.53 and a cladding index of 1.39. See drawing. Such clad fibers are used frequently in applications involving communication, sensing, and imaging.



What is the maximum acceptance angle θ_m for a cone of light rays incident on the fiber face such that the refracted ray in the core of the fiber is incident on the cladding at the critical angle?

Solution: First find the critical angle θ_c in the core, at the core-cladding interface. Then, from geometry, identify θ_r and use Snell's law to find θ_m .

- (1) From Equation 3-3, at the core-cladding interface

$$\theta_c = \sin^{-1} \left(\frac{1.39}{1.53} \right) = 65.3^\circ$$

- (2) From right-triangle geometry, $\theta_r = 90 - 65.3 = 24.7^\circ$

- (3) From Snell's law, at the fiber face,

$$n_{air} \sin \theta_m = n_{core} \sin \theta_r$$

$$\text{and} \quad \sin \theta_m = \left(\frac{n_{core}}{n_{air}} \right) \sin \theta_r = \left(\frac{1.53}{1.00} \right) \sin (24.7^\circ)$$

$$\text{from which} \quad \sin \theta_m = 0.639$$

$$\text{and} \quad \theta_m = \sin^{-1} 0.639 = 39.7^\circ$$

Thus, the maximum acceptance angle is 39.7° and the acceptance cone is twice that, or $2\theta_m = 79.4^\circ$. The acceptance cone indicates that any light ray incident on the fiber face within the acceptance angle will undergo total internal reflection at the core-cladding face and remain trapped in the fiber as it propagates along the fiber.

D. Refraction in prisms

Glass prisms are often used to bend light in a given direction as well as to bend it back again (retroreflection). The process of refraction in prisms is understood easily with the use of light rays and Snell's law. Look at Figure 3-11a. When a light ray enters a prism at one face and exits at another, the exiting ray is *deviated* from its original direction. The prism shown is isosceles in cross section with *apex angle* $A = 30^\circ$ and refractive index $n = 1.50$. The incident angle θ and the *angle of deviation* δ are shown on the diagram.

Figure 3-11b shows how the angle of deviation δ changes as the angle θ of the incident ray changes. The specific curve shown is for the prism described in Figure 3-11a. Note that δ goes through a minimum value, about 23° for this specific prism. Each prism material has its own unique *minimum angle of deviation*.

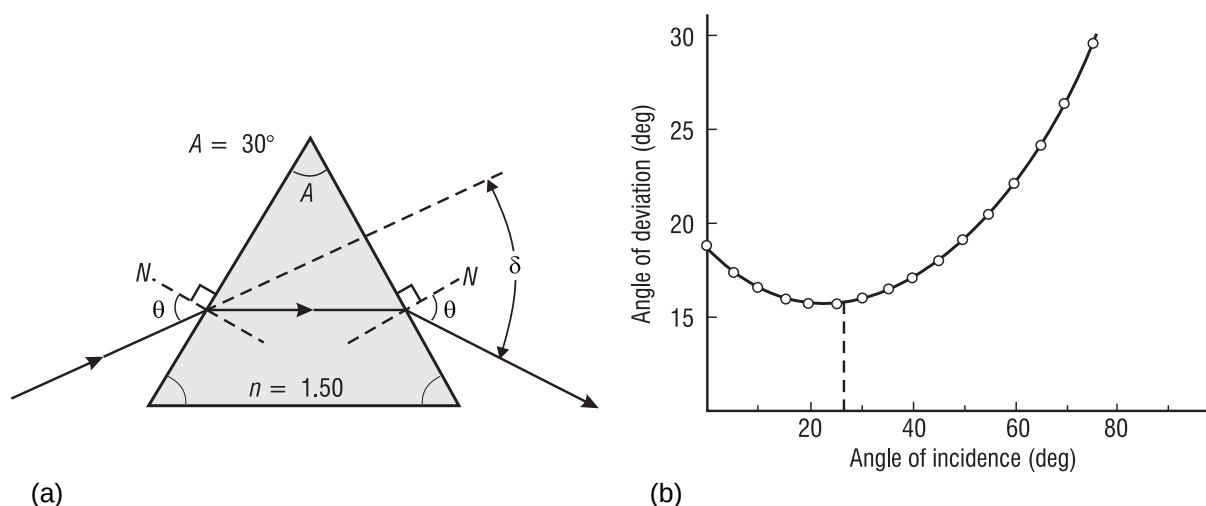


Figure 3-11 Refraction of light through a prism

1. Minimum angle of deviation. It turns out that we can determine the refractive index of a transparent material by shaping it in the form of an isosceles prism and then measuring its *minimum angle of deviation*. With reference to Figure 3-11a, the relationship between the refractive index n , the prism apex angle A , and the minimum angle of deviation δ_m is given by

$$n = \frac{\sin \left(\frac{A + \delta_m}{2} \right)}{\sin \frac{A}{2}} \quad (3-4)$$

where both A and δ_m are measured in degrees.

The derivation of Equation 3-4 is straightforward, but a bit tedious. Details of the derivation—making use of Snell's law and geometric relations between angles at each refracting surface—can be found in most standard texts on geometrical optics. (See suggested references at the end of the module.) Let's show how one can use Equation 3-4 in Example 4 to determine the index of refraction of an unknown glass shaped in the form of a prism.

Example 4

A glass of unknown index of refraction is shaped in the form of an isosceles prism with an apex angle of 25° . In the laboratory, with the help of a laser beam and a prism table, the *minimum* angle of deviation for this prism is measured carefully to be 15.8° . What is the refractive index of this glass material?

Solution: Given that $\delta_m = 15.8^\circ$ and $A = 25^\circ$, we use Equation 3-4 to calculate the refractive index.

$$n = \frac{\sin\left(\frac{A + \delta_m}{2}\right)}{\sin\left(\frac{A}{2}\right)} = \frac{\sin\left(\frac{25^\circ + 15.8^\circ}{2}\right)}{\sin\left(\frac{25^\circ}{2}\right)} = \frac{\sin(20.4^\circ)}{\sin(12.5^\circ)} = \frac{0.3486}{0.2164}$$

$$n = 1.61$$

(Comparing this value with refractive indexes given in Table 3-1, the unknown glass is probably flint glass.)

2. Dispersion of light. Table 3-1 lists indexes of refraction for various substances independent of the wavelength of the light. In fact, the refractive index is slightly wavelength dependent. For example, the index of refraction for flint glass is about 1% higher for blue light than for red light. The variation of refractive index n with wavelength λ is called *dispersion*. Figure 3-12a shows a normal dispersion curve of n_λ versus λ for different types of optical glass. Figure 3-12b shows the separation of the individual colors in white light—400 nm to 700 nm—after passing through a prism. Note that n_λ decreases from short to long wavelengths, thus causing the red light to be less deviated than the blue light as it passes through a prism. This type of dispersion that accounts for the colors seen in a rainbow, the “prism” there being the individual raindrops.

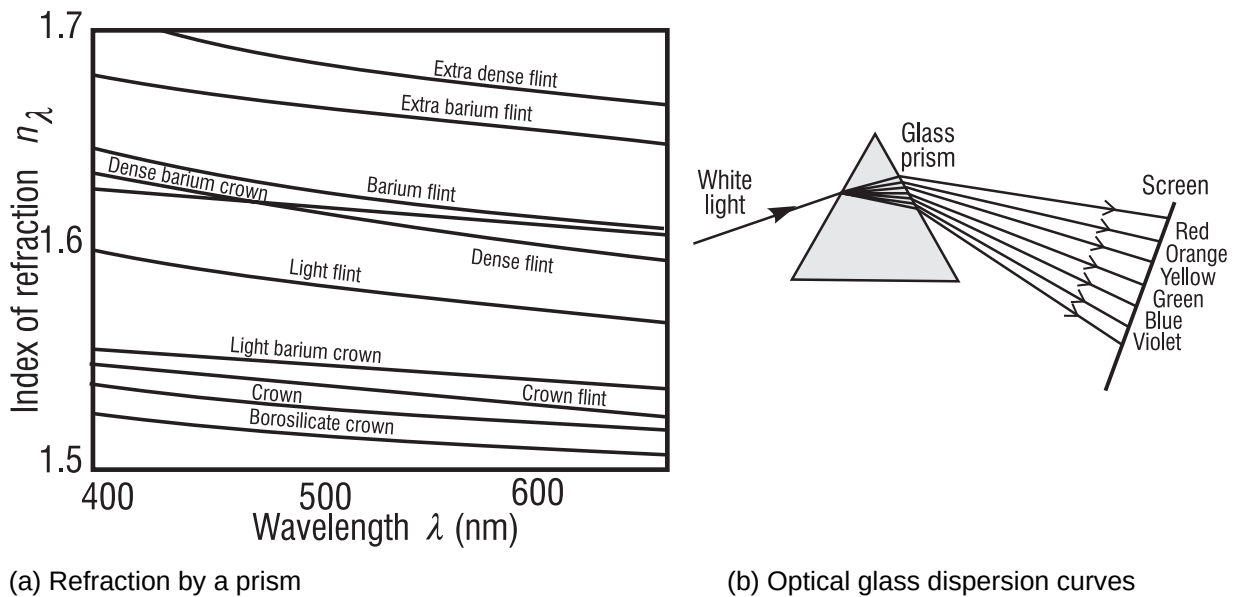


Figure 3-12 Typical dispersion curves and separation of white light after refraction by a prism

3. Special applications of prisms. Prisms that depend on *total internal reflection* are commonly used in optical systems, both to change direction of light travel and to change the orientation of an image. While mirrors can be used to achieve similar ends, the reflecting faces of a prism are easier to keep free of contamination and the process of total internal reflection is capable of higher reflectivity. Some common prisms in use today are shown in Figure 3-13, with details of light redirection and image reorientation shown for each one. If, for example, the Dove prism in Figure 3-13b is rotated about its long axis, the image will also be rotated.

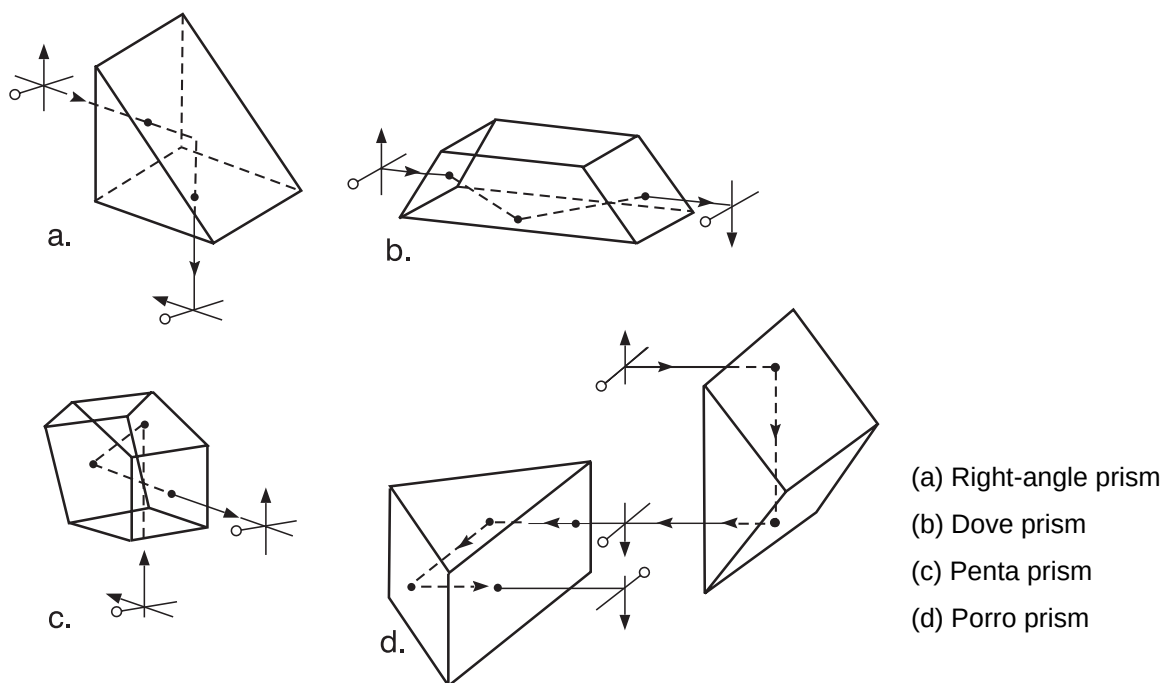


Figure 3-13 Image manipulation with refracting prisms

The Porro prism, consisting of two right-angle prisms, is used in binoculars, for example, to produce erect final images and, at the same time, permit the distance between the object-viewing lenses to be greater than the normal eye-to-eye distance, thereby enhancing the stereoscopic effect produced by ordinary binocular vision.

II. IMAGE FORMATION WITH MIRRORS

Mirrors, of course, are everywhere—in homes, auto headlamps, astronomical telescopes, and laser cavities, and many other places. Plane and spherical mirrors are used to form three-dimensional images of three-dimensional objects. If the size, orientation, and location of an object relative to a mirror are known, the *law of reflection* and ray tracing can be used to locate the image graphically. Appropriate mathematical formulas can also be used to calculate the locations and sizes of the images formed by mirrors. In this section we shall use both graphical ray tracing and formulas.

A. Images formed with plane mirrors

Images with mirrors are formed when many nonparallel rays from a given point on a source are reflected from the mirror surface, converge, and form a corresponding image point. When this happens, point by point for an extended object, an image of the object, point by point, is formed. Image formation in a plane mirror is illustrated in several sketches shown in Figure 3-14.

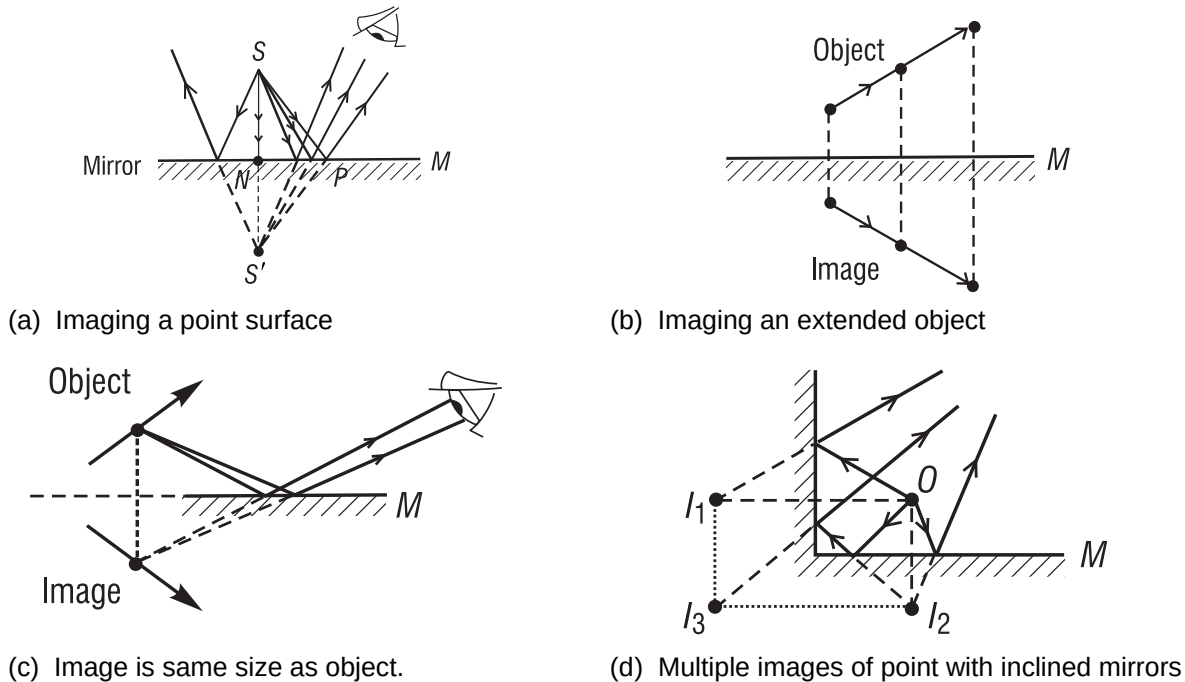


Figure 3-14 Image formation in a plane mirror

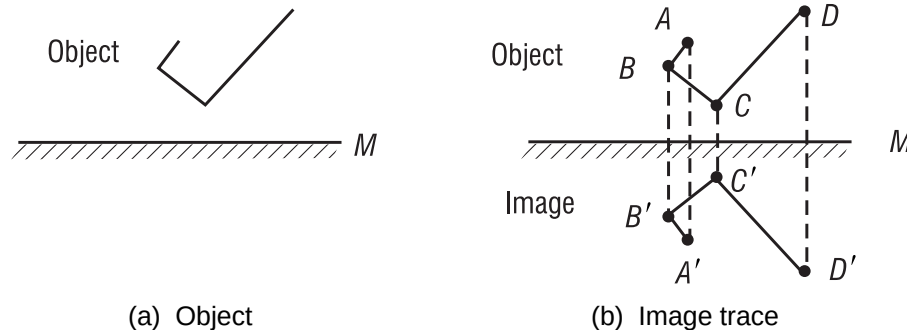
In Figure 3-14a, point object S sends nonparallel rays toward a plane mirror, which reflects them as shown. The *law of reflection* ensures that pairs of triangles like SNP and $S'NP$ are equal, so that all reflected rays appear to originate at the *image point* S' , which lies along the normal line SN , and at such depth that the *image distance* $S'N$ equals the *object distance* SN . The eye sees a point image at S' in exactly the same way it would see a real point object placed there. Since the actual rays do not exist below the mirror surface, the image is said to be a *virtual image*. The image S' cannot be projected on a screen as in the case of a *real image*. An extended object, such as the arrow in Figure 3-14b, is imaged point by point by a plane mirror surface in similar fashion. Each object point has its image point along its normal to the mirror surface and *as far below the reflecting surface as the object point lies above the surface*. Note that image position does not depend on the position of the eye.

The construction in Figure 3-14b also makes clear that the image size is identical to the object size, giving a *magnification* of unity. In addition, the transverse orientations of object and image are the same. A right-handed object, however, appears left-handed in its image. In Figure 3-14c, where the mirror does not lie directly below the object, the mirror plane may be extended to determine the position of the image as seen by an eye positioned to receive reflected rays originating at the object. Figure 3-14d illustrates multiple images of a point object O formed by two perpendicular mirrors. Each image, I_1 and I_2 , results from a single reflection in one of the

two mirrors, but a third image I_3 is also present, formed by sequential reflections from both mirrors. All parts of Figure 3-14 and the related discussion above should be understood clearly because they are fundamental to the optics of images. Look at Example 5.

Example 5

Making use of the *law of reflection* and the conclusions drawn from Figure 3-14, draw the image of the letter L positioned above a plane mirror as shown below in (a).



Solution: Make use of the fact that each point on the image is as far below the mirror—along a line perpendicular to the mirror—as the actual object point is above the mirror. Indicate key points on the object and locate corresponding points on the image. Sketch in the image as shown in (b).

B. Images formed with spherical mirrors

As we showed earlier in Figure 3-6, the *law of reflection* can be used to determine the direction along which any ray incident on a spherical mirror surface will be reflected. Using the *law of reflection*, we can trace rays from any point on an object to the mirror, and from there on to the corresponding image point. This is the method of *graphical ray tracing*.

1. Graphical ray-trace method. To employ the method of ray tracing, we agree on the following:

- Light will be incident on a mirror surface *initially* from the left.
- The axis of symmetry normal to the mirror surface is its *optical axis*.
- The point where the optical axis meets the mirror surface is the *vertex*.

To locate an image we use two points common to each mirror surface, the *center of curvature* C and the *focal point* F . They are shown in Figure 3-15, with the mirror vertex V , for both a *concave* and a *convex* spherical mirror.

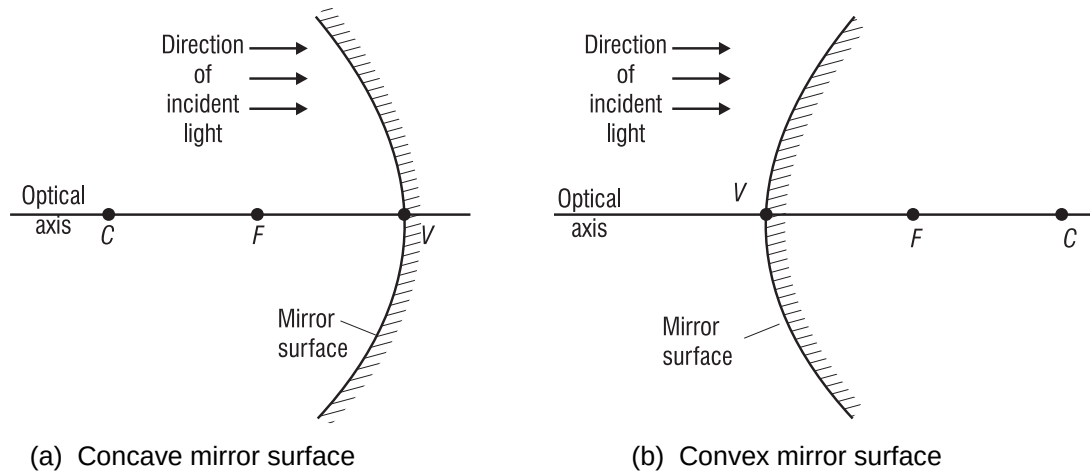


Figure 3-15 Defining points for concave and convex mirrors

The edges of *concave* mirrors always bend toward the oncoming light. Such mirrors have their center of curvature C and focal point F located to the *left* of the vertex as seen in Figure 3-15a. The edges of *convex* mirrors always bend away from the oncoming light, and their center of curvature C and focal point F are located to the *right* of the vertex. See Figure 3-15b.

The important connection between parallel rays and the focal points for mirror surfaces is shown in Figure 3-16 a, b. Parallel rays are light rays coming from a very distant source (such as the sun) or from a collimated laser beam. The *law of reflection*, applied at each point on the mirror surface where a ray is incident, requires that the ray be reflected so as to pass through a focal point F in front of the mirror (Figure 3-16a) or be reflected to appear to come from a focal point F behind the mirror (Figure 3-16b). Notice that a line drawn from the center of curvature C to any point on the mirror is a *normal* line and thus bisects the angle between the incident and reflected rays. As long as the transverse dimension of the mirror is not too large, simple geometry shows that the point F , for either mirror, is located at the midpoint between C and V , so that the distance FV is one-half the radius of curvature CV . The distance FV is called the *focal length* and is commonly labeled as f .

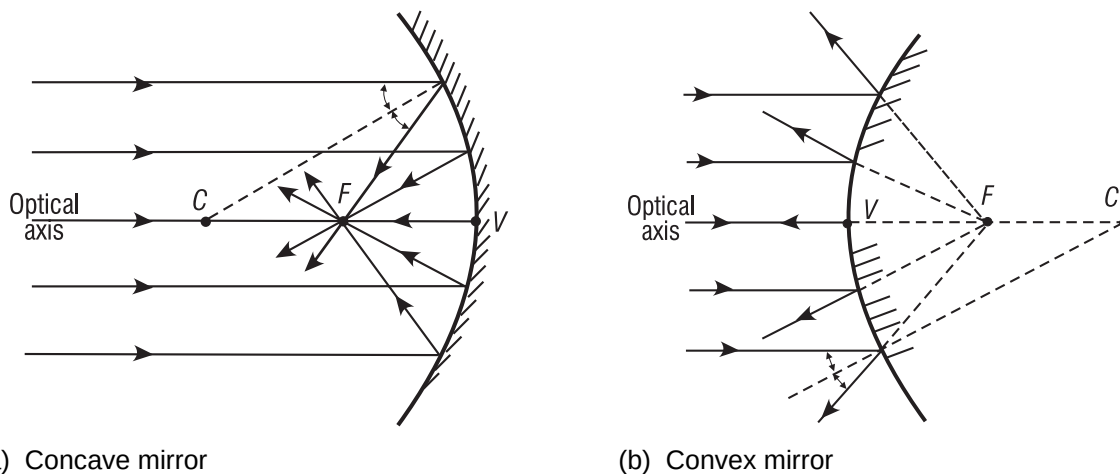


Figure 3-16 Parallel rays and focal points

2. Key rays used in ray tracing. Figure 3-17 shows three key rays—for each mirror—that are used to locate an image point corresponding to a given object point. They are

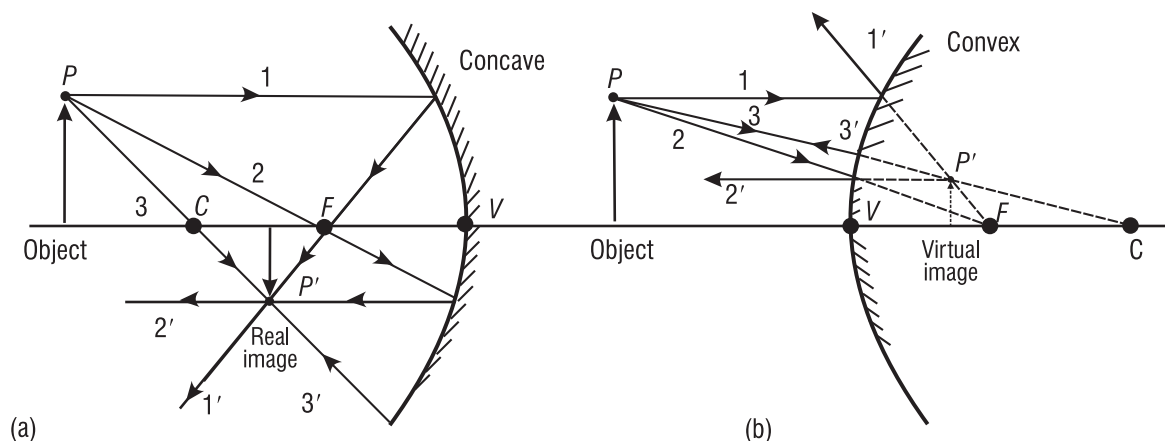


Figure 3-17 Key rays for graphical ray tracing with spherical mirrors

labeled 1, 2, and 3. Any two, drawn from object point P , will locate the corresponding image point P' . In most cases it is sufficient to locate one point, like P' , to be able to draw the entire image. Note carefully, with reference to Figure 3-17a, b, the following facts:

For a concave mirror:

- The ray from object point P parallel to the axis, such as ray 1, reflects from the mirror and *passes through the focal point F* (labeled ray 1').
- The ray from P passing through the focal point F , such as ray 2, reflects from the mirror *as a ray parallel to the axis* (labeled ray 2').
- The ray from P passing through the center of curvature C , such as ray 3, *reflects back along itself* (labeled ray 3').
- Reflected rays 1', 2', and 3' converge to locate point P' on the image. This image is a *real image* that can be formed on a screen located there.

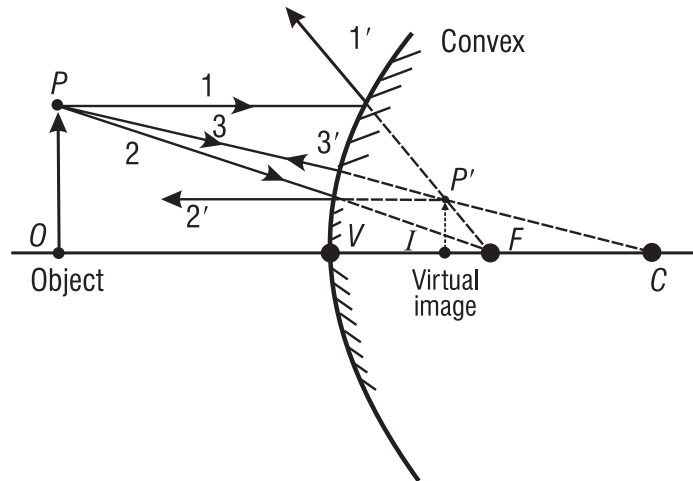
For a convex mirror:

- The ray from object point P , parallel to the axis, such as ray 1, reflects from the mirror *as if to come from the focal point F behind the mirror* (labeled ray 1').
- The ray from P , such as ray 2, headed toward the focal point F behind the mirror, reflects from the mirror in a direction parallel to the optical axis (labeled ray 2').
- The ray from P , such as ray 3, headed toward the center of curvature C behind the mirror, reflects back along itself (labeled ray 3').
- Rays 1', 2', and 3' diverge after reflection. A person looking toward the mirror intercepts the diverging rays and sees them appearing to come from their common intersection point P' , behind the mirror. The image is *virtual* since it cannot be formed on a screen placed there.

Example 6

The passenger-side mirror on an automobile is a *convex* mirror. It provides the driver with a wide field of view, but significantly reduced images. Assume that object OP is part of an automobile trailing the driver's car. See diagram below. Use three key rays to locate the reduced, virtual image of the trailing auto.

Solution: Using key rays 1, 2, and 3 incident on the mirror from point P on object OP , in conjunction with points C and F , draw the appropriate reflected rays, as show below, to locate P' on image IP' .



The three reflected rays 1', 2', and 3' *diverge* after reflection. They appear to come from a common point P' *behind* the mirror. This locates *virtual image* IP' , reduced in size, about one-third as large as object OP . As a result, drivers are always cautioned that images seen in the passenger-side mirror are actually **NEARER** than they appear to be.

C. Mirror formulas for image location

In place of the graphical ray-tracing methods described above, we can use formulas to calculate the image location. We shall derive below a “mirror formula” and then use the formula to determine image location. The derivation is typical of those found in geometrical optics, and is instructive in its combined use of algebra, geometry, and trigonometry. (If the derivation is not of interest to you, you may skip to the next section, where the derived formula is used in typical calculations. Be sure, though, that you learn about the sign convention discussed below.)

1. Derivation of the mirror formula. The drawing we need to carry out the derivation is shown in Figure 3-18. The important quantities are the object distance p , the image distance q , and the radius of curvature r . Both p and q are measured relative to the mirror vertex, as shown, and the sign on r will indicate whether the mirror is concave or convex. All other quantities in Figure 3-18 are used in the derivation but will not show up in the final “mirror formula.”

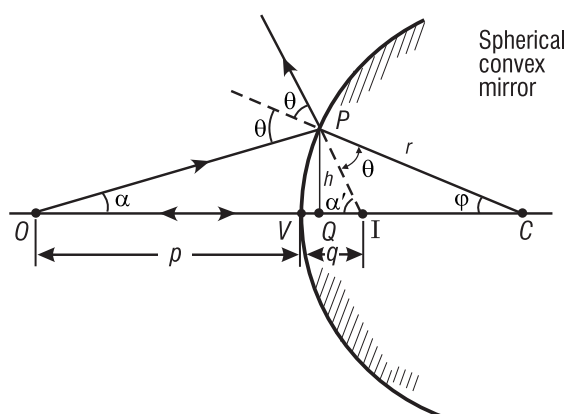


Figure 3-18 Basic drawing for deriving the mirror formula

The mirror shown in Figure 3-18 is convex with center of curvature C on the right. Two rays of light originating at object point O are drawn, one normal to the convex surface at its vertex V and the other an arbitrary ray incident at P . The first ray reflects back along itself; the second reflects at P as if incident on a plane tangent at P , according to the *law of reflection*. Relative to each other, the two reflected rays diverge as they leave the mirror. The intersection of the two rays (extended backward) determines the image point I corresponding to object point O . The image is virtual and located behind the mirror surface.

Object and image distances measured from the vertex V are shown as p and q , respectively. A perpendicular of height h is drawn from P to the axis at Q . We seek a relationship between p and q that depends on only the radius of curvature r of the mirror. As we shall see, such a relation is possible only to a first-order approximation of the sines and cosines of angles such as α and ϕ made by the object and image rays at various points on the spherical surface. This means that, in place of expansions of $\sin \phi$ and $\cos \phi$ in series as shown here,

$$\sin \phi = \phi - \frac{\phi^3}{3!} + \frac{\phi^5}{5!} - \dots$$

$$\cos \phi = 1 - \frac{\phi^2}{2!} + \frac{\phi^4}{4!} - \dots$$

we consider the first terms only and write

$$\sin \phi \cong \phi \quad \text{and} \quad \cos \phi \cong 1, \quad \text{so that} \quad \tan \phi = \frac{\sin \phi}{\cos \phi} \cong \phi$$

These relations are accurate to 1% or less if the angle ϕ is 10° or smaller. This approximation leads to *first-order (or Gaussian) optics*, after Karl Friedrich Gauss, who in 1841 developed the foundations of this subject. Returning now to the problem at hand—that of relating p , q , and r —notice that two angular relationships may be obtained from Figure 3-18, because the exterior angle of a triangle equals the sum of its interior angles. Thus,

$$\theta = \alpha + \phi \text{ in } \triangle OPC \quad \text{and} \quad 2\theta = \alpha + \alpha' \text{ in } \triangle OPI$$

which combine to give

$$\alpha - \alpha' = 2\phi$$

Using the small-angle approximation, the angles α , α' , and ϕ above can be replaced by their tangents, yielding

$$\frac{h}{p} - \frac{h}{q} = -2 \frac{h}{r}$$

Note that we have neglected the axial distance VQ , small when ϕ is small. Cancellation of h produces the desired relationship,

$$\frac{1}{p} - \frac{1}{q} = -\frac{2}{r} \quad (3-5)$$

If the spherical surface is chosen to be concave instead, the center of curvature will be to the left. For certain positions of the object point O , it is then possible to find a real image point, also to the left of the mirror. In these cases, the resulting geometric relationship analogous to Equation 3-5 consists of the same terms, but with different algebraic signs, depending on the *sign convention* employed. We can choose a sign convention that leads to a single equation, the *mirror equation*, valid for both types of mirrors. It is Equation 3-6.

$$\boxed{\frac{1}{p} + \frac{1}{q} = -\frac{2}{r}} \quad (3-6)$$

2. Sign convention. The sign convention to be used in conjunction with Equation 3-6 and Figure 3-18 is as follows.

- Object and image distances p and q are both *positive* when located to the *left* of the vertex and both *negative* when located to the *right*.
- The radius of curvature r is *positive* when the center of curvature C is to the *left* of the vertex (concave mirror surface) and *negative* when C is to the *right* (convex mirror surface).
- Vertical dimensions are positive above the optical axis and negative below.

In the application of these rules, light is assumed to be directed initially, as we mentioned earlier, from left to right. According to this sign convention, *positive* object and image distances correspond to *real* objects and images, and *negative* object and image distances correspond to *virtual* objects and images. *Virtual objects* occur only with a sequence of two or more reflecting or refracting elements.

3. Magnification of a mirror image. Figure 3-19 shows a drawing from which the *magnification*—ratio of image height h_i to object height h_o —can be determined. Since angles θ_i , θ_r , and α are equal, it follows that triangles VOP and VIP' are similar. Thus, the sides of the two triangles are proportional and one can write

$$\frac{h_i}{h_o} = \frac{q}{p}$$

This gives at once the magnification m to be

$$m \equiv \frac{h_i}{h_o} = \frac{q}{p}$$

When the *sign convention* is taken into account, one has, for the *general case*, a single equation, Equation 3-7, valid for both convex and concave mirrors.

$$m = -\frac{q}{p} \quad (3-7)$$

If, after calculation, the value of m is positive, the image is erect. If the value is negative, the image is inverted.

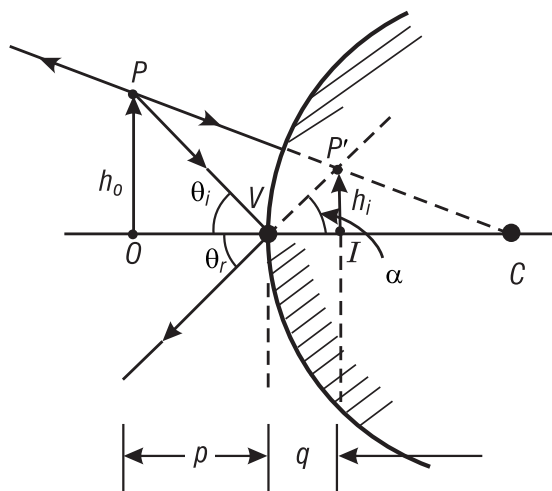
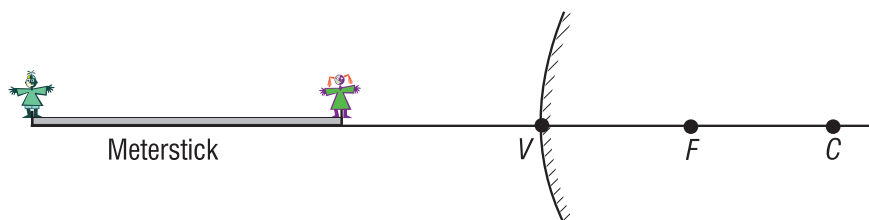


Figure 3-19 Construction for derivation of mirror magnification formula

Let us now use the mirror formulas in Equations 3-6 and 3-7, and the sign convention, to locate an image and determine its size.

Example 7

A meterstick lies along the optical axis of a convex mirror of focal length 40 cm, with its near end 60 cm from the mirror surface. Five-centimeter toy figures stand erect on both the near and far ends of the meterstick. (a) How long is the virtual image of the meterstick? (b) How tall are the toy figures in the image, and are they erect or inverted?



Solution: Use the mirror equation $\frac{1}{p} + \frac{1}{q} = -\frac{2}{r}$ twice, once for the near end and once for the far

end of the meterstick. Use the magnification equation $m = -\frac{q}{p}$ for each toy figure.

- (a) **Near end:** Sign convention gives $p = +60$ cm, $r = 2f = -(2 \times 40) = -80$ cm

$$\therefore \frac{1}{60} + \frac{1}{q_n} = -\frac{2}{80}, \text{ so } q_n = -24 \text{ cm}$$

Negative sign indicates image is *virtual*, 24 cm to the right of V.

- Far end:** $p = +160$ cm, $r = -80$ cm

$$\frac{1}{160} + \frac{1}{q_f} = -\frac{1}{40}, \text{ so } q_f = -32 \text{ cm}$$

Far-end image is *virtual*, 32 cm to the right of V.

- $\therefore$ Meterstick image is $32 \text{ cm} - 24 \text{ cm} = 8 \text{ cm}$ long.

- (b) **Near-end toy figure:**

$$m_n = \frac{-q}{p} = \frac{-(-24)}{60} = +0.4 \text{ (Image is erect since } m \text{ is positive.)}$$

The toy figure is $5 \text{ cm} \times 0.4 = 2 \text{ cm}$ tall, at near end of the meterstick image.

- Far-end toy figure:**

$$m_f = \frac{-q}{p} = \frac{-(-32)}{160} = +0.2 \text{ (Image is erect since } m \text{ is positive.)}$$

The toy figure is $5 \text{ cm} \times 0.2 = 1 \text{ cm}$ tall, at far end of the meterstick image.

III. IMAGE FORMATION WITH LENSES

Lenses are at the heart of many optical devices, not the least of which are cameras, microscopes, binoculars, and telescopes. Just as the *law of reflection* determines the imaging properties of mirrors, so *Snell's law of refraction* determines the imaging properties of lenses. Lenses are essentially light-controlling elements, used primarily for image formation with visible light, but also for ultraviolet and infrared light. In this section we shall look first at the types and properties of lenses, then use graphical ray-tracing techniques to locate images, and finally use mathematical formulas to locate the size, orientation, and position of images in simple lens systems.

A. Function of a lens

A lens is made up of a transparent refracting medium, generally of some type of glass, with spherically shaped surfaces on the front and back. A ray incident on the lens refracts at the front surface (according to Snell's law) propagates through the lens, and refracts again at the rear surface. Figure 3-20 shows a rather thick lens refracting rays from an object OP to form an image $O'P'$. The ray-tracing techniques and lens formulas we shall use here are based again on *Gaussian optics*, just as they were for mirrors.

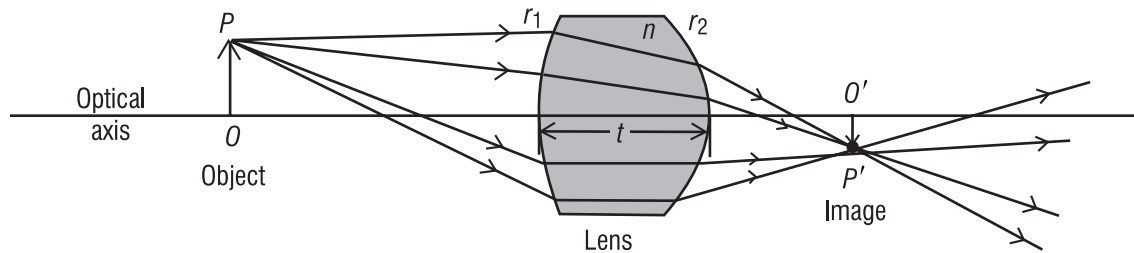


Figure 3-20 Refraction of light rays by a lens

As we have seen, *Gaussian optics*—sometimes called *paraxial optics*—arises from the basic approximations $\sin \varphi \cong \varphi$, $\tan \varphi \cong \varphi$, and $\cos \varphi \cong 1$. These approximations greatly simplify ray tracing and lens formulas, but they do restrict the angles the light rays make with the optical axis to rather small values of 20° or less.

B. Types of lenses

If the axial thickness of a lens is small compared with the radii of curvature of its surfaces, it can be treated as a *thin* lens. Ray-tracing techniques and lens formulas are relatively simple for thin lenses. If the thickness of a lens is not negligible compared with the radii of curvature of its faces, it must be treated as a *thick* lens. Ray-tracing techniques and lens-imaging formulas are more complicated for thick lenses, where computer programs are often developed to trace the rays through the lenses or make surface-by-surface calculations. In this basic introduction of geometrical optics, we shall deal with only thin lenses.

1. Converging and diverging thin lenses. In Figure 3-21, we show the shapes of several common “thin” lenses. Even though a “thickness” is shown, the use of thin lenses assumes that the rays simply refract at the front and rear faces without a translation through the lens medium. The first three lenses are *thicker in the middle than at the edges* and are described as *converging* or *positive* lenses. They are converging because they cause parallel rays passing through them to bend toward one another. Such lenses give rise to positive focal lengths. The last three lenses are *thinner in the middle than at the edges* and are described as *diverging* or *negative* lenses. In contrast with converging lenses, they cause parallel rays passing through them to spread as they leave the lens. These lenses give rise to negative focal lengths. In Figure 3-21, names associated with the different shapes are noted.

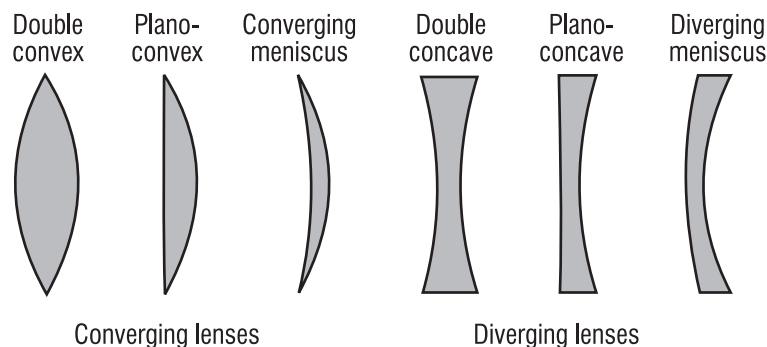


Figure 3-21 Shapes of common thin lenses

2. Focal points of thin lenses. Just as for mirrors, the *focal points* of lenses are defined in terms of their effect on parallel light rays and plane wave fronts. Figure 3-22 shows parallel light rays and their associated plane wave fronts incident on a *positive* lens (Figure 3-22a) and a *negative* lens (Figure 3-22b). For the positive lens, refraction of the light brings it to focal point F (real image) to the right of the lens. For the negative lens, refraction of the light causes it to diverge as if it is coming from focal point F (virtual image) located to the left of the lens. Note how the plane wave fronts are changed to converging spherical wave fronts by the positive lens and to diverging spherical wave fronts by the negative lens. This occurs because light travels more slowly in the lens medium than in the surrounding air, so the thicker parts of the lens retard the light more than do the thinner parts.

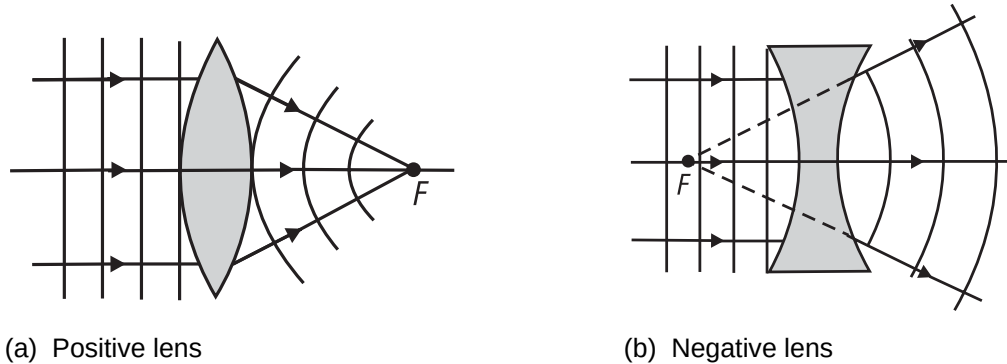


Figure 3-22 Focal points for positive and negative lenses

Recall that, for mirrors, there is but a *single* focal point for each mirror surface since light remains always on the same side of the mirror. For thin lenses, there are *two* focal points, symmetrically located on each side of the lens, since light can approach from either side of the lens. The sketches in Figure 3-23 indicate the role that the two focal points play, for positive lenses (Figure 3-23a) and negative lenses (Figure 3-23b). Study these figures carefully.

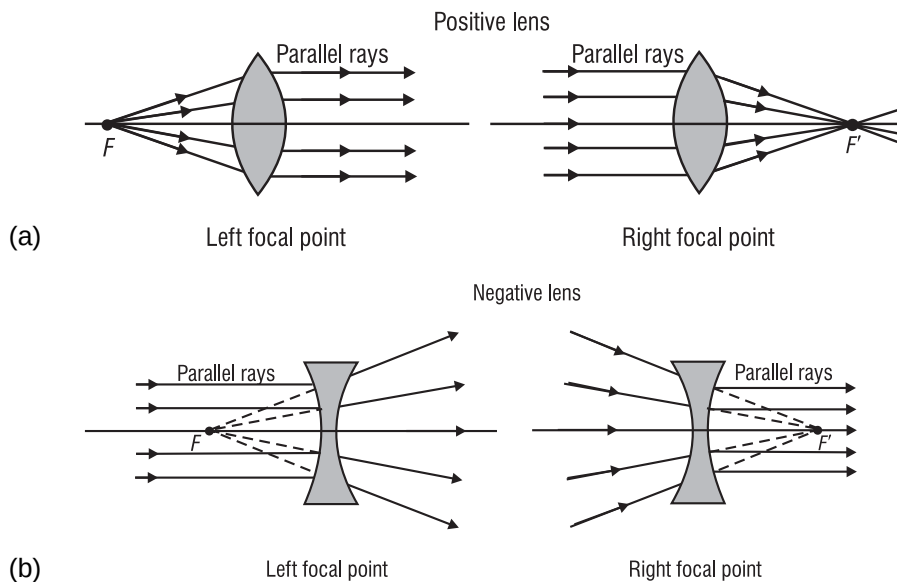


Figure 3-23 Relationship of light rays to right and left focal points in thin lenses

3. *f*-number and numerical aperture of a lens. The size of a lens determines its light-gathering power and, consequently, the brightness of the image it forms. Two commonly used indicators of this special characteristic of a lens are called the *f*-number and the *numerical aperture*.

The *f*-number, also referred to as the *relative aperture* and the *f/stop*, is defined simply as the ratio of the focal length *f* of the lens, to its diameter *D*, as given in Equation 3-8.

$$\boxed{f\text{-number} = \frac{f}{D}} \quad (3-8)$$

For example, a lens of focal length 4 cm *stopped down* to an aperture of 0.5 cm has an *f*-number of $4/0.5 = 8$. Photographers usually refer to this situation as a lens with an *f/stop* of *f*/8. Before the advent of fully automated cameras (“point and shoot”), a photographer would routinely select an aperture size for a given camera lens—(thereby setting the *f/stop*), a shutter speed, and a proper focus to achieve both the desired image brightness and sharpness.

Table 3-2 lists the usual choices of *f/stops* (*f*-numbers) available on cameras and the corresponding image irradiance or “brightness”—in watts per square meter. The listing gives the irradiance E_0 as the value for an *f/stop* of 1 and shows how the image irradiance decreases as the lens is “stopped down,” that is, as the adjustable aperture size behind the camera lens is made smaller. From Equation 3-8, it should be clear that, for a given camera lens of focal length *f*, the *f/stop* or *f*-number increases as *D* decreases, that is, as the aperture size decreases. Clearly then, increasing the *f*-number of a lens decreases its light-gathering power.

**Table 3-2. Relative Image Irradiance (Brightness)
as a Function of *f/stop* Setting**

<i>f/stop</i> or <i>f</i> -number	Relative Image Irradiance in watts/m ²
1	E_0
1.4	$E_0/2$
2	$E_0/4$
2.8	$E_0/8$
4	$E_0/16$
5.6	$E_0/32$
8	$E_0/64$
11	$E_0/128$
16	$E_0/256$
22	$E_0/512$

Since the *total exposure* in joules/m² on the film is the product of the irradiance in joules/(m²-s) and the exposure time (shutter speed) in seconds, a desirable film exposure can be obtained in a variety of ways. Accordingly, if a particular film—whose speed is described by an ASA number—is perfectly exposed by light from a particular scene with a shutter speed of 1/50 second and an *f/stop* of *f*/8 (irradiance equals $E_0/64$ from Table 3-2), it will also be perfectly exposed by any other combination that gives the same total exposure. For example, by choosing

a shutter speed of 1/100 second and an f /stop of $f/5.6$, the exposure time is cut in half while the irradiance ($E_0/32$) is doubled, thereby leaving *no net change* in the film exposure (J/m^2).

The *numerical aperture* is another important design parameter for a lens, related directly to how much light the lens gathers. If the focal length of a design lens increases and its diameter decreases, the solid angle (cone) of useful light rays from object to image for such a lens decreases. For example, the concept of a numerical aperture finds immediate application in the design of the objective lens (the lens next to the specimen under observation) for a *microscope*, as we show below. Light-gathering capability is crucial for microscopes.

Figure 3-24 depicts the light-gathering power of a lens relative to a point O on a specimen covered by a glass slide. Lens L is the objective lens of a microscope focused on the specimen. On the right side of the symmetry axis of the lens, the light-gathering power of the lens—with *air between the cover slide and the lens*—is depicted in terms of half-angle α_{air} . On the left side, by contrast, the increased light-gathering power of the lens—with *oil situated between the cover slide and the lens*—is shown in terms of the *larger* half-angle α_{oil} . The oil is chosen so as to

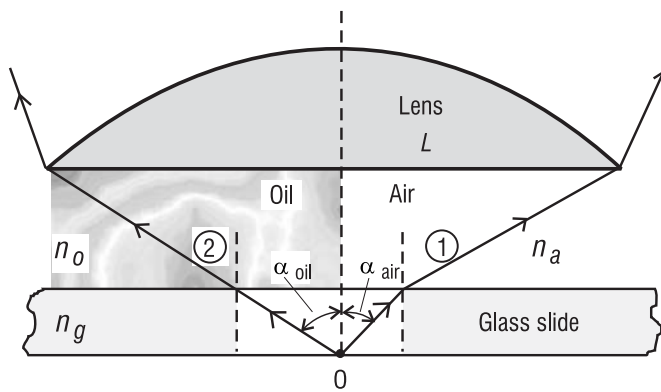


Figure 3-24 Light-gathering power of oil-immersion and air-immersion lens, showing that α_{oil} is greater than α_{air}

have an index of refraction (n_o) very near that of the cover slide (n_g) so that little or no refraction occurs for limiting ray 2 at the glass-oil interface. Consequently the half-angle α_{oil} is greater than the half-angle α_{air} . As Figure 3-24 shows, ray 1 suffers refraction at the glass-air interface, thereby restricting the cone of rays accepted by the lens to the smaller half-angle α_{air} .

The *numerical aperture* of a lens is defined so as to exhibit the difference in solid angles (cones) of light accepted, for example, by an “oil-immersion” arrangement versus an air-immersion setup.

The definition of numerical aperture (N.A.) is given in Equation 3-9 as

$$N.A. = n \sin \alpha \quad (3-9)$$

where n is the index of refraction of the intervening medium between object and lens and α is the half-angle defined by the limiting ray (α_{air} or α_{oil} in Figure 3-24). The “light-gathering” power of the microscope’s objective lens is thus increased by increasing the refractive index of the intervening medium.

In addition, the numerical aperture is closely related to the acceptance angle discussed in Example 3 for both graded-index and step-index optical fibers, as will be shown in Module 1-7, *Optical Waveguides and Fibers*. Since the rays entering the fiber face are in air, the numerical aperture $N.A.$ is equal simply to $N.A. = \sin \alpha$.

It is shown in most basic books on optics (see references listed at end of this module) that image brightness is dependent on values of the f -number or numerical aperture, in accordance with the following proportionalities:

$$\text{image brightness} \propto \frac{1}{(f\text{-number})^2}$$

$$\text{image brightness} \propto (N.A.)^2$$

In summary, one can *increase* the light-gathering power of a lens and the brightness of the image formed by a lens by *decreasing* the f -number of the lens (increasing lens diameter) or by *increasing* the numerical aperture of the lens (increasing the refraction index and thus making possible a larger acceptance angle).

C. Image location by ray tracing

To locate the image of an object formed by a thin lens, we make use of three key points for the lens and associate each of them with a defining ray. The three points are the left focal point F , the right focal point F' , and the lens vertex (center) V . In Figure 3-25 the three rays are shown locating an image point P' corresponding to a given object point P , for both a positive and a negative lens. The object is labeled OP and the corresponding image IP' . The defining rays are labeled to show clearly their connection to the points F , F' , and V . In practice, of course, only two of the three rays are needed to locate the desired image point. Note also that the location of image point P' is generally sufficient to sketch in the rest of the image IP' , to correspond with the given object OP .

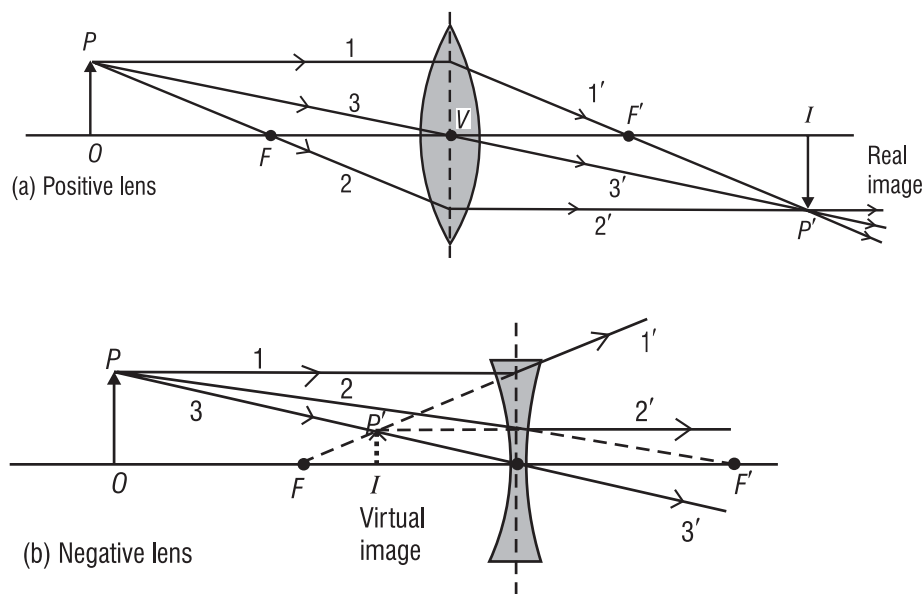


Figure 3-25 Ray diagrams for image formation by positive and negative lenses

The behavior of rays 1 and 2—connected with the left and right focal points for both the positive and negative lenses—should be apparent from another look at Figure 3-23. The behavior of ray 3—going straight through the lens at its center V —is a consequence of assuming that the lens has zero thickness. Note, in fact, that, for both Figures 3-23 and 3-25, all the bending is assumed to take place at the dashed vertical line that splits the drawn lenses in half. Also, it should be clear in Figure 3-25 that the positive lens forms a real image while the negative lens forms a virtual image.

One can apply the principles of ray tracing illustrated in Figure 3-25 to a train of thin lenses. Figure 3-26 shows a ray trace through an “optical system” made up of a positive and a negative lens. For accuracy in drawing, a common practice used here is to show the positive lens as a vertical line with normal arrowheads and the negative lens as a vertical line with inverted arrowheads, and to show all ray bending at these lines. Note that the primary object is labeled RO_1 (real object 1) and its image formed by the positive lens is labeled RI_1 (real image 1). The image RI_1 then serves as a real object (RO_2) for the negative lens, leading finally to a virtual image VI_2 .

Test your understanding of ray tracing through thin lenses by accounting for each numbered ray drawn in the figure. What happens to rays 1 and 3 relative to the negative lens? Why are rays 4 and 5 introduced? Is this a “fair” practice?

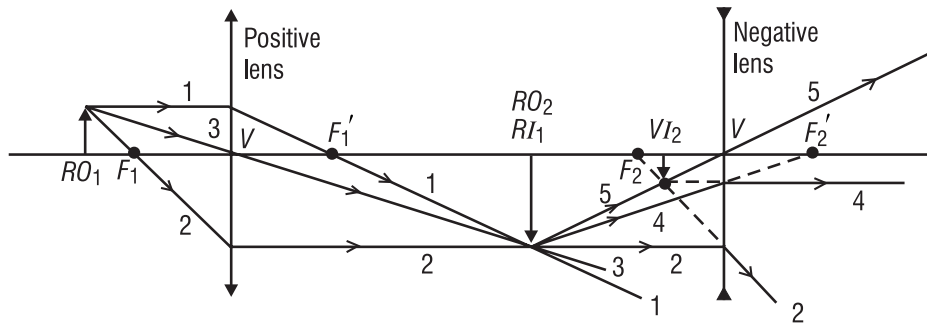


Figure 3-26 Ray diagram for image formation through two lenses

D. Lens formulas for thin lenses

As with mirrors, convenient formulas can be used to locate the image mathematically. The derivation of such formulas—as was carried out for spherical mirrors in the previous section—can be found in most texts on geometrical optics. The derivation essentially traces an arbitrary ray *geometrically* and *mathematically* from an object point through the two surfaces of a thin lens to the corresponding image point. *Snell's law* is applied for the ray at each spherical refracting surface. The details of the derivation involve the geometry of triangles and the approximations mentioned earlier— $\sin \varphi \cong \varphi$, $\tan \varphi \cong \varphi$, and $\cos \varphi \cong 1$ —to simplify the final results. Figure 3-27 shows the essential elements that show up in the final equations, relating object distance p to image distance q , for a lens of focal length f with radii of curvature r_1 and r_2 and refractive index n_g . For generality, the lens is shown situated in an arbitrary medium of refractive index n . If the medium is air, then, of course, $n = 1$.

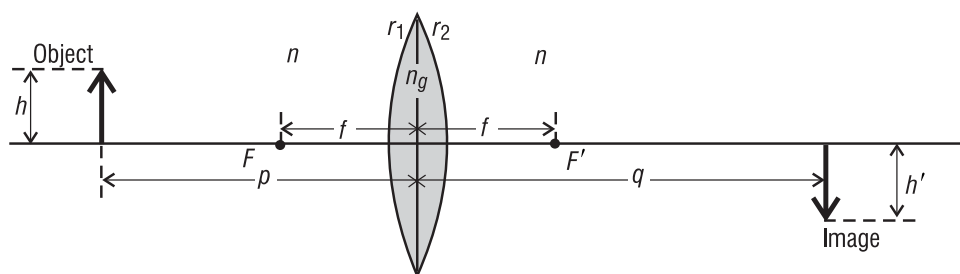


Figure 3-27 Defining quantities for image formation with a thin lens

1. Equations for thin lens calculations. The *thin lens equation* is given by Equation 3-10.

$$\boxed{\frac{1}{p} + \frac{1}{q} = \frac{1}{f}} \quad (3-10)$$

where p is the object distance (from object to lens vertex V)

q is the image distance (from image to lens vertex V)

and f is the focal length (from either focal point F or F' to the lens vertex V)

For a lens of refractive index n_g situated in a medium of refractive index n , the relationship between the parameters n , n_g , r_1 , r_2 and the focal length f is given by the *lensmaker's equation* in Equation 3-11.

$$\boxed{\frac{1}{f} = \left(\frac{n_g - n}{n} \right) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)} \quad (3-11)$$

where n is the index of refraction of the surrounding medium

n_g is the index of refraction of the lens materials

r_1 is the radius of curvature of the front face of the lens

r_2 is the radius of curvature of the rear face of the lens

The magnification m produced by a thin lens is given in Equation 3-12.

$$\boxed{m = \frac{h_i}{h_o} = -\frac{q}{p}} \quad (3-12)$$

where m is the magnification (ratio of image size to object size)

h_i is the transverse size of the image

h_o is the transverse size of the object

p and q are object and image distance respectively

2. Sign convention for thin lens formulas. Just as for mirrors, we must agree on a sign convention to be used in the application of Equations 3-10, 3-11, and 3-12. It is:

- Light travels initially from left to right toward the lens.

- Object distance p is *positive* for *real* objects located to the *left* of the lens and *negative* for *virtual* objects located to the *right* of the lens.
- Image distance q is *positive* for *real* images formed to the *right* of the lens and *negative* for *virtual* images formed to the *left* of the lens.
- The focal length f is *positive* for a *converging* lens, *negative* for a *diverging* lens.
- The radius of curvature r is *positive* for a *convex* surface, *negative* for a *concave* surface.
- Transverse distances (h_o and h_i) are *positive* above the optical axis, *negative* below.

Now let's apply Equations 3-10, 3-11, and 3-12 in several examples, where the use of the sign convention is illustrated and where the size, orientation, and location of a final image are determined.

Example 8

A double-convex thin lens such as that shown in Figure 3-21 can be used as a simple “magnifier.” It has a front surface with a radius of curvature of 20 cm and a rear surface with a radius of curvature of 15 cm. The lens material has a refractive index of 1.52. Answer the following questions to learn more about this simple magnifying lens.

- What is its focal length in air?
- What is its focal length in water ($n = 1.33$)?
- Does it matter which lens face is turned toward the light?
- How far would you hold an index card from this lens to form a sharp image of the sun on the card?

Solution:

- Use the lensmaker's equation. With the sign convention given, we have $n_g = 1.52$, $n = 1.00$, $r_1 = +20$ cm, and $r_2 = -15$ cm. Then

$$\frac{1}{f} = \left(\frac{n_g - n}{n} \right) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) = \left(\frac{1.52 - 1}{1} \right) \left(\frac{1}{20} - \frac{1}{-15} \right) = 0.0607$$

So $f = +16.5$ cm (a converging lens, so the sign is positive, as it should be)

- $\frac{1}{f} = \left(\frac{1.52 - 1.33}{1.33} \right) \left(\frac{1}{20} - \frac{1}{-15} \right) = 0.0167$

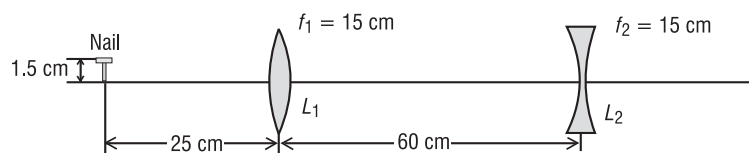
$f = 60$ cm (converging but less so than in air)

- No, the magnifying lens behaves the same, having the same focal length, no matter which surface faces the light. You can prove this by reversing the lens and repeating the calculation with Equation 3-11. Results are the same. But **note carefully**, reversing a *thick* lens changes its effect on the light passing through it. The two orientations are *not* equivalent.

(d) Since the sun is very far away, its light is collimated (parallel rays) as it strikes the lens and will come to a focus at the lens focal point. Thus, one should hold the lens about 16.5 cm from the index card to form a sharp image on the card.

Example 9

A two-lens system is made up of a converging lens followed by a diverging lens, each of focal length 15 cm. The system is used to form an image of a short nail, 1.5 cm high, standing erect, 25 cm from the first



lens. The two lenses are separated by a distance of 60 cm. See accompanying diagram. (Refer to Figure 3-26 for a ray-trace diagram of what's going on in this problem.)

Locate the final image, determine its size, and state whether it is real or virtual, erect or inverted.

Solution: We apply the thin lens equations to each lens in turn, making use of the correct sign convention at each step.

$$\text{Lens } L_1: \quad \frac{1}{p_1} + \frac{1}{q_1} = \frac{1}{f_1} \quad \text{or} \quad \frac{1}{25} + \frac{1}{q_1} = \frac{1}{15} \quad (f_1 \text{ is } + \text{ since lens } L_1 \text{ is converging.})$$

$q_1 = +37.5$ cm (Since the sign is positive, the image is real and located 37.5 cm to the right of lens L_1 .)

$$\text{Lens } L_2: \quad \frac{1}{p_2} + \frac{1}{q_2} = \frac{1}{f_2} \quad \text{where } p_2 = (60 - 37.5) = 22.5 \text{ cm}$$

Since the first image, a distance q_1 from L_1 , serves as the object for the lens L_2 , this object is to the left of lens L_2 , and thus its distance p_2 is positive. The focal length for L_2 is negative since it is a diverging lens. So, the thin lens equation becomes

$$\frac{1}{22.5} + \frac{1}{q_2} = \frac{1}{-15}, \quad \text{giving } q_2 = -9 \text{ cm}$$

Since q_2 is negative, it locates a *virtual* image, 9 cm to the left of lens L_2 . (See Figure 3-26.)

The overall magnification for the two-lens system is given by the combined magnification of the lenses. Then

$$m_{\text{sys}} = m_1 \times m_2 = \left(-\frac{q_1}{p_1} \right) \left(-\frac{q_2}{p_2} \right) = \left(-\frac{37.5}{25} \right) \left(-\frac{-9}{22.5} \right) = -0.6$$

Thus, the final image is inverted (since overall magnification is negative) and is of final size $(0.6 \times 1.5 \text{ cm}) = 0.9 \text{ cm}$.

Laboratory

In this laboratory you will perform the following simple experiments with prisms and lenses:

- Determine the index of refraction of a prism material.
- Demonstrate total internal reflection (TIR) with right-angle prisms and show how to use the prisms to produce (a) 90° bending, (b) retroreflection, and (c) “periscope-type” bending.
- Determine the index of refraction of a thin-lens material.
- Determine the focal lengths of convex and concave lenses.

Equipment List

The following equipment is needed to complete this laboratory:

- 1 equilateral prism^a (25-mm faces by 25 mm long works well)
- 2 45°-45°-90° prisms^a (25-mm legs, 35-mm hypotenuse, 25 mm thick)
- 2 diode laser pointers^b (5 mW or less)
- 1 spherometer^b
- 1 double-convex lens^a (75-mm diameter by 150-mm focal length)
- 1 double-concave lens^a (75-mm diameter by 150-mm focal length)
- 1 protractor
- 1 white cardboard screen
- Index cards, white paper sheet, (8½" × 11" and 11" × 17"), masking tape, and ruler

Procedure

A. Index of Refraction of a Prism Material

1. Arrange the laser pointer, equilateral prism, and white cardboard screen on a flat tabletop as shown in Figure L-1. Center the prism over a sheet of white paper. Fasten down the white paper, cardboard screen, and laser with tape.

^a These items are readily available from Edmund Scientific, Barrington, New Jersey, 609-573-6250, as parts of their “Educational Quality Demonstration Optics,” at reasonable prices. See their Industrial Optics Division catalog.

^b These items are also available from Edmund Scientific but are more expensive.

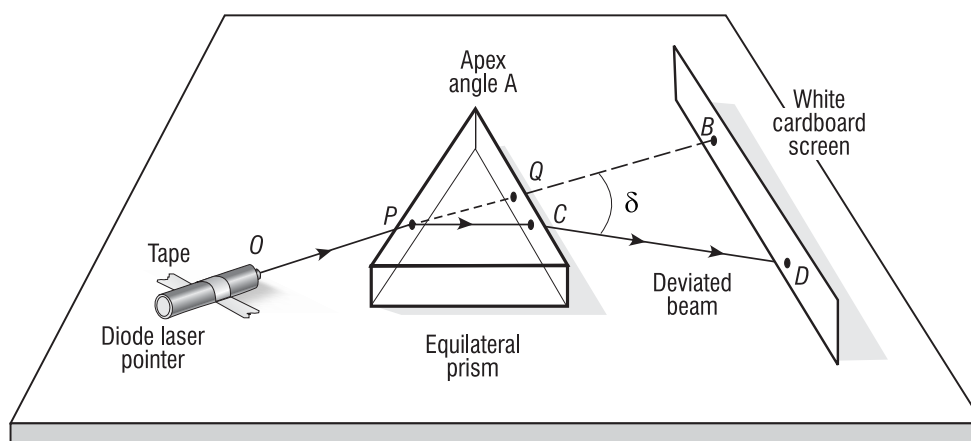


Figure L-1 Setup for measuring minimum angle of deviation

2. As you rotate the prism relative to the incident laser beam, the laser spot D on the screen moves, so the angle of deviation δ will become larger or smaller. By experimentation, determine the *smallest angle of deviation* (δ_m) between an original beam direction $OPQB$ and the deviated beam CD . (It should be clear that the farther the screen is from the prism the more precise will be your determination of δ_m , since small changes in spot D will then be more exaggerated.)
3. When you have achieved the minimum angle for δ , carefully tape the prism in place. Trace the prism edges on the paper, the straight segments OP and QB along the *original direction*, and the segment CD . (**Note:** Location of laser spots Q , C on the exit face of the prism and B , D on the screen are needed to be able to draw segments QB and CD .) With the line segments drawn, remove the prism and measure the minimum angle δ_m with a protractor. Complete a ray trace of the incident beam through the prism, deviated at angle δ_m . Is the segment DC parallel to the prism base? Should it be?
4. Record the measured angle δ_m and the apex angle A . Use the formula

$$n = \frac{\sin\left(\frac{A + \delta_m}{2}\right)}{\sin\frac{A}{2}}$$

to calculate the index of refraction n . Compare your value with values given in Table 3-1. Does it agree with any value given there? What is your best guess for the prism material?

B. Total Internal Reflection (TIR)

(When you have finished this part, you will have three different traces of laser light interacting with right-angle prisms, all on an 11" × 17" sheet of white tracing paper.

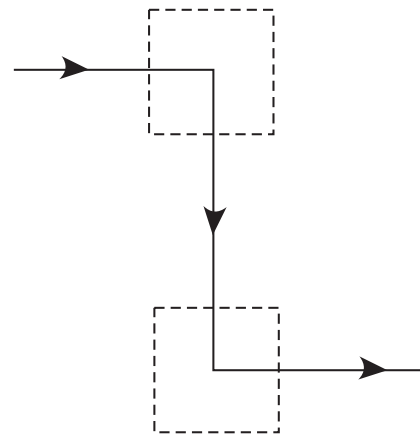
1. Set a right-angle prism on one of its parallel sides on a sheet of 11" × 17" white tracing paper. Tape the paper and prism in position. Shine a diode laser beam onto an appropriate face of the prism so that it *undergoes total internal reflection* (TIR) and exits the prism at 90° to its original direction of entry. Use index cards as a screen to "locate" the laser beam

outside the prism. On the paper, trace the edges of the prism, a line along the incident beam, a line along the path through the prism, and a line along the exit beam. Label the angles of incidence and reflection and their values at the face where TIR takes place. What would you need to know to determine the critical angle at this face? Is the incident angle on the face where TIR occurs larger or smaller than the critical angle?

2. Move the right-angle prism to a different position on the 11" × 17" paper and tape it down. Direct the diode laser beam onto an appropriate face so that the beam returns along a direction parallel to its entering direction. Use index cards to "locate" the beam path. When you have achieved this condition of *retroreflection*, trace the edges of the prism, the entering beam, the path through the prism, and the exit beam. Draw appropriate angles at the faces where TIR occurs and give their correct values.

3. Move two right-angle prisms to a new location on the 11" × 17" paper. Arrange them to produce "periscope action." This action requires, for example, that a horizontal beam that enters at one level be deflected downward 90° and exit horizontally at a different level, as shown in the accompanying sketch. Here the dashed squares indicate the locations of the two prisms. Use index cards to locate the beam through the prism arrangement.

When you have achieved the "periscope" geometry, tape the prisms down. Trace their edges, and trace the laser beam path from initial entry to final exit. Show where TIR occurs and label the incident and reflected angles there correctly, at each position.



C. Index of Refraction of a Thin Lens

Use the lensmaker's equation $\left[\frac{1}{f} = \left(\frac{n_g - n}{n} \right) \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \right]$ to determine the value of n_g for the

double-convex lens. Use a ruler, overhead lights, and an index card to obtain a good approximation for the focal length of the lens. (Going outside and imaging the sun would be even better.) Use a spherometer to measure the radii of curvature r_1 and r_2 . (You will have to be especially creative to get r_1 and r_2 if you don't have access to a spherometer.) With the values of f , r_1 , and r_2 , solve the lensmaker's equation for n_g , the index of refraction of the lens glass. Compare your value with values given in Table 3-1. Do you have a match?

D. Measuring the Focal Lengths of Thin Lenses

Set up the two diode lasers on a stand or optical bench so that they emit beams parallel to one another and normal to the plane defining the vertical position of the thin lens. See Figure L-2. (To see the beams converging on the image side of the lens, you will have to use chalk-dust particles or smoke—some form of cloud—to illuminate the path.) By moving the screen forward and backward, you can locate a position where the beams cross to form the "smallest" spot. This is the *focal point* for the lens. Measuring the optical bench distance from lens to focal

point gives the *focal length*. Compare this value to the value you obtained in part C, when you simply imaged a “distant” object on an index card. Which method is more accurate? Which method is easier?

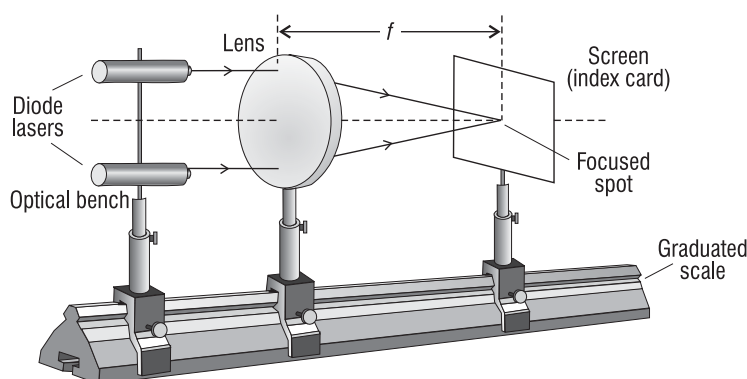


Figure L-2 Setup for determining focal length of a positive lens

Replace the positive lens in Figure L-2 with the negative lens. The challenge now is greater since the two laser beams diverge on the right side of the lens—and do not form a real image anywhere. Can you design a method to locate the spots of the two parallel beams at the lens and the spots for the two diverging beams on the right of the lens, then “trace” your way back to locate the focal point on the left side of the lens? If you can locate the focal point on the left, you can then measure its distance from the lens to get the focal length of the negative lens.

Student Project (optional)

Design a 10X beam expander using first a combination of two positive lenses and next a combination of a positive and a negative lens. Carefully draw each design to scale. Refer to publications such as the *Melles-Griot Catalog* or the *Edmund Scientific Industrial Optics Catalog* to obtain lens diameters, focal lengths, and approximate costs for each beam-expander design (less housing). Test each design on an optical bench and measure the size of the incident and exit beams. Determine how closely each beam expander meets the 10X specification. Is there any reason for choosing one design over the other?

Other Resources

- The Education Council of the Optical Society of America (OSA) has prepared a *discovery kit* designed to introduce students to modern optical science and engineering. The kit includes two thin lenses, a Fresnel lens, a mirror, a hologram, an optical illusion slide, a diffraction grating, one meter of optical fiber, two polarizers, four color filters, and instructions for eleven detailed experiments. OSA offers teacher membership opportunities. Contact the Optical Society of America, 2010 Massachusetts Avenue, NW, Washington, D.C. 20036, 800-762-6960.
- K-12 Optics Outreach kit, available from SPIE, Bellingham, Washington.

- Atneosen, R., and R. Feinberg. “Learning Optics with Optical Design Software,” *American Journal of Physics*, Vol 59, March 1991: pp 242-47.
- “Teaching Optics with an O/H Projector,” Douglas S. Goodman, Polaroid Corporation, 38 Henry St., Cambridge, Maryland.

References

Textbooks

- Beiser, Arthur. *Physics*, 3rd Edition, Menlo Park, California: The Benjamin/Cummings Publishing Company, 1982.
- Hecht, E., and A. Zajac. *Optics*, 2nd Edition. Reading, Massachusetts: Addison Wesley Publishing Company, 1987.
- Pedrotti, F., and L. Pedrotti. *Introduction to Optics*, 2nd Edition. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1993.
- Pedrotti, F., and L. Pedrotti. *Optics and Vision*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1998.
- Serway, R. A. *Principles of Physics*. Orlando, Florida: Saunders College Publishing, 1992.
- Waldman, Gary. *Introduction to Light*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1983.

Articles

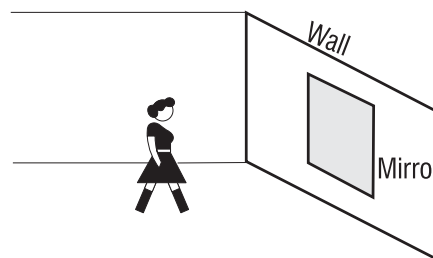
- Englert, B-G., M. O. Scully, and H. Walther. “The Duality in Matter and Light,” *Scientific American* (December 1994), 86.
- Weisskopf, Victor F. “How Light Interacts with Matter,” *Lasers and Light*, Readings from *Scientific American*. W. H. Freeman and Company, 1969, pp 14-26.

Optical Manufacturing Companies

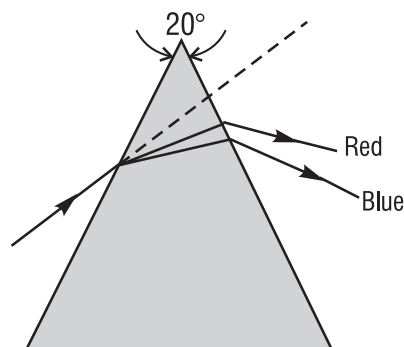
- Optics and Optical Instruments Catalog*. Edmund Scientific, Industrial Optics Division, Barrington, New Jersey.
- Melles-Griot Catalog*. Melles-Griot Optical Systems, Rochester, New York, and Melles-Griot Photonics Components, Irvine, California.

Problem Exercises

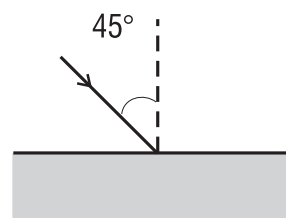
1. Use the *law of reflection* to determine the (a) minimum height and (b) position for a plane mirror that just allows a 5'6" woman standing on the floor in front of the mirror to see both her head and feet in the mirror. See sketch.



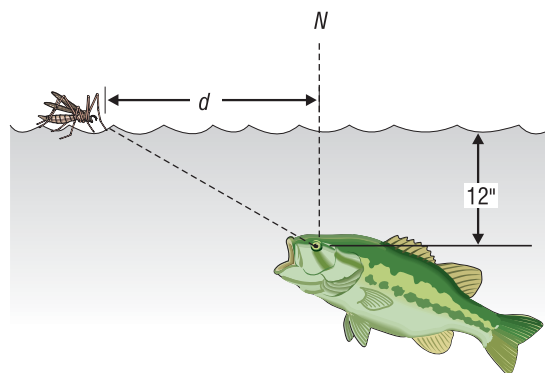
2. White light contains all wavelengths from deep blue at 400 nm to deep red at 700 nm. A narrow beam of collimated white light is sent through a prism of apex angle 20° as shown. The prism is made of light flint glass whose refractive index at 400 nm is 1.60 and at 700 nm is 1.565. What is the angular spread between the red and blue light at the minimum angle of deviation for each?



3. A ray of sodium light at 589 nm is incident on a rectangular slab of crown glass at an angle of 45° with the normal. (a) At what angle to the normal does this ray exit the slab? (b) What is the direction of the exiting ray relative to the entering ray? (c) Sketch an accurate trace of the ray through the slab.

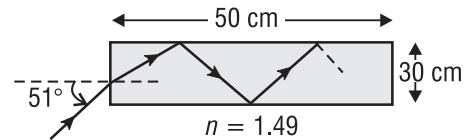


4. An object 3 cm high is placed 20 cm to the left of (a) a convex and (b) a concave spherical mirror, each of focal length 10 cm. Determine the position and nature of the image for each mirror.
5. Make a ray-trace diagram—on an $8\frac{1}{2}'' \times 11''$ sheet of white paper—that locates the image of a 2-cm object placed 10 cm in front of a concave spherical mirror of curvature 6 cm. Make your drawing *to scale*. Where is the image located and what are its orientation and its size? Repeat this for a convex spherical mirror of the same curvature.
6. A fish in a lake looks up at the surface of the water. At what distance d along the surface, measured from the normal, is a water-skimming insect safe from the roving eye of the fish?

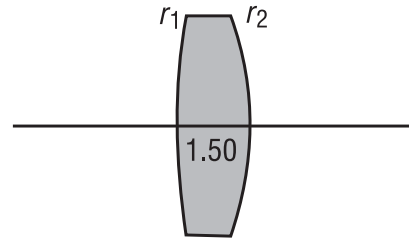


7. What is the light cone acceptance angle for an optical fiber of diameter $100\ \mu$, located in air, having a plastic core of index 1.49 and a plastic cladding of index 1.39? Make a sketch of the fiber, showing a limiting ray along the surface of the acceptance cone entering the fiber and refracting appropriately.

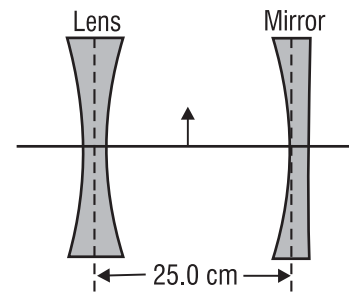
8. A laser beam is incident on the end face of a cylindrical rod of material as shown in the sketch. The refractive index of the rod is 1.49. How many internal reflections does the laser beam experience before it exits the rod?



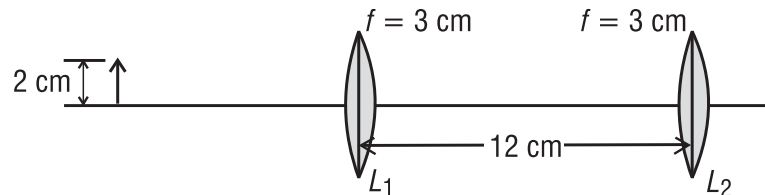
9. A thin, double-convex lens has a refractive index of 1.50. The radius of curvature of the front surface is 15 cm and that of the rear surface is 10 cm. See sketch. (a) How far from the lens would an image of the sun be formed? (b) How far from the lens would an image of a toy figure 24 cm from the lens be formed? (c) How do the answers to (a) and (b) change if you flip the lens over?



10. The object shown in the accompanying sketch is midway between the lens and the mirror. The radius of curvature of the mirror is 20 cm. The concave lens has a focal length of 16.7 cm. (a) Where is the light that travels first to the mirror and then to the lens finally imaged? (b) Where is the light finally imaged that travels first to the lens? (Note: Be especially careful of applying the sign convention!)



11. A ray of light makes an angle of incidence of 45° at the center of one face of a transparent cube of refractive index 1.414. Trace the ray through the cube, providing backup calculations to support your answer.
12. Two positive thin lenses, each of focal length $f = 3\text{ cm}$, are separated by a distance of 12 cm. An object 2 cm high is located 6 cm to the left of the first lens. See sketch.



On an $8\frac{1}{2}'' \times 11''$ sheet of paper, make a drawing of the two-lens system, *to scale*. (a) Use ray-tracing techniques to locate the final image and describe its size and nature. (b) Use the thin-lens equation to locate the position and size of the final image. How well do your results for (a) and (b) agree?

BASIC GEOMETRICAL OPTICS

13. A plano-convex lens of focal length 25.0 cm is to be made with crown glass of refractive index 1.52. Calculate the radius of curvature of the grinding and polishing tools to be used in making this lens.
14. An eyepiece is made of two positive thin lenses, each of focal length $f = 20$ mm, separated by a distance of 16 mm. (a) Where must a small object viewed by the eyepiece be placed so that the eye receives parallel light from the eyepiece? (b) Does the eye see an erect image relative to the object? Is it magnified? (c) Use a ray-trace diagram to answer these questions by inspection.

Basic Physical Optics

Leno S. Pedrotti

CORD

Waco, Texas

In Module 1-3, *Basic Geometrical Optics*, we made use of *light rays* to demonstrate *reflection* and *refraction* of light and the imaging of light with mirrors and lenses. In this module, we shift the emphasis from *light rays* to *light waves*—from *geometrical optics* to *physical optics*. In so doing, we move from a concern over the *propagation* of light energy along straight-line segments to one that includes the *spreading* of light energy—a fundamental behavior of all wave motion.

With *wave optics*—commonly referred to as *physical optics*—we are able to account for important phenomena such as *interference*, *diffraction*, and *polarization*. The study of these phenomena lays the foundation for an understanding of such devices and concepts as holograms, interferometers, thin-film interference, coatings for both antireflection (AR) and high reflection (HR), gratings, polarizers, quarter-wave plates, and laser beam divergence in the near and far field.

Prerequisites

Before you begin your study of this module, you should have completed a study of Module 1-1, *Nature and Properties of Light*, and Module 1-3, *Basic Geometrical Optics*. In addition, you should be able to use algebra, plane geometry, and trigonometry—especially the use and interpretation of the trigonometric functions ($\sin$, $\cos$, $\tan$) as they relate to sides and angles in triangles.

Objectives

When you finish this module you will be able to:

- Describe a *wave front*.
- Describe the relationship between *light rays* and *wave fronts*.
- Define *phase angle* and its relationship to a *wave front*.
- Calculate *water wave displacement* on a sinusoid-like waveform as a *function of time and position*.
- Describe how *electromagnetic waves* are *similar* to and *different* from *water waves*.
- State the *principle of superposition* and show how it is used to combine two overlapping waves.
- State *Huygens' principle* and show how it is used to predict the shape of succeeding wave fronts.
- State the *conditions required* for producing *interference patterns*.
- Define *constructive* and *destructive* interference.
- Describe a *laboratory setup* to produce a *double-slit interference pattern*.
- State the *conditions* for an *automatic phase shift of 180° at an interface between two optical media*.
- *Calculate the thickness of thin films* designed to *enhance or suppress* reflected light.
- Describe how *multilayer stacks* of quarter-wave films are used to *enhance or suppress* reflection over a *desired wavelength region*.
- Describe how *diffraction* differs from *interference*.
- Describe *single-slit diffraction* and *calculate positions of the minima* in the diffraction pattern.
- Distinguish between *Fraunhofer* and *Fresnel* diffraction.
- Sketch typical *Fraunhofer diffraction patterns* for a *single slit*, *circular aperture*, and *rectangular aperture*, and use equations to calculate *beam spread* and *fringe locations*.
- *Describe a transmission grating and calculate positions of different orders of diffraction*.
- Describe what is meant by *diffraction-limited optics* and describe the difference between a *focal point in geometrical optics* and a *focal-point diffraction pattern in wave optics*.
- Describe how *polarizers/analyzers* are used with polarized light.
- State the *Law of Malus* and explain how it is used to calculate intensity of polarized light passing through a polarizer with a tilted transmission axis.
- Calculate *Brewster's angle of incidence* for a given interface between two optical media.
- Describe how Brewster windows are used in a laser cavity to produce a linearly polarized laser beam.

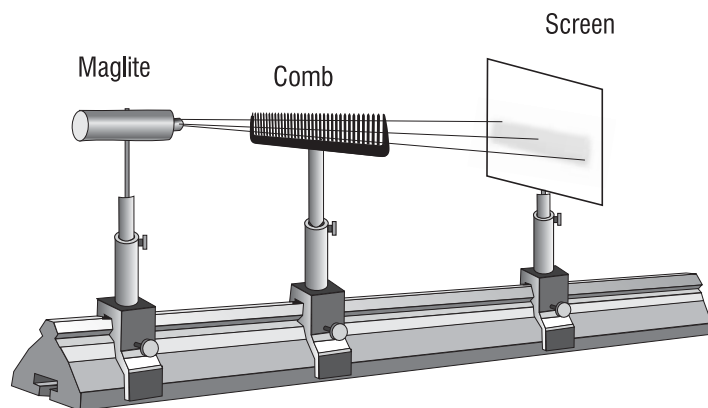
Scenario—Using Wave Optics in the Workplace

Letitia works for an optical coating company that produces highly transmissive and highly reflecting optics. For the past several weeks she has been working on protective overcoats for metallic gold mirrors. The overcoats are made of multilayer dielectric stacks that preserve the required reflective properties of the mirrors while protecting the soft gold surface from scratches and digs. Letitia remembers her work in wave optics at school, where she learned about quarter-wave plates, AR and HR coats, and surface properties of metallic reflectors. She is both pleased and surprised at how much she remembers about light interference in thin films and how much more interesting this makes her work. Today she is working in the coating lab with other technicians, preparing a multilayer dielectric quarter-wave stack, made up of alternate layers of high- and low-index-of-refraction materials to enhance the reflection of light near 550 nm. Letitia knows that her time in school prepared her to understand the principles of wave optics and also to learn valuable hands-on skills in the laboratory. She feels that she is becoming a “coating” expert.

Opening Demonstrations

Note: The hands-on exercises that follow are to be used as short, introductory laboratory demonstrations. They are intended to provide you with a glimpse of several phenomena that are dependent on wave optics and stimulate your interest in the study of optics and photonics

1. Shining White Light Through a Comb. In an appropriately darkened room, shine light from a focusable mini Mag-Lite (Mag Instrument, Ontario, California, 909-947-1006) through the narrowly spaced teeth of an ordinary comb. Mount the Mag-Lite and comb firmly on an optical bench with appropriate holders. Examine the light pattern on a white screen, securely mounted several feet from the comb. See sketch below. Describe in detail what is seen on the screen. Can *geometrical optics* account for what is observed?



D-1 Setup for observing white light through the teeth of a comb

2. Shining Laser Light Through a Transmission Grating. Replace the Mag-Lite above with an ordinary low-power (5 mW or less) diode laser and the comb with a transmission grating (around 5000 lines/inch). Observe the pattern produced by the light passing through the grating,

first on the screen and then on a distant wall. Describe in detail what is observed. Can *geometrical optics* account for the patterns observed?

3. Shining Laser Light Through a Pinhole. Arrange a 5-mW diode laser, pinhole (50 micrometers or so in diameter), and screen along an optical bench. Carefully align the laser beam so that it falls perpendicularly on the tiny pinhole. Observe the light that passes through the pinhole on a white cardboard screen. Make minor adjustments to the relative positions of the laser and pinhole to obtain the brightest pattern on the screen. Move the screen far enough away so you can see clearly (in a darkened room) the details of the light pattern. Describe what you see. Can *geometrical optics* account for the light pattern?

Basic Concepts

I. LIGHT WAVES AND PHYSICAL OPTICS

In our study of ray optics and image formation, we represented image points as “geometrical points,” without physical extent. That, of course, followed logically since light rays were used to locate the image points and light rays are *lines* that intersect clearly at geometrical points. But in reality, if you were to examine such image points with a microscope, you would see *structure* in the “point,” a structure explained only when you invoke the true *wave nature* of light.

In effect, then, we are saying that, with large objects such as prisms, mirrors, and lenses—large in the sense that their dimensions are millions of times that of the wavelength of light—interference and diffraction effects are still present in the imaging process, but they occur on so small a scale as to be hardly observable to the naked eye. To a good approximation, then, with “large” objects we are able to describe light imaging quite satisfactorily with geometrical (ray) optics and obtain fairly accurate results. But when light waves pass around small objects, such as a 100- μ -diameter human hair, or through small openings, such as a 50- μ pinhole, ray optics *cannot* account for the light patterns produced on a screen beyond these objects. Only wave optics leads to the correct interpretation of such patterns.

And so now we turn to a study of the wave nature of light and to the fascinating phenomena of *interference*, *diffraction*, and *polarization*—and of such devices as gratings and thin-film coatings. We shall see that interference occurs when two or more light waves pass through the same region and add to or subtract from each other. Diffraction occurs when light waves pass through small openings or around small obstacles and spread, and polarization occurs due to the transverse nature of the electric field vibration in a propagating electromagnetic wave. Before we look at these phenomena, let’s review briefly the nature of waves, wave fronts, and wave motion.

A. Physics of waves and wave motion

Wave optics treats light as a series of propagating electric and magnetic field oscillations. While we cannot see these extremely rapid oscillations, their wave behavior is similar to that of water waves. Thus, we find it useful to picture waves and wave motion in terms of simple water waves, such as those created by a bobbing cork on an otherwise quiet pond. See Figure 4-1a.

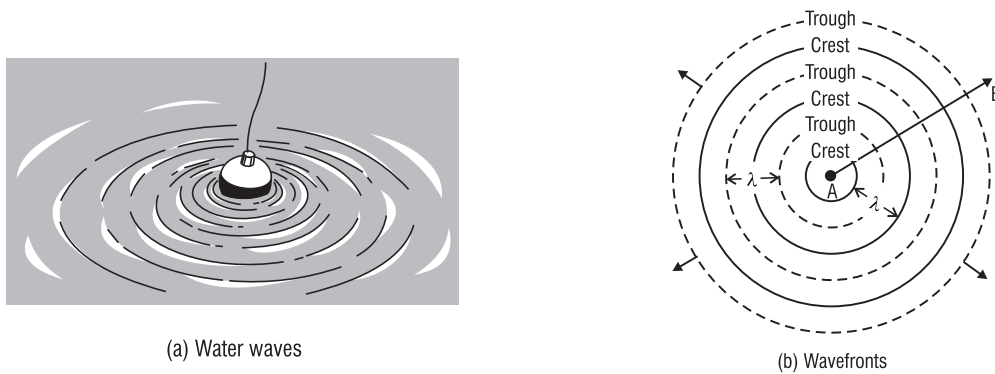
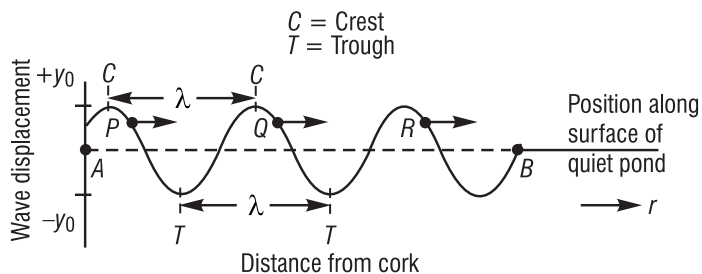


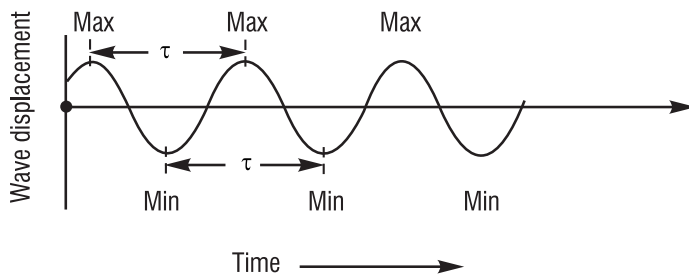
Figure 4-1 Water waves and wave fronts

The bobbing cork generates a series of surface disturbances that travel outward from the cork. Figure 4-1b shows the same disturbances traveling away from point A (the cork) as a series of successive *wave fronts* labeled *crests* and *troughs*. Recall that a *wave front* is a locus of points along which all phases and displacements are identical. The solid circles in Figure 4-1b depict the outward-moving wave crests; the dashed circles represent wave troughs. Adjacent crests are always a *wavelength* apart, as are the adjacent troughs.

If we were able to look along the surface of the pond, we would see a *sinusoid-like* profile of the traveling wave such as that shown in Figure 4-2a. The profile is a *snapshot* of the water displacement at a *certain instant of time* along a direction such as AB, labeled back in Figure 4-1b. The water surface rises to a maximum displacement ($+y_0$) and falls to a minimum displacement ($-y_0$) along the profile. As time varies, the “snapshot” profile in Figure 4-2a moves to the right with its characteristic wave speed. The radial distance outward from the cork at position A, shown in Figure 4-1b, is denoted by the variable r in Figure 4-2a.



(a) Wave profile along the pond at a *certain instant of time*



(b) Wave displacement at a *fixed* position on the pond as a function of time

Figure 4-2 Two aspects of wave motion for a traveling wave

Now suppose that—instead of looking along the surface of the pond—we look at the moving wave at one *definite position* on the pond, such as at point Q in Figure 4-2a. What happens to the wave displacement at this fixed position as the wave disturbances move away from the cork? We know from experience that the surface of the pond at Q rises and falls, repeatedly—as long as the wave disturbances move past this position. This wave displacement as a function of time—at a *fixed position*—is shown in Figure 4-2b. Note again that the shape is *sinusoid-like*.

Since we’re concentrating on one position in Figure 4-2b, we cannot “see” the whole wave. All we see is the *up and down* motion of point Q. The time between successive maxima or successive minima is defined as the *period* (τ) of the wave. The number of times point Q goes from max to min to max per second is called the frequency (f) of the wave. The period τ and the frequency f are related by the simple relationship $f = 1/\tau$, as presented in Module 1-1, *Nature and Properties of Light*.

B. The mathematics of sinusoidal waveforms (optional)*

The two aspects of wave motion depicted in Figures 4-2a and 4-2b—one at a fixed time, the other at a fixed position—are addressed in a mathematical equation that describes a *sinusoidally varying* traveling wave. Refer to Equation 4-1,

$$y(r, t) = y_0 \sin \left[\frac{2\pi}{\lambda} (r - vt) \right] \quad (4-1)$$

where: $y(r, t)$ is the wave displacement at position r and time t

y_0 is the wave amplitude as shown in Figure 4-2a

λ is the wavelength

r is the position along the traveling wave

v is the wave speed, equal to $\lambda \times f$, and

t is the time

If we “freeze” time at some value t_0 , for example, we obtain the specialized equation

$y(r, t_0) = y_0 \sin \left[\frac{2\pi}{\lambda} (r - \text{constant}) \right]$. This is a mathematical description of the wave profile

shown in Figure 4-2a. On the other hand, if we select a fixed position r_0 , we obtain another

specialized equation $y(r_0, t) = y_0 \sin \left[\frac{2\pi}{\lambda} (\text{constant} - vt) \right]$. This is a mathematical description of

the waveform shown in Figure 4-2b.

The factor in brackets in Equation 4-1 defines the *phase angle* ϕ of the wave at position r and time t . Thus,

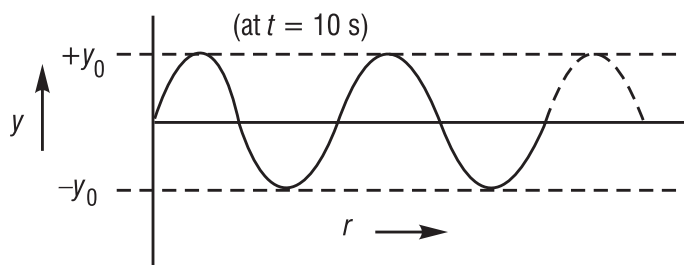
*The text material in this section, through Example 1, is optional. Depending on the background of the class, this section may or may not be covered.

$$\phi = \left[\frac{2\pi}{\lambda}(r - vt) \right] \quad (4-2)$$

The phase angle is the *same for any point on a given wave front*, as mentioned earlier. For example, for successive wave fronts whose values of ϕ are $[\pi/2]$, $[\pi/2 + 2\pi]$, $[\pi/2 + 4\pi]$, and so on—always 2π radians (360°) apart— $\sin \phi$ for *each of these angles* equals $+1$, so that $y(r, t)$ equals $+y_0$, a maximum *positive* displacement. Such wave fronts are crests. Similarly, for successive wave fronts whose values of ϕ are $[3\pi/2]$, $[3\pi/2 + 2\pi]$, $[3\pi/2 + 4\pi]$, etc., always 2π radians apart, $\sin \phi$ for *each of these angles* equals -1 , so that $y(r, t)$ equals $-y_0$, the maximum *negative* wave displacement. Such wave fronts are troughs. And so it goes for all other wave fronts between the crests and troughs. For example, points P , Q , and R in Figure 4-2a, all with the same wave displacement, represent wave fronts a wavelength apart with phase angles of values differing by 2π . Example 1 provides an application of Equations 4-1 and 4-2 to circular water waves on a quiet pond.

Example 1

Circular water waves such as those shown in Figures 4-1a and 4-1b move outward from a bobbing cork at A . The cork bobs up and down and back again—a complete cycle—once per second, and generates waves that measure 10 cm from crest to crest. Some time after the wave motion has been established, we begin to time the motion with a stopwatch. At a certain time $t = 10$ s on the watch, we notice that the wave profile has the shape shown below.



- What is the wave frequency f for this water wave?
- What is its wavelength λ ?
- What is its wave speed v ?
- What is the phase angle ϕ for a wave front at position $r = 102.5$ cm at time $t = 10$ s?
- What is the wave displacement y on the wave front at $r = 102.5$ cm?
- What is the phase angle ϕ for a wave front at $r = 107.5$ cm at $t = 10$ s?
- What is the wave displacement y on the wave front at $r = 107.5$ cm?
- If we focus on the wave motion at the position $r = 105$ cm and let time vary, what kind of motion do we observe?

Solution:

- The wave frequency is 1 cycle/s; (therefore, the period $\tau = 1/f$ is 1 second).

(b) The wavelength λ is the crest-to-crest distance, thus $\lambda = 10$ cm.

(c) The wave speed $v = \lambda \times f = 10 \text{ cm} \times 1/\text{s} = 10 \text{ cm/s}$.

(d) At $t = 10$ s, $r = 102.5$ cm and $v = 10$ cm/s. Using $\phi = \left[\frac{2\pi}{\lambda}(r - vt) \right]$, we get

$$\phi = \frac{2\pi}{10}(102.5 - 10 \times 10) = \frac{2\pi}{10}(2.5) = \pi/2 \text{ rad, an angle of } 90^\circ$$

(e) $y = y_0 \sin \phi = y_0 \sin(\pi/2) = y_0 \sin(90^\circ) = y_0$, since $\sin 90^\circ = 1$. Since $y = y_0$ at this location and y_0 is the maximum positive displacement, the circular wave front is a *crest*.

(f) At $t = 10$ s, $r = 107.5$ cm and $v = 10$ cm/s. Using the expression for the phase angle ϕ , we get

$$\phi = \frac{2\pi}{10}(107.5 - 10 \times 10) = \frac{2\pi}{10}(7.5) = 3\pi/2, \text{ an angle of } 270^\circ$$

(g) $y = y_0 \sin \phi = y_0 \sin\left(\frac{3\pi}{2}\right) = y_0 \sin(270^\circ) = -y_0$, since $\sin 270^\circ = -1$. Since $y = -y_0$, a maximum negative displacement, the circular wave front at $r = 107.5$ cm is a *trough*.

(h) At $r = 105$ cm, we see the water move up and down, repeatedly, between displacements of $(+y_0)$ and $(-y_0)$, completing a cycle of motion once per second. Thus, the frequency of this vertical motion is 1 cycle/s and its period is 1 s.

Before we leave this section, we need to make a connection between the wave motion we are studying here with water waves and the wave motion of light waves. For light waves it is the electric field and magnetic field that vary between positive and negative maxima—in a direction *transverse* to (perpendicular to) the direction of propagation just as the vertical displacement of the water does for water waves. Figure 4-3 shows a profile of the transverse electric field E and magnetic field B at one instant of time. It is easy to see the *sinusoidal form* of the varying E and B values, much like the sinusoidal form of the varying displacement values for the water wave in Figure 4-2a. When we study interference, diffraction, and polarization, we can ignore the B-field and concentrate only on the varying E-field.

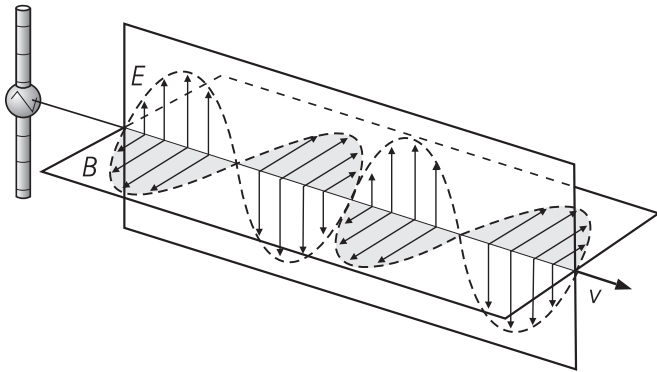


Figure 4-3 Profiles of the electric and magnetic fields in a light wave at an instant of time

II. INTERACTION OF LIGHT WAVES

A. The principle of superposition

An understanding of light wave interference begins with an answer to the question, “What happens at a certain position in space when two light waves pass through that position at the same time? To answer this question, we invoke the *principle of superposition*, which states:

When two or more waves move simultaneously through a region of space, each wave proceeds independently as if the other were not present. The resulting wave “displacement” at any point and time is the vector sum of the “displacements” of the individual waves.

This principle holds for water waves, mechanical waves on strings and on springs (the Slinky!), and for sound waves in gases, liquids and solids. Most important for us, it holds for all electromagnetic waves in free space. So, if we have two light waves passing through some common point P , where Wave 1 alone causes a “displacement” Y_1 and Wave 2 alone a displacement Y_2 , the *principle of superposition* states that the resultant displacement Y_{RES} is given by a *vector sum* of the two displacements. If both displacements are along the same direction—as they will be for most applications in this module—we can add the two displacements *algebraically*, as in Equation 4-3.

$$Y_{RES} = Y_1 + Y_2 \quad (4-3)$$

An application of Equation 4-3 is shown in Figure 4-4, where Wave 1 and Wave 2 are moving along the x -direction to the right. Wave 2 is drawn with $\frac{3}{4}$ the amplitude and $\frac{1}{2}$ the wavelength of Wave 1. The resultant wave, obtained by applying Equation 4-3 at each point along the x -direction, is shown by the solid waveform, Y_{RES} .

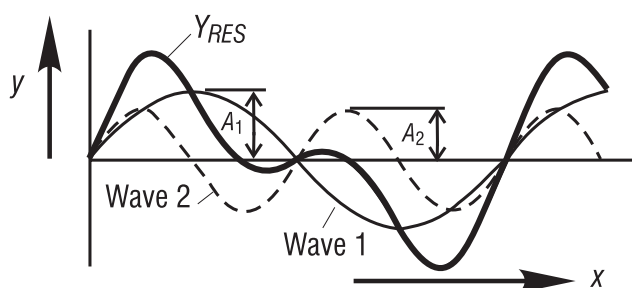


Figure 4-4 Superposition of two waves moving along the same direction

In Figure 4-5, we show the interference of two sinusoidal waves of the same amplitude and same frequency, traveling in the same direction. The two waves are represented by the light solid and broken curves, the resultant by the solid heavy curve. In Figure 4-5a the two waves are *exactly in phase*, with their maximum and minimum points matching perfectly. Applying the *principle of superposition* to the two waves, the resultant wave is seen to have the same amplitude and frequency but *twice* the amplitude $2A$ of either initial wave. This is an example of *constructive interference*. In Figure 4-5b the two curves are *exactly out of phase*, with the crest of one falling on the trough of the other, and so on. Since one wave effectively cancels the

effect of the other at each point, the resultant wave has zero *displacement* everywhere, as indicated by the solid black line. This is an example of *destructive interference*. In Figure 4-5c, the two waves are neither completely in phase nor completely out of phase. The resultant wave then has an amplitude somewhere between A and $2A$, as shown.

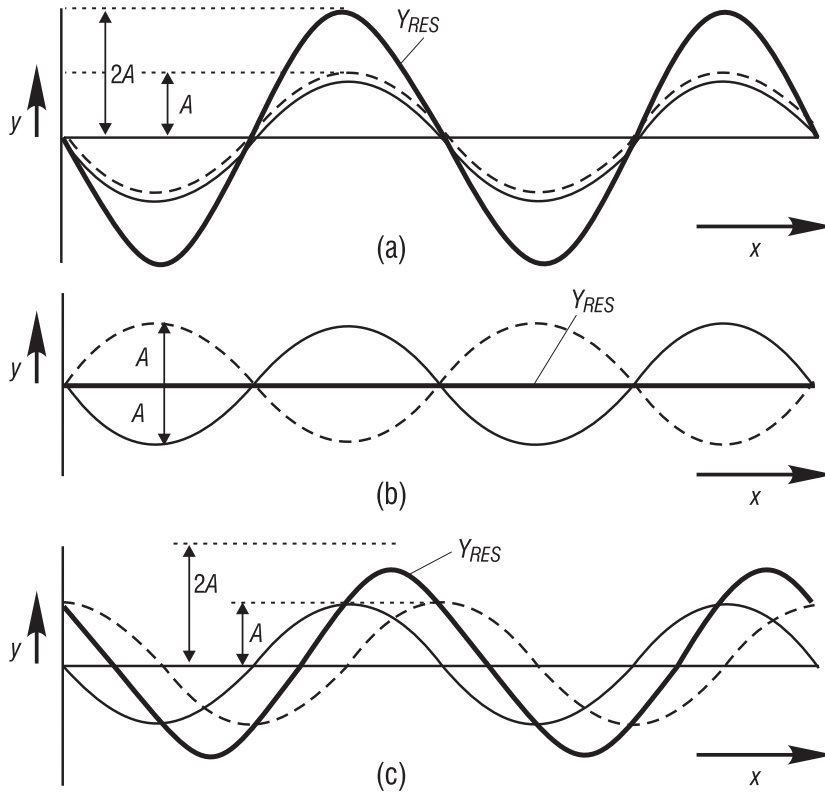


Figure 4-5 Interference of two identical sinusoidal waves

B. Huygens' wavelets

Long before people understood the electromagnetic character of light, Christian Huygens—a 17th-century scientist—came up with a technique for propagating waves from one position to another, determining, in effect, the shapes of the developing wave fronts. This technique is basic to a quantitative study of interference and diffraction, so we cover it here briefly. Huygens claimed that:

Every point on a known wave front in a given medium can be treated as a point source of secondary wavelets (spherical waves “bubbling” out of the point, so to speak) which spread out in all directions with a wave speed characteristic of that medium. The developing wave front at any subsequent time is the envelope of these advancing spherical wavelets.

Figure 4-6 shows how Huygens' principle is used to demonstrate the propagation of successive (a) *plane* wave fronts and (b) *spherical* wave fronts. Huygens' technique involves the use of a series of points $P_1 \dots P_8$, for example, on a given wave front defined at a time $t = 0$. From these points—as many as one wishes, actually—spherical wavelets are assumed to emerge, as shown in Figures 4-6a and 4-6b. Radiating outward from each of the P -points, with a speed v , the

series of secondary wavelets of radius $r = vt$ defines a new wave front at some time t later. In Figure 4-6a the new wave front is drawn as an *envelope tangent* to the secondary wavelets at a distance $r = vt$ from the initial plane wave front. It is, of course, another *plane* wave front. In Figure 4-6b, the new wave front at time t is drawn as an *envelope tangent* to the secondary wavelets at a distance $r = vt$ from the initial spherical wave front. It is an advancing *spherical* wave front.

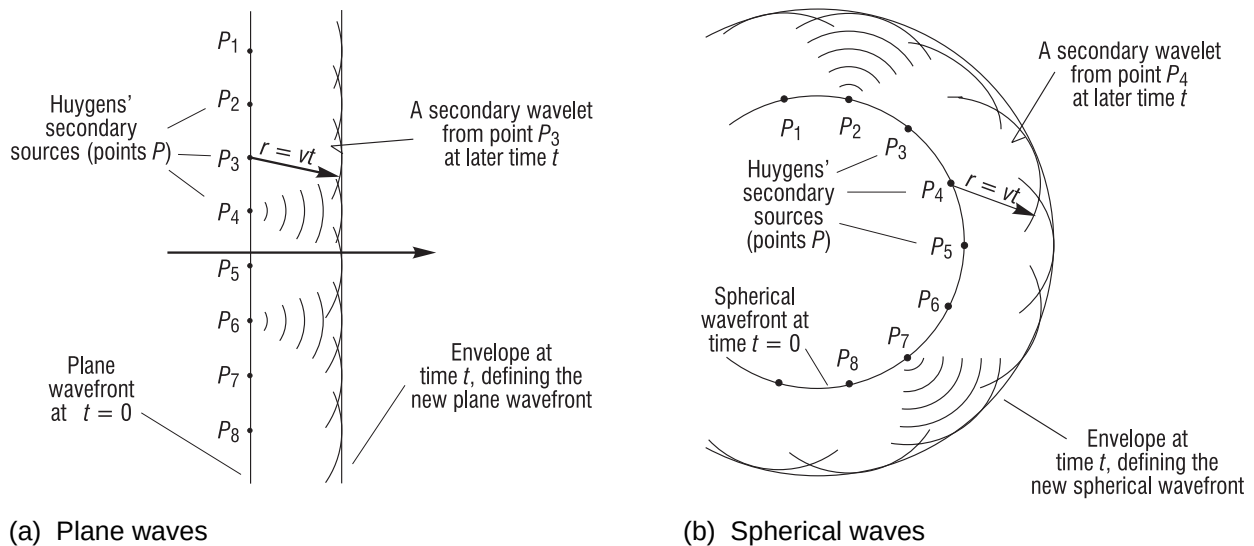


Figure 4-6 Huygens' principle applied to the propagation of plane and spherical wave fronts

While there seems to be no physical basis for the existence of Huygens' "secondary" point sources, Huygens' technique has enjoyed extensive use, since it does predict accurately—with waves, not rays—both the *law of reflection* and *Snell's law of refraction*. In addition, Huygens' principle forms the basis for calculating, for example, the diffraction pattern formed with multiple slits. We shall soon make use of Huygens' secondary sources when we set up the problem for diffraction from a single slit.

III. INTERFERENCE

Today we produce interference effects with little difficulty. In the days of Sir Isaac Newton and Christian Huygens, however, light interference was not easily demonstrated. There were several reasons for this. One was based on the extremely short wavelength of visible light—around 20 millionths of an inch—and the obvious difficulty associated with seeing or detecting interference patterns formed by overlapping waves of so short a wavelength, and so rapid a vibration—around a million billion cycles per second! Another reason was based on the difficulty—before the laser came along—of creating *coherent* waves, that is, waves with a phase relationship with each other that remained *fixed* during the time when interference was observed.

It turns out that we *can* develop phase coherence with *nonlaser* light sources to demonstrate interference, but we must work at it. We must "prepare" light from readily available incoherent light sources—which typically emit individual, uncoordinated, short wave trains of fixed phase

of no longer than 10^{-8} seconds—so that the light from such sources remains coherent over periods of time long enough to overlap and produce visible interference patterns. There are generally two ways to do this.

- Develop several coherent *virtual* sources from a single incoherent “point” source with the help of mirrors. Allow light from the two virtual sources to overlap and interfere. (This method is used, for example, in the Lloyd’s mirror experiment.)
- Take monochromatic light from a single “point” source and pass it through two small openings or slits. Allow light from the two slits to overlap on a screen and interfere.

We shall use the second of these two methods to demonstrate Thomas Young’s famous *double-slit experiment*, worked out for the first time at the very beginning of the 19th century. But first, let’s consider the basics of interference from two point sources.

A. Constructive and destructive interference

Figure 4-7 shows two “point” sources of light, S and S' , whose radiating waves maintain a fixed phase relationship with each other as they travel outward. The emerging waves are in effect spherical, but we show them as circular in the two-dimensional drawing. The solid circles represent crests, the dashed circles, troughs.

Earlier, in Figure 4-5a, we saw the effect of *constructive interference* for waves perfectly in phase and, in Figure 4-5b, the effect of *destructive interference* for waves perfectly out of phase. In Figure 4-7, along directions OP , OP_2 , and OP_2' (emphasized by solid dots) crests from S and S' meet (as do the troughs), thereby creating a condition of *constructive interference*. As a result, light striking the screen at points P , P_2 , and P_2' is at a maximum intensity and a bright spot appears. By contrast, along directions OP_1 and OP_1' (emphasized by open circles) crests and troughs meet each other, creating a condition of *destructive interference*. So at points P_1 and P_1' on the screen, no light appears, leaving a dark spot.

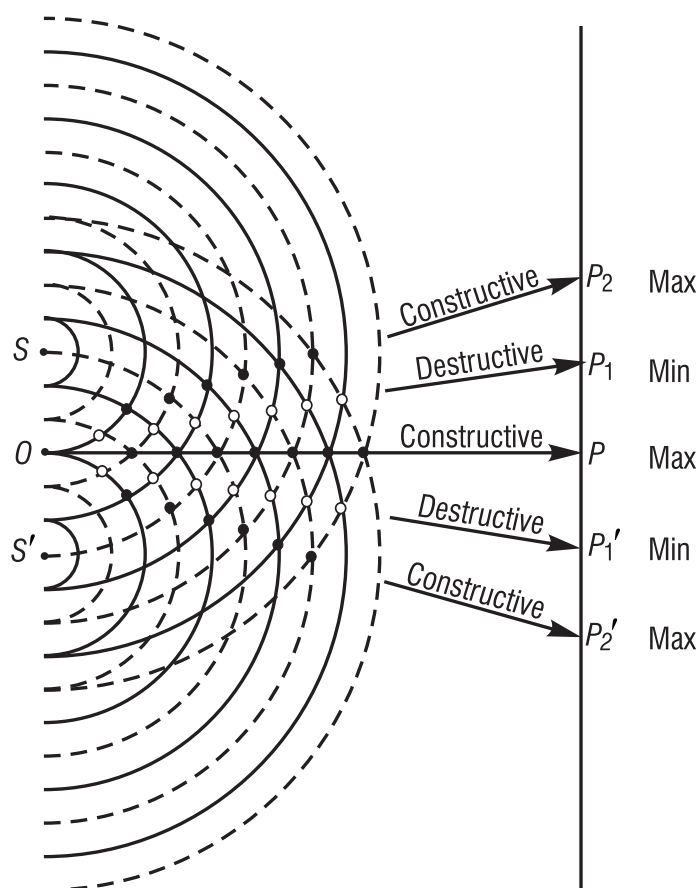


Figure 4-7 Wave interference created by overlapping waves from coherent sources S and S'

The requirement of *coherent sources* is a stringent requirement if interference is to be observed. To see this clearly, suppose for a moment that sources S and S' in Figure 4-7 are, in fact, *two corks* bobbing up and down on a quiet pond. As long as the two corks maintain a *fixed* relationship between their vertical motions, each will produce a series of *related* crests and troughs, and observable interference patterns in the overlap region will occur. But if the two corks bob up and down in a *random, disorganized manner*, no series of related, fixed-phase crests and troughs will form and no interference patterns of sufficiently long duration can develop, and so interference will not be observed.

B. Young's double-slit interference experiment

Figure 4-8a shows the general setup for producing *interference* with coherent light from two slits S_1 and S_2 . The source S_0 is a monochromatic point source of light whose spherical wave fronts (circular in the drawing) fall on the two slits to create secondary sources S_1 and S_2 . Spherical waves radiating out from the two secondary sources S_1 and S_2 maintain a fixed phase relationship with each other as they spread out and overlap on the screen, to produce a series of alternate bright and dark regions, as we saw in Figure 4-7. The alternate regions of bright and dark are referred to as *interference fringes*. Figure 4-8b shows such interference fringes, greatly expanded, for a small central portion of the screen shown in Figure 4-8a.

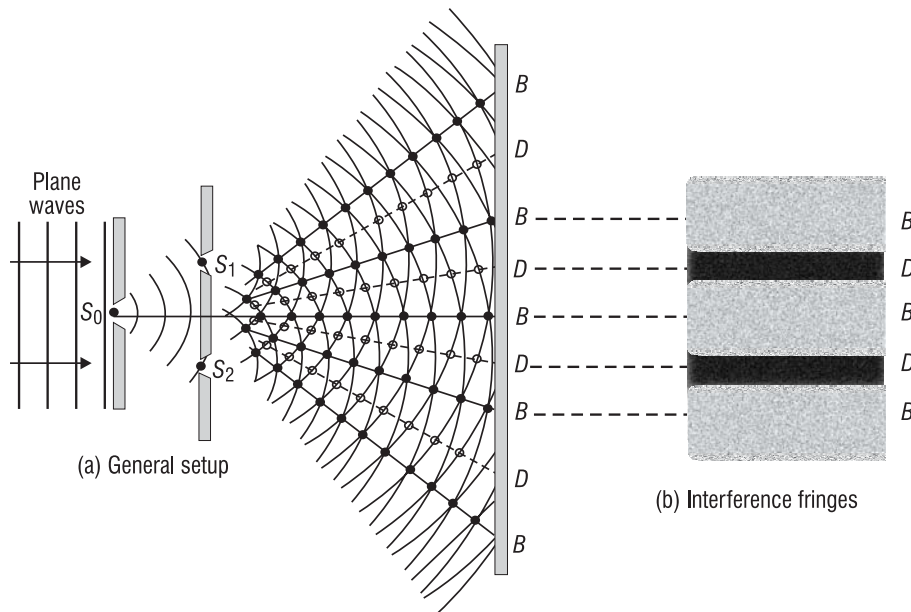


Figure 4-8 Young's double-slit interference experiment showing (a) general setup and (b) typical interference fringes

1. Detailed analysis of interference from a double slit: With the help of the *principle of superposition*, we can calculate the positions of the alternate maxima (bright regions) and minima (dark regions) shown in Figure 4-8. To do this we shall make use of Figure 4-9 and the following conditions:

- (a) Light from slits S_1 and S_2 is coherent; that is, there exists a fixed phase relationship between the waves from the two sources.
- (b) Light from slits S_1 and S_2 is of the same wavelength.

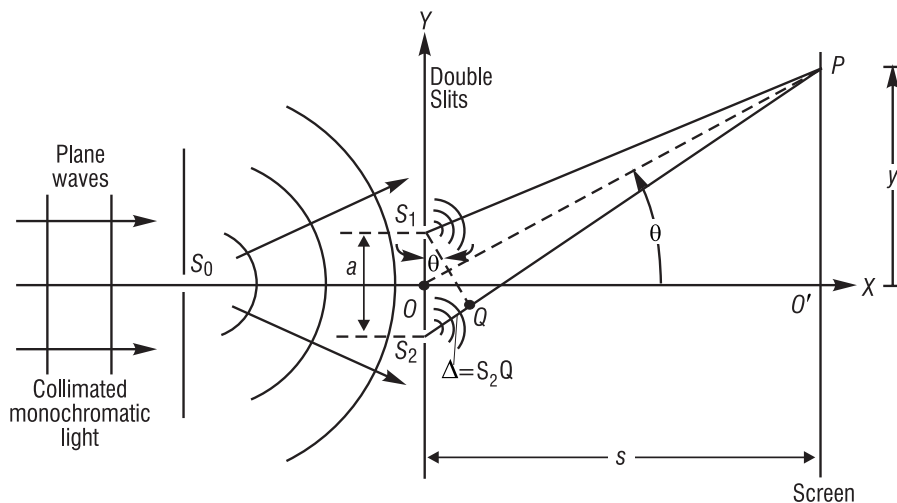


Figure 4-9 Schematic for double-slit interference calculations. Source S_0 is generally a small hole or narrow slit; sources S_1 and S_2 are generally long, narrow slits perpendicular to the page.

In Figure 4-9, light waves from S_1 and S_2 spread out and overlap at an arbitrary point P on the screen. If the overlapping waves are in phase, we expect a bright spot at P ; if they are out of phase, we expect a dark spot. So the *phase difference* between the two waves arriving at point P is a key factor in determining what happens there. We shall express the phase difference in terms of the *path difference*, which we can relate to the wavelength λ .

For clarity, Figure 4-9 is not drawn to scale. It will be helpful in viewing the drawing to know that, in practice, the distance s from the slits to the screen is about *one meter*, the distance a between slits is *less than a millimeter*, so that the angle θ in triangle S_1S_2Q , or triangle OPO' , is quite small. And on top of all this, the wavelength of light is a fraction of a micrometer.

The path difference Δ between S_1P and S_2P , as seen in Figure 4-9, is given by Equation 4-4, since the distances PS_1 and PQ are equal and since $\sin \theta = \Delta/a$ in triangle S_1S_2Q .

$$\Delta = S_2P - S_1P = S_2Q = a \sin \theta \quad (4-4)$$

If the path difference Δ is equal to λ or some integral multiple of λ , the two waves arrive at P in phase and a bright fringe appears there (constructive interference). The condition for bright (B) fringes is, then,

$$\Delta_B = a \sin \theta = m\lambda \text{ where } m = 0, \pm 1, \pm 2, \dots \quad (4-5)$$

The number m is called the *order number*. The central bright fringe at $\theta = 0$ (point O' on the screen) is called the zeroth-order maximum ($m = 0$). The first maximum on either side, for which $m = \pm 1$, is called the *first-order maximum*, and so on.

If, on the other hand, the path difference at P is an odd multiple of $\lambda/2$, the two waves arrive out of phase and create a dark fringe (destructive interference). The condition for dark (D) fringes is given by Equation 4-6.

$$\Delta_D = a \sin \theta = (m + \frac{1}{2})\lambda \text{ where } m = 0, \pm 1, \pm 2, \dots \quad (4-6)$$

Since the angle θ exists in both triangles S_1S_2Q and OPO' , we can find an expression for the *positions* of the bright and dark fringes along the screen. Because θ is small, as mentioned above, we know that $\sin \theta \cong \tan \theta$, so that for triangle OPO' we can write

$$\sin \theta \cong \tan \theta = \frac{y}{\lambda s} \quad (4-7)$$

Combining Equation 4-7 with Equations 4-5 and 4-6 in turn, by substituting for $\sin \theta$ in each, we obtain expressions for the position y of bright and dark fringes on the screen.

$$y_B = \frac{\lambda s}{a} m \text{ where } m = 0, \pm 1, \pm 2, \dots \quad (4-8)$$

and

$$y_D = \frac{\lambda s}{a} \left(m + \frac{1}{2}\right) \text{ where } m = 0, \pm 1, \pm 2, \dots \quad (4-9)$$

In Example 2, through the use of Equation 4-8, we recreate the method used by Thomas Young to make the *first* measurement of the wavelength of light.

Example 2

A double-slit source with slit separation 0.2 mm is located 1.2 m from a screen. The distance between successive bright fringes on the screen is measured to be 3.30 mm. What is the wavelength of the light?

Solution: Using Equation 4-8 for any two adjacent bright fringes, we can obtain an equation for Δy , the fringe separation. Thus,

$$\Delta = (y_B)_{m+1} - (y_B)_m = \frac{\lambda s(m+1)}{a} - \frac{\lambda s(m)}{a} = \frac{\lambda s}{a}$$

$$\therefore \Delta y = \frac{\lambda s}{a}, \text{ so that } \lambda = \frac{(\Delta y)a}{s}, \text{ giving}$$

$$\lambda = \frac{(3.30 \times 10^{-3} \text{ m})(2 \times 10^{-4} \text{ m})}{1.2 \text{ m}} = 5.5 \times 10^{-7} \text{ m} = 550 \times 10^{-9} \text{ m}$$

So the wavelength is about 550 nm and the light is yellowish green in color.

2. Intensity variation in the interference pattern. Knowing how to locate the positions for the fringes on a screen, we might now ask, “How does the brightness (intensity) of the fringes vary as we move, in either direction, from the central bright fringe ($m = 0$)?” We obtain a satisfactory answer to this question by representing the two separate electric fields at point P , the one coming from S_1 as $E_1 = E_0 \sin 2\pi ft$ and the one from S_2 as $E_2 = E_0 \sin (2\pi ft + \delta)$. The waves are assumed to have the same amplitude E_0 . Here δ is the *phase angle* difference between the two waves arriving at P . The *path difference* Δ is related to the *phase angle* δ by the relationship

$$\frac{\delta}{\Delta} = \frac{2\pi}{\lambda} \quad (4-10)$$

so that if $\Delta = \lambda$, $\delta = 2\pi \text{ rad} = 360^\circ$, if $\Delta = \lambda/2$, $\delta = \pi \text{ rad} = 180^\circ$, and so on.

Then, by using the *principle of superposition*, we can add the two electric fields at point P to obtain $E_{RES} = E_1 + E_2$. (Carrying out this step involves some trigonometry, the details of which can be found in most optics texts.) Since the intensity I of the light goes as the *square* of the electric field E , we square E_{RES} and *average* the result over one cycle of wave oscillation at P , obtaining, finally, an expression for the average intensity, I_{AV} .

$$I_{AV} = I_0 \cos^2 \frac{\delta}{2} \quad (4-11)$$

Here δ is the critical *phase angle difference* at point P . For all points P for which $\delta = 0, 2\pi, 4\pi$, and so on, corresponding to $\Delta = 0, \lambda, 2\lambda$, etc., $\cos^2\left(\frac{\delta}{2}\right) = 1$ and $I_{AV} = I_0$, the maximum possible “brightness.” At these points, bright fringes form. For $\delta = \pi, 3\pi, 5\pi$, and so on, corresponding to $\Delta = \lambda/2, 3\lambda/2, 5\lambda/2$, etc., $\cos^2\left(\frac{\delta}{2}\right) = 0$, and dark fringes form.

The maximum intensity I_0 is equal to $(E_0 + E_0)^2$ or $4E_0^2$, since each wave has amplitude E_0 . Further, from Equations 4-10 and 4-4, we see that

$$\delta = \frac{2\pi}{\lambda} \Delta = \frac{2\pi}{\lambda} a \sin \theta \quad (4-12)$$

so that the phase angle δ is connected clearly through the angle θ to different points P on the screen. Going one step further, replacing $\sin \theta$ by $\frac{y}{s}$ in Equation 4-12, we have the connection between δ and any position y on the screen, such that

$$\delta = \frac{2\pi a}{\lambda s} y \quad (4-13)$$

With Equation 4-13 and $I_0 = 4E_0^2$, we can rewrite Equation 4-11 in a form that relates I_{AV} directly to a position y on the screen.

$$I_{AV} = 4E_0^2 \cos^2\left(\frac{\pi a}{\lambda s} y\right) \quad (4-14)$$

where: I_{AV} = intensity of light along screen at position y

E_0 = amplitude of light wave from S_1 or S_2

s = distance from the plane of the double slit to the screen

a = slit separation

λ = wavelength of monochromatic light

y = distance above (or below) central bright fringe on the screen

Example 3

Using Equation 4-14 and the double-slit arrangement described in Example 2, determine how I_{AV} varies along the screen as a function of y .

Solution:

$$I_{AV} = 4E_0^2 \cos^2\left(\frac{\pi a}{\lambda s} y\right), \text{ where } a = 2 \times 10^{-4} \text{ m, } \lambda = 550 \times 10^{-9} \text{ m, and } s = 1.2 \text{ m}$$

$$I_{AV} = 4E_0^2 \cos^2 \left(\frac{\pi(2 \times 10^{-4})y}{550 \times 10^{-9}(1.2)} \right)$$

$$I_{AV} = 4E_0^2 \cos^2(303 \pi y)$$

Note that, when $y = \frac{1}{303}, \frac{2}{303}, \frac{3}{303}$ and so on, the angle $(303 \pi y)$ becomes π rad, 2π rad, 3π rad, and so on, for which $\cos^2(303 \pi y)$ is always 1. At these values of y , we have the first order, second order and third order of bright fringes—each of intensity $I_{AV} = 4E_0^2$. Since the interval Δy between successive fringes is $\frac{1}{303}$ meter, we get $\Delta y = 3.3 \times 10^{-3}$ m or 3.3 mm, in agreement with the value of Δy given in Example 2.

C. Thin-film interference

Interference effects provide us with the rainbow of colors we often see on thin-film soap bubbles and “oil slicks.” Each is an example of the interference of white light reflecting from opposite surfaces of the thin film. When thin films of different refractive indexes and thicknesses are judiciously stacked, coatings can be created that either enhance reflection greatly (HR coats) or suppress reflection (AR coats). A basic appreciation of these phenomena begins with an understanding of interference in a *single thin film*.

1. Single-film interference. The geometry for thin-film interference is shown in Figure 4-10. We assume that the light strikes the film—of thickness t and refractive index n_f —at near-perpendicular incidence. In addition we take into account the following established facts:

- A light wave traveling from a medium of lower refractive index to a medium of higher refractive index *automatically undergoes a phase change of π (180°)* upon reflection. A light wave traveling from a medium of higher index to one of lower index *undergoes no phase change* upon reflection. (We state this without proof.)
- The wavelength of light λ_n in a medium of refractive index n is given by $\lambda_n = \lambda_0/n$, where λ_0 is the wavelength in a vacuum or, approximately, in air.

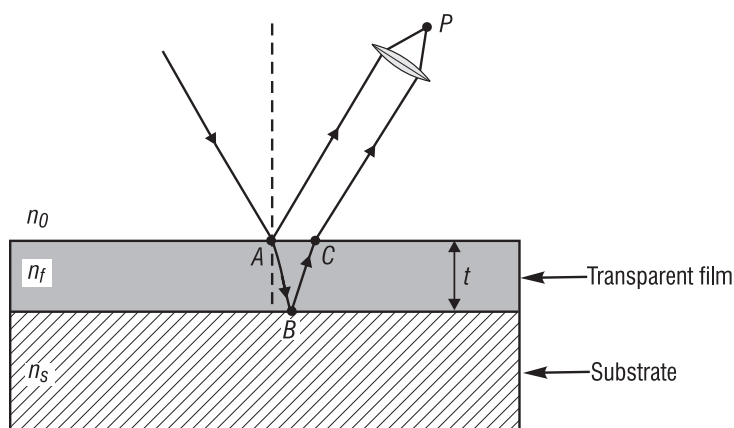


Figure 4-10 Two-beam interference from a thin film. Rays reflected from the film's top and bottom plane surfaces are brought together at P by a lens.

In Figure 4-10, we show a light beam in medium of index n_0 incident on the transparent film of index n_f . The film itself rests on a substrate of index n_s . Generally, the initial medium is air, so that $n_0 = 1$. The beam incident on the film surface at A divides into reflected and refracted portions. The refracted beam reflects again at the film-substrate interface at B and leaves the film at C, in the same direction as the beam reflected at A. Part of the beam may reflect internally again at C and continue to experience multiple reflections within the film layer until it has lost its intensity. There will thus exist multiple parallel beams emerging from the top surface, although with rapidly diminishing amplitudes.

Unless the reflectance of the film is large, a *good approximation* to the more complex situation of multiple reflection is to consider only *the first two emerging beams*. The two parallel beams leaving the film at A and C can be brought together by a converging lens, the eye, for example. The two beams intersecting at P overlap and interfere. Since the two beams travel different paths from point A onward, one in air, the other partly in the film, a relative phase difference develops that can produce *constructive* or *destructive* interference at P. The optical path difference Δ —in the case of *normal incidence*—is the additional path length ABC traveled by the refracted ray. The *optical path difference* in the film is equal to the product of the geometrical path difference $(AB + BC)$ times the refractive index of the film. If the incident ray is nearly perpendicular to the surface, the path difference $(AB + BC)$ is approximately equal to twice the film thickness $2t$. Then,

$$\Delta = n(AB + BC) = n(2t) \quad (4-15)$$

where t is the film thickness. For example, if $2nt = \lambda_0$, the wavelength of the light in air, the two interfering beams—on the basis of optical path difference alone—would be in phase and produce constructive interference.

However, an additional phase difference, due to the phenomenon mentioned above—*phase change on reflection*—must be considered. Suppose that $n_f > n_0$ and $n_f > n_s$. Often, in practice, $n_0 = n_s$, because the two media bounding the film are identical, as in the case of a water film (soap bubble) in air. Then the reflection at A occurs with light going from a lower index n_0 (air) toward the higher index n_f (film). The reflection at B, on the other hand, occurs for light going from a higher index n_f (film) toward a lower index n_s (air). Thus, the light reflecting at A shifts

phase by 180° (equivalent to one-half wavelength) while the light reflecting at B does not. As a result, if $2nt = \lambda_0$ and we add to this the additional $\lambda_0/2$ phase shift for the beam reflecting at A , we have a total optical path difference of $(\lambda_0 + \lambda_0/2)$, leading to *destructive*—rather than constructive—interference. So, in addition to the phase change introduced by path differences, we must always consider the *possible phase change upon reflection* at the interfaces.

If we denote Δ_p as the optical path difference due to the film and Δ_r as the equivalent path difference introduced upon reflection, the condition for *constructive interference* becomes

$$\Delta_p + \Delta_r = m\lambda, \quad (m = 1, 2, 3, \dots) \quad (4-17)$$

where m equals the order of interference.

For a thin film of thickness t and refractive index n_f , located in air, $\Delta_p = 2n_ft$ (according to Equation 4-15), and $\Delta_r = \lambda_0/2$. Thus, Equation 4-17—for *constructive interference*—becomes

$$\text{normal incidence: } 2n_ft + \frac{\lambda_0}{2} = m\lambda_0, \quad (m = 1, 2, 3, \dots) \quad (4-18)$$

where λ_0 is the wavelength in air. For *destructive interference*, Equation 4-18 changes slightly to

$$\text{normal incidence: } 2n_ft + \frac{\lambda_0}{2} = (m + 1/2)\lambda_0, \quad (m = 1, 2, 3, \dots) \quad (4-19)$$

Let's apply these ideas to the results of interference seen in soap-bubble films.

Example 4

White light is incident normally on the surface of a soap bubble. A portion of the surface reflects green light of wavelength $\lambda_0 = 540$ nm. Assume that the refractive index of the soap film is near that of water, so that $n_f = 1.33$. Estimate the thickness (in nanometers) of the soap bubble surface that appears green in second order.

Solution: Since the soap-bubble film is surrounded by air, Equation 4-18 applies. Rearranging Equation 4-18 to solve for the thickness t gives

$$t = \frac{\left(m\lambda_0 - \frac{\lambda_0}{2}\right)}{2n_f}$$

where $m = 2$, $n_f = 1.33$, and $\lambda_0 = 540$ nm. Thus,

$$t = \frac{\frac{3}{2}\lambda_0}{2n_f} = \frac{1.5(540 \text{ nm})}{2(1.33)} \cong 305 \text{ nm}$$

The soap film thickness is about 0.3 thousandths of a millimeter.

2. Single-layer antireflection (AR) coat. A common use of single-layer films deposited on glass substrates occurs in the production of *antireflecting (AR) coatings* on optical surfaces, often found in lenses for cameras and binoculars. The arrangement of a single-layer AR coat is shown in Figure 4-11, with the film made of magnesium fluoride (MgF_2) coated on top of a glass substrate.

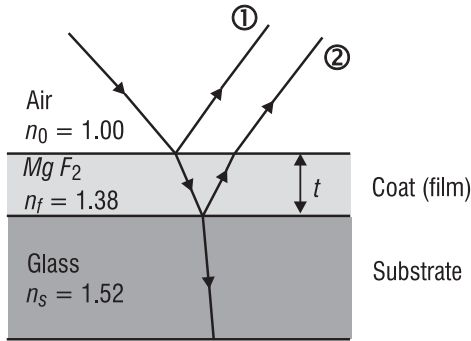


Figure 4-11 Single-layer AR coat on glass substrate

According to the rules for phase change upon reflection, *both* rays 1 and 2 undergo 180° shifts equal to $\lambda_0/2$, since both reflections occur at interfaces separating lower-to-higher refractive indexes. So the difference in phase between rays 1 and 2 comes from only the optical path difference due to the coating thickness t . If the thickness t is such that ray 2 falls behind ray 1 by $\lambda_{coat}/2$, the two rays interfere destructively, minimizing the reflected light. At *near-normal incidence* this requires that the distance $2t$, down and back, equal $\lambda_{coat}/2$. The mathematical condition for antireflection is then given by $2t = \frac{\lambda_{coat}}{2}$, and, since $\lambda_{coat} = \frac{\lambda_{air}}{n_f}$, we have finally

$$t = \frac{\lambda_{air}}{4n_f} \quad (4-20)$$

Example 5

Determine the minimum thickness of an AR coat of magnesium fluoride, MgF_2 , deposited on a glass substrate ($n_s = 1.52$) if the coating is to be highly antireflective for the center of the white light spectrum, say at $\lambda_{air} = 550$ nm. The refractive index for MgF_2 is near 1.38.

Solution: Application of Equation 4-20 gives

$$t_{min} = \frac{\lambda_{air}}{4n_f} = \frac{550 \text{ nm}}{4(1.38)}$$

$$t_{min} = 99.6 \text{ nm, about } 100 \text{ nm}$$

Without a coating (bare lens surface) the amount of light reflected is around 30% of the incident light. With a single-layer AR coat of 100 nm of MgF_2 on the lens surface, the light reflected drops to around 10%. Thus, the transmission of light through the lens increases from 70% to 90%.

3. Interference with multilayer films. As an extension of single-layer interference, consider the multilayer stack shown in Figure 4-12.

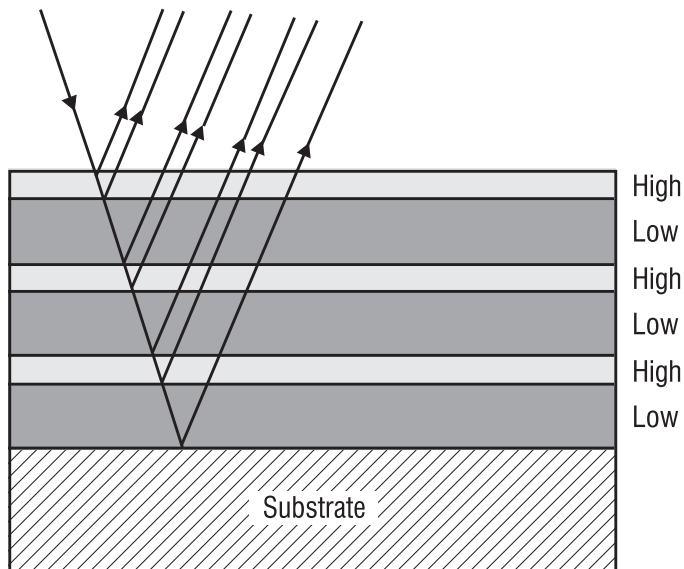


Figure 4-12 Multilayer stack of quarter-wave thin films of alternating high and low refractive indexes. Each film has an optical thickness of $\lambda_f/4$.

The stack is composed of alternate layers of identical *high* index and *low* index films. If each film has an optical thickness of $\lambda_f/4$, a little analysis shows that all emerging beams are in phase. Multiple reflections in the region of λ_0 increase the total reflected intensity, and the quarter-wave stack performs as an efficient mirror. Such multilayer stacks can be designed to satisfy extinction of reflected light—AR effect—or enhancement of reflected light—HR effect—over a greater portion of the spectrum than with a single-layer film. Such multilayer stacks are used in the design of *narrow-band interference filters* that filter out unwanted light, transmitting only light of the desired wavelength. For antireflection over *broadband* regions, the optical industry produces *HEBBAR™ coatings* (High Efficiency Broadband Anti Reflection) for regions of ultraviolet and infrared light, as well as for visible light. The coating industry also produces *V-coatings*, which reduce reflectance to near zero at one specific wavelength for an optical component. *High-reflection* coatings are produced over broadbands with multilayer stacks of thin films—just as for the antireflection coatings. In addition *HR* coats are used as overcoatings on *metallic reflectors*, which typically use aluminum, silver, and gold as the base metals. The overcoats protect the metals from oxidation and scratching.

IV. DIFFRACTION

The ability of light to bend around corners, a consequence of the wave nature of light, is fundamental to both interference and diffraction. *Diffraction* is simply any deviation from geometrical optics resulting from the *obstruction* of a wave front of light by some obstacle or some opening. Diffraction occurs when light waves pass through small openings, around obstacles, or by sharp edges.

Several common diffraction patterns—as sketched by an artist—are shown in Figure 4-13. Figure 4-13a is a typical diffraction pattern for HeNe laser light passing through a circular pinhole. Figure 4-13b is a typical diffraction pattern for HeNe laser light passing through a narrow (vertical) slit. And Figure 4-13c is a typical pattern for diffraction by a sharp edge.

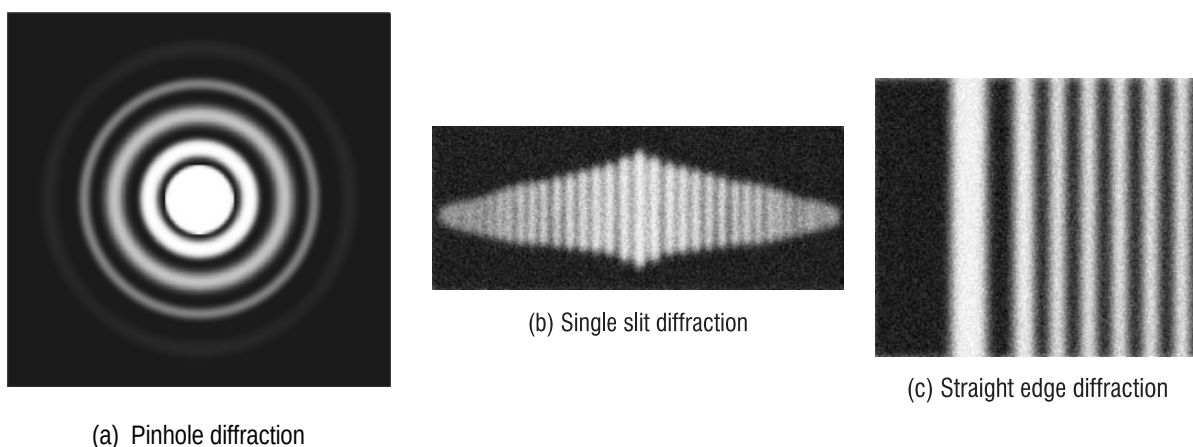


Figure 4-13 Sketches of several common diffraction patterns

The intricacy of the patterns should convince us—once and for all—that geometrical ray optics is incapable of dealing with diffraction phenomena. To demonstrate how wave theory does account for such patterns, we now examine the phenomenon of diffraction of waves by a single slit.

A. Diffraction by a single slit

The overall geometry for diffraction by a single slit is shown in Figure 4-14. The slit opening, seen in cross section, is in fact a long, narrow slit, perpendicular to the page. The shaded “humps” shown along the screen give a rough idea of intensity variation in the pattern, and the sketch of bright and dark regions to the right of the screen simulates the actual fringe pattern seen on the screen. We observe a wide central bright fringe, bordered by narrower regions of dark and bright. The angle θ shown connects a point P on the screen to the center of the slit.

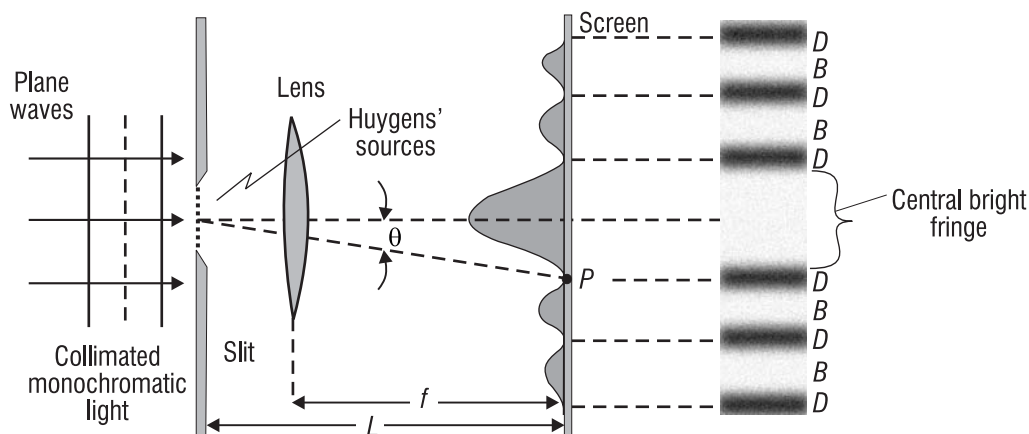


Figure 4-14 Diffraction pattern from a single slit

Since plane waves are incident on the screen, the diffraction pattern—in the absence of the focusing lens—would be formed far away from the slit and be much more spread out than that shown in Figure 4-14. The lens serves to focus the light passing through the slit onto the screen, just a focal length f away from the lens, while preserving faithfully the relative details of the diffraction pattern that would be formed on a distant screen without the lens.

To determine the location of the minima and maxima on the screen, we divide the slit opening through which a plane wave is passing into many point sources (Huygens' sources), as shown by the series of tiny dots in the slit opening of Figure 4-14. These numerous point sources send out Huygens' spherical waves, all in phase, toward the screen. There, at a point such as P , light waves from the various Huygens' sources overlap and interfere, forming the variation in light intensity shown in Figure 4-14. Thus, diffraction considers the contribution from *every part of the wave front* passing through the aperture. By contrast, when we looked at interference from Young's double slit, we considered *each slit* as a point source, ignoring details of the portions of the wave fronts in the slit openings themselves.

The mathematical details involved in adding the contributions at point P from each of the Huygens' sources can be found in basic texts on *physical optics*. Here we give only the end result of the calculation. Equation 4-21 locates the minima, y_{min} , on the screen, in terms of the slit width b , slit-to-screen distance L , wavelength λ , and order m .

$$y_{min} = \frac{m\lambda L}{b} \quad \text{where } m = 1, 2, 3, \dots \quad (4-21)$$

Figure 4-15 shows the positions of several orders of minima and the essential parameters associated with the single-slit diffraction pattern. (The positions of the *maxima* are mathematically more complicated to express, so we typically work with the positions of the well-defined minima.)

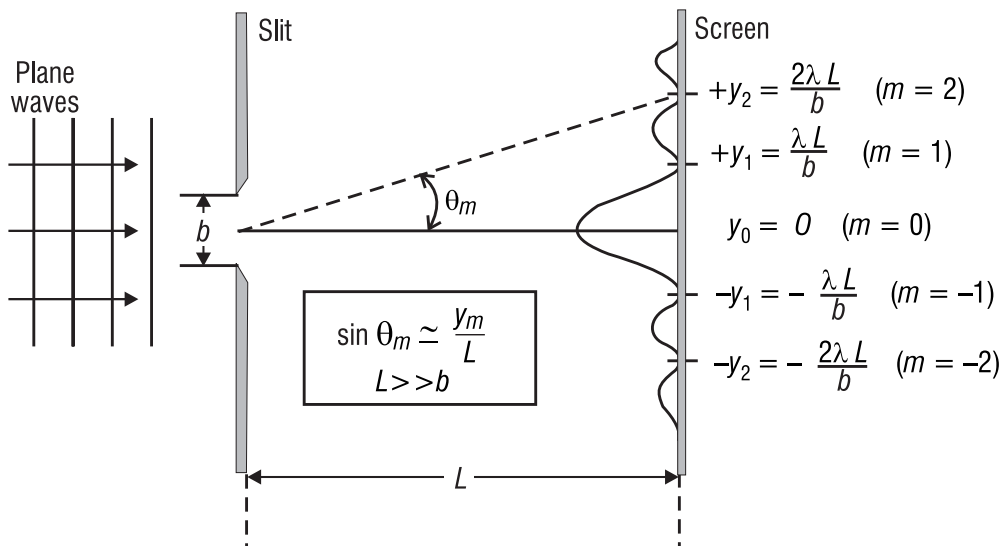


Figure 4-15 Positions of adjacent minima in the diffraction patterns (Drawing is not to scale.)

Now let's use Equation 4-21 to work several sample problems.

Example 6

Coherent laser light of wavelength 633 nm is incident on a single slit of width 0.25 mm. The observation screen is 2.0 m from the slit. (a) What is the width of the central bright fringe? (b) What is the width of the bright fringe between the 5th and 6th minima?

Solution:

(a) The width of the central bright fringe is $2y_1$, where y_1 is the distance to the first minimum ($m = 1$) on either side. Thus, using Equation 4-21,

$$\text{Width} = 2y_1 = 2\left(\frac{m\lambda L}{b}\right) = \frac{(2)(1)(633 \times 10^{-9} \text{ m})(2.0 \text{ m})}{2.5 \times 10^{-4} \text{ m}} = 0.01 \text{ m}$$

The width of the central bright fringe is about 1 cm.

$$(b) \text{ Width} = y_6 - y_5 = \frac{6\lambda L}{b} - \frac{5\lambda L}{b} = \frac{\lambda L}{b}$$

$$\text{Width} = \frac{(633 \times 10^{-9} \text{ m})(2.0 \text{ m})}{2.5 \times 10^{-4} \text{ m}} = 5.06 \times 10^{-3} \text{ m} \cong 0.5 \text{ cm}$$

The width of bright fringe between the 5th and 6th minima is about half the width of the central bright fringe.

Example 7

Monochromatic light is incident on a single slit of width 0.30 mm. On a screen located 2.0 m away, the width of the central bright fringe is measured and found to be near 7.8 mm. What is the wavelength of the incident light?

Solution: Since the width of the central bright fringe is 7.8 mm, equal to $2y_1$, we see that

$$y_1 = 3.9 \text{ mm. Then, rearranging Equation 4-21 to find } \lambda, \text{ we have } \lambda = \frac{y_{\min} b}{mL}, \text{ where}$$

$y_{\min} = y_1 = 3.9 \text{ mm}$, $m = 1$, $L = 2.0 \text{ m}$, and $b = 0.30 \text{ mm}$. Thus,

$$\lambda = \frac{(3.9 \times 10^{-3})(3 \times 10^{-4})}{(1)(2.0)} = 5.85 \times 10^{-7} \text{ m}$$

$\lambda \cong 585 \text{ nm}$, very near the principal wavelengths of light from sodium lamps.

B. Fraunhofer and Fresnel diffraction

In general, if the observation screen is far removed from the slit on which plane waves fall (as in Figure 4-15) or a lens is used to focus the collimated light passing through the slit onto the screen (as in Figure 4-14), the diffraction occurring is described as *Fraunhofer diffraction*, after Joseph von Fraunhofer (1787-1826), who first investigated and explained this type of so-called

far-field diffraction. If however, no lens is used and the observation screen is near to the slit, for either incident plane or spherical waves, the diffraction is called *Fresnel diffraction*, after Augustin Fresnel (1788-1829), who explained this type of *near-field* diffraction. The mathematical calculations required to determine the details of a diffraction pattern and account for the variations in intensity on the pattern are considerably more complicated for Fresnel diffraction than for Fraunhofer diffraction, so typically one studies first the Fraunhofer diffraction patterns, as we have.

Without going into the details of how to distinguish mathematically between Fresnel and Fraunhofer diffraction we can give results that help you decide whether the diffraction pattern formed is Fraunhofer or Fresnel in origin. Knowing this distinction helps you choose which equations to use in describing a particular diffraction pattern arising from a particular optical setup.

1. Criteria for far-field and near-field diffraction. Figure 4-16 shows the essential features of a general diffraction geometry, involving a source of light of wavelength λ , an opening to “obstruct” the light, and a screen to form the diffraction pattern.

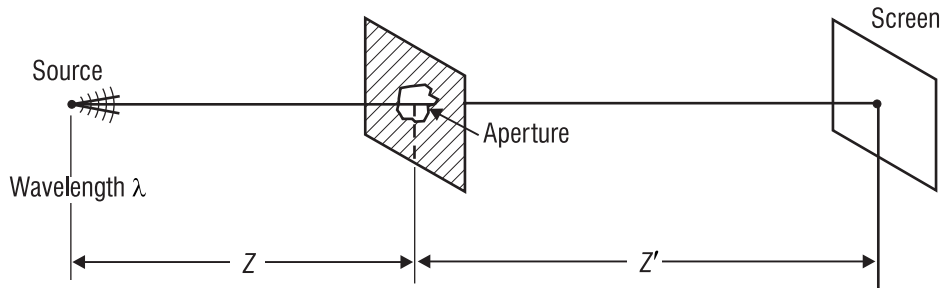


Figure 4-16 General diffraction geometry involving source, aperture, and screen

The distance from source to aperture is denoted as Z and that from aperture to screen as Z' . Calculations based on geometries that give rise to Fraunhofer and Fresnel diffraction patterns verify the following:

- If the distance Z from source to aperture **and** the distance Z' from aperture to screen are **both** greater than the ratio $\left(\frac{\text{aperture area}}{\lambda}\right)$ by a factor of 100 or so, the diffraction pattern on the screen is characteristic of *Fraunhofer diffraction*—and the screen is said to be in the *far field*. For this situation, all Fraunhofer-derived equations apply to the details of the diffraction pattern.
- If **either** distance— Z or Z' —is of the order of, or less than, the ratio $\left(\frac{\text{aperture area}}{\lambda}\right)$, the diffraction pattern on the screen is characteristic of *Fresnel diffraction* and is said to be in the *near field*. For this situation, all Fresnel-derived equations apply to the details of the diffraction pattern.
- Equation 4-22 indicates the “rule-of-thumb” conditions to be satisfied for both Z and Z' for *Fraunhofer diffraction*.

$$\text{Far - field condition: (Fraunhofer)} \quad \left\{ \begin{array}{l} Z > 100 \left[\frac{\text{aperture area}}{\lambda} \right] \\ Z' > 100 \left[\frac{\text{aperture area}}{\lambda} \right] \end{array} \right\} \quad (4-22)$$

Figure 4-18 illustrates these conditions and shows the locations of the *near field*, *far field*, and a *gray area* in between. If the screen is in the *gray area* and accuracy is important, a Fresnel analysis is usually applied. If the screen is in the *gray area* and approximate results are acceptable, a Fraunhofer analysis (significantly simpler than a Fresnel analysis) can be applied.

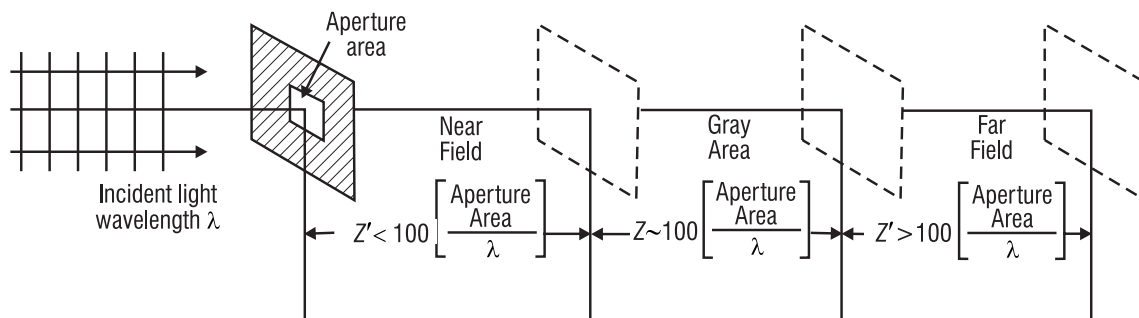


Figure 4-17 Defining near-field and far-field regions for diffraction

Figure 4-18 shows how we can satisfy the conditions for Fraunhofer diffraction, as spelled out in Equation 4-22, through the use of focusing lenses *on both sides* of the aperture (Figure 4-18a)—or with a *laser illuminating the aperture* and a *focusing lens* located on the screen side of the aperture (Figure 4-18b). Either optical arrangement has plane waves approaching and leaving the aperture, guaranteeing that the diffraction patterns formed are truly Fraunhofer in nature.

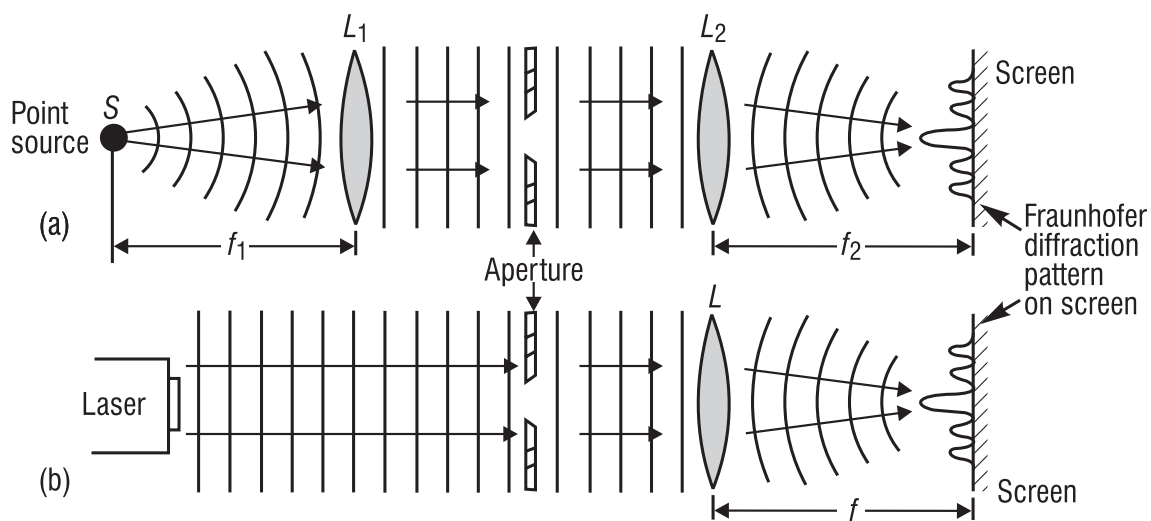


Figure 4-18 Optical arrangements for Fraunhofer diffraction

Now let's see how Equation 4-22 and Figure 4-18 are applied in a real situation.

Example 8

Minati, a photonics technician, has been asked to produce a Fraunhofer diffraction pattern formed when light from a HeNe laser ($\lambda = 633 \text{ nm}$) passes through a pinhole of $150\text{-}\mu\text{m}$ diameter. In order to set up the correct geometry for Fraunhofer diffraction, Minati needs to know (a) the distance Z from the laser to the pinhole and (b) the distance Z' from the pinhole to the screen.

Solution: Minati needs first to test the conditions given in Equation 4-22 so she calculates the ratio of $\left(\frac{\text{aperture area}}{\lambda}\right)$ assuming the pinhole to be circular.

$$\text{Ratio} = \frac{\text{aperture area}}{\lambda} = \frac{\pi D^2}{4\lambda} = \frac{(3.14)(150 \times 10^{-6})^2}{(4)(633 \times 10^{-9})}$$

$$\text{Ratio} = 0.0279 \text{ m}$$

(a) Minati knows that light from the HeNe laser is fairly well collimated, so that nearly plane waves are incident on the pinhole, as illustrated in Figure 4-18b. She knows that plane waves are those that come—or appear to come—from very distant sources. So she concludes that, with the laser, the distance Z is much greater than 100 (0.0279 m)—that is, greater than about 2.8 m —and so the “ Z -condition” for Fraunhofer diffraction is automatically satisfied.

(b) From her calculation of the ratio $\left(\frac{\text{aperture area}}{\lambda}\right)$ she knows also that the distance Z' must be greater than 2.8 m . So she can place the screen 3 meters or so from the aperture and form a Fraunhofer diffraction pattern—**OR** she can place a positive lens just beyond the aperture—as in Figure 4-18b—and focus the diffracting light on a screen a focal length away. With the focusing lens in place she obtains a much reduced—but valid—Fraunhofer diffraction pattern located nearer the aperture. She chooses to use the latter setup, with a positive lens of focal length 10 cm , enabling her to arrange the laser, pinhole, and screen, all on a convenient 2-meter optical bench.

2. Several typical Fraunhofer diffraction patterns. In successive order, we show the far-field diffraction pattern for a *single slit* (Figure 4-19), a *circular aperture* (Figure 4-20), and a *rectangular aperture* (Figure 4-21). Equations that describe the locations of the bright and dark fringes in the patterns accompany each figure.

Single Slit

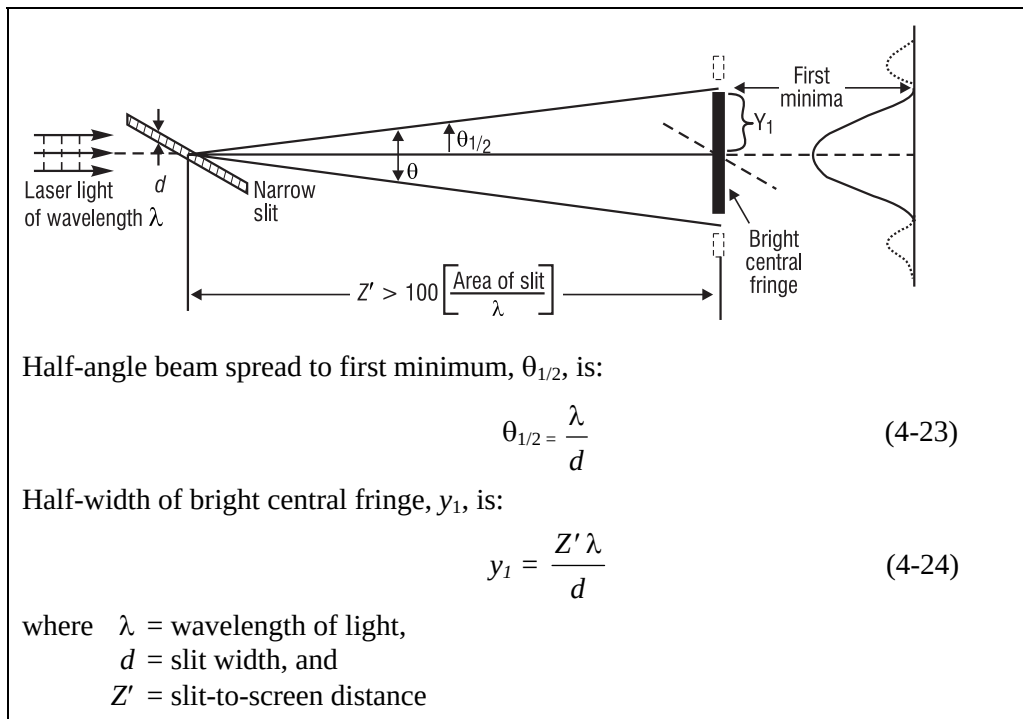


Figure 4-19 Fraunhofer diffraction pattern for a single slit

Circular Aperture

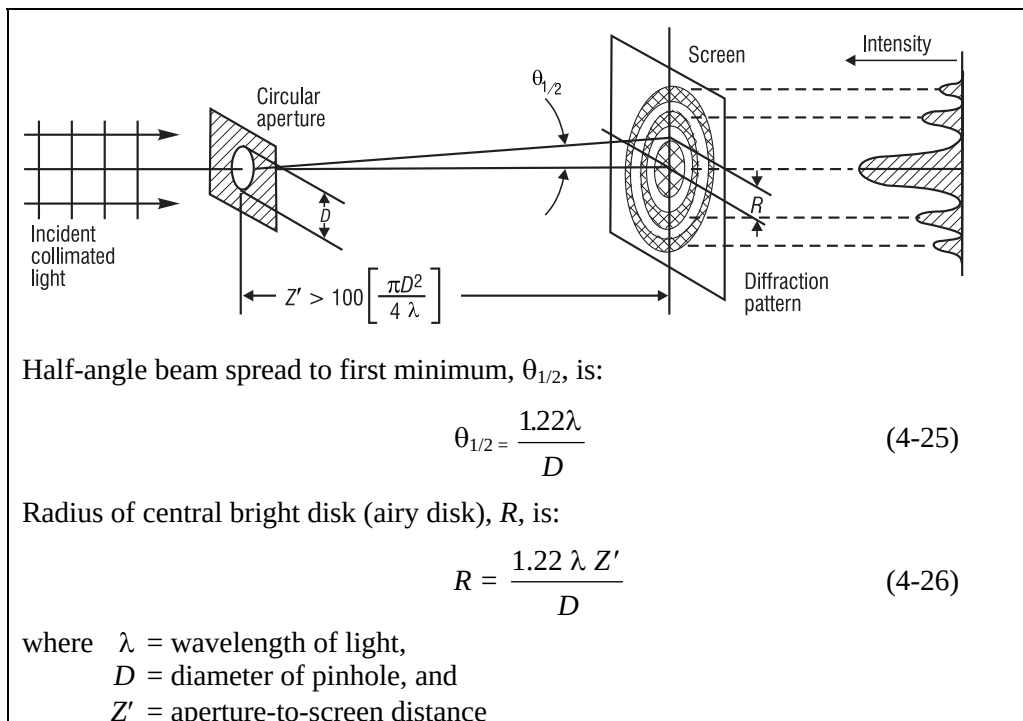


Figure 4-20 Fraunhofer diffraction pattern for a circular aperture

Rectangular aperture

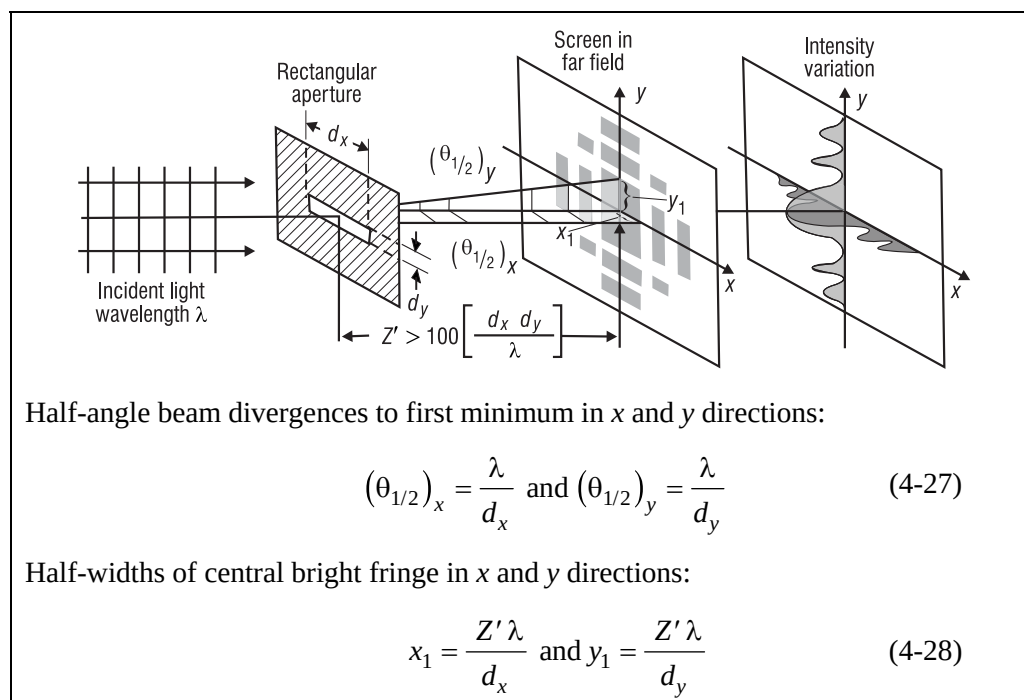


Figure 4-21 Fraunhofer diffraction pattern for a rectangular aperture

C. Diffraction Grating

If we prepare an aperture with thousands of adjacent slits, we have a so-called *transmission-diffraction grating*. The width of a single slit—the opening—is given by d , and the distance between slit centers is given by ℓ (see Figure 4-22). For clarity, only a few of the thousands of slits normally present in a grating are shown. Note that the spreading of light occurs always in a direction perpendicular to the direction of the long edge of the slit opening—that is, since the long edge of the slit opening is *vertical* in Figure 4-22, the spreading is in the *horizontal* direction—along the screen.

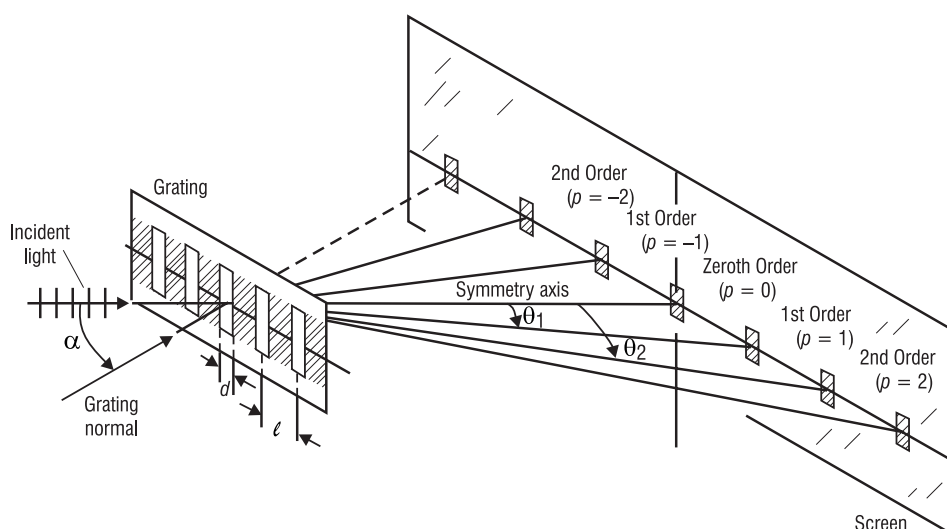


Figure 4-22 Diffraction of light through a grating under Fraunhofer conditions

The resulting diffraction pattern is a series of sharply defined, widely spaced fringes, as shown. The central fringe, on the symmetry axis, is called the *zeroth-order* fringe. The successive fringes on either side are called *1st order*, *2nd order*, etc., respectively. They are numbered according to their positions relative to the central fringe, as denoted by the letter p .

The intensity pattern on the screen is a *superposition* of the *diffraction effects* from each slit as well as the *interference effects* of the light from all the adjacent slits. The combined effect is to cause overall cancellation of light over most of the screen with marked enhancement over only limited regions, as shown in Figure 4-22. The location of the bright fringes is given by the following expression, called the *grating equation*, assuming that Fraunhofer conditions hold.

$$\ell (\sin \alpha + \sin \theta_p) = p\lambda \text{ where } p = 0, \pm 1, \pm 2, \dots \quad (4-29)$$

where ℓ = distance between slit centers

α = angle of incidence of light measured with respect to the normal to the grating surface

θ_p = angle locating the p th-order fringe

p = an integer taking on values of 0, ± 1 , ± 2 , etc.

λ = wavelength of light

Note that, if the light is incident on the grating along the grating normal ($\alpha = 0$), the grating equation, Equation 4-29, reduces to the more common form shown in Equation 4-30.

$$\ell (\sin \theta_p) = p\lambda \quad (4-30)$$

If, for example, you shine a HeNe laser beam perpendicularly onto the surface of a transmission grating, you will see a series of brilliant red dots, spread out as shown in Figure 4-22. A complete calculation would show that less light falls on each successively distant red dot or fringe, the $p = 0$ or central fringe being always the brightest. Nevertheless, the location of each

bright spot, or fringe, is given accurately by Equation 4-29 for either normal incidence ($\alpha = 0$) or oblique incidence ($\alpha \neq 0$). If light containing a mixture of wavelengths (white light, for example) is directed onto the transmission grating, Equation 4-29 holds for *each* component color or wavelength. So each color will be spread out on the screen according to Equation 4-29, with the longer wavelengths (red) spreading out farther than the shorter wavelengths (blue). In any case, the central fringe ($p = 0$) always remains the same color as the incident beam, since all wavelengths in the $p = 0$ fringe have $\theta_p = 0$, hence all overlap to re-form the “original” beam and therefore the original “color.” Example 9 shows calculations for a typical diffraction grating under Fraunhofer conditions.

Example 9

Michael has been handed a transmission grating by his supervisor who wants to know how widely the red light and blue light fringes—in *second* order—are separated on a screen one meter from the grating. Michael is told that the separation distance between the red and blue colors is a critical piece of information needed for an experiment with a grating spectrometer. The transmission grating is to be illuminated at *normal incidence* with red light at $\lambda = 632.8$ nm and blue light at $\lambda = 420.0$ nm. Printed on the frame surrounding the ruled grating, Michael sees that there are 5000 slits (lines) per centimeter on this grating. Michael decides he must, in turn:

- Determine the distance ℓ between the slit centers.
- Determine the angular deviation θ_p in 2nd order for both the red and the blue light.
- Determine the separation distance on the screen between the red and blue fringes.

Solution:

(a) Since there are 5000 slits or grooves per centimeter, Michael knows that the distance ℓ between the slits, center to center, must be $\ell = \frac{1 \text{ cm}}{5000} = 2 \times 10^{-4} \text{ cm}$.

(b) At normal incidence ($\alpha = 0$), Equation 4-29 reduces to Equation 4-30, so, for 2nd order ($p = 2$), Michael writes the following two equations and solves them for the deviation angles θ_2^{red} and θ_2^{blue} :

$$\begin{aligned}\sin \theta_2^{\text{red}} &= \frac{p\lambda_{\text{red}}}{\ell} = \frac{(2)(632.8 \times 10^{-9} \text{ m})}{2 \times 10^{-6} \text{ m}} = 0.6328 \\ \therefore \theta_2^{\text{red}} &= \sin^{-1}(0.6328) = 39.3^\circ \\ \sin \theta_2^{\text{blue}} &= \frac{p\lambda_{\text{blue}}}{\ell} = \frac{(2)(420 \times 10^{-9} \text{ m})}{2 \times 10^{-6} \text{ m}} = 0.4200 \\ \therefore \theta_2^{\text{blue}} &= \sin^{-1}(0.4200) = 24.8^\circ\end{aligned}$$

(c) From the geometry shown in Figure 4-22, Michael sees that the screen distances y_2^{red} and y_2^{blue} to the red and blue fringes in 2nd order respectively, and the grating-to-screen distance Z' are related to deviation angles by the equation

$$\tan \theta_2 = \frac{y_2}{Z'}, \text{ where here, } Z' = 1 \text{ meter.}$$

Thus

$$\Delta y = y_2^{red} - y_2^{blue} = (Z' \tan \theta_2^{red}) - (Z' \tan \theta_2^{blue})$$

which becomes

$$\Delta y = (1 \text{ m}) (\tan 39.3^\circ - \tan 24.8^\circ)$$

$$\Delta y = (100 \text{ cm}) (0.8185 - 0.4621)$$

$$\Delta y = 35.6 \text{ cm}$$

Michael reports his finding of $\Delta y = 35.6 \text{ cm}$ to his supervisor, who decides that this grating will work in the proposed experiment.

D. Diffraction-Limited Optics

A lens of diameter D is in effect a large circular aperture through which light passes. Suppose a lens is used to focus plane waves (light from a distant source) to form a “spot” in the focal plane of the lens, much as is done in *geometrical optics*. Is the focused spot truly a *point*? Reference to Figure 4-20 indicates that the focused spot is actually a tiny diffraction pattern—with a bright disk at the center (the so-called *airy disk*) surrounded by dark and bright rings, as pictured earlier in Figure 4-13a.

In Figure 4-23, we see collimated light incident on a lens of focal length f . The lens serves as both a circular aperture of diameter D to intercept the plane waves and a lens to focus the light on the screen, as shown in Figure 4-18b. Since the setup in Figure 4-23 matches the conditions shown in Figure 4-18b, we are assured that a Fraunhofer diffraction pattern will form at the “focal spot” of the lens.

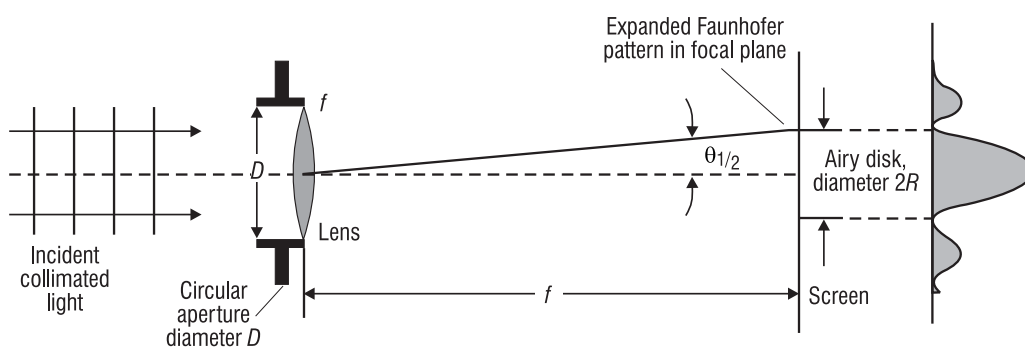


Figure 4-23 Fraunhofer diffraction pattern formed in the focal plane of a lens of focal length f (Drawing is not to scale.)

The diffraction pattern is, in truth then, an array of alternate bright and dark rings, with a bright spot at the center, even though the array is very small and hardly observable to the human eye. From the equations given with Figure 4-20, we see that the diameter of the central bright spot—inside the surrounding rings—is itself of size $2R$, where, from Equation 4-26,

$$2R = 2 \left(\frac{1.22 \lambda Z'}{D} \right) \quad (4-31)$$

where $Z' = f$

While indeed small, the diffraction pattern overall is greater than $2R$, demonstrating clearly that a lens focuses collimated light to a small diffraction pattern of rings and not to a point. However, when the lens is inches in size, we do justifiably refer to the focal plane pattern as a “point,” ignoring all structure within the “point.” Example 10 provides us with a “feel” for the *size of the structure* in the focused spot, when a lens of nominal size becomes the circular aperture that gives rise to the airy disk diffraction pattern.

Example 10

Determine the size of the airy disk at the center of the diffraction pattern formed by a lens such as that shown in Figure 4-23, if the lens is 4 cm in diameter and its focal length is 15 cm. Assume a wavelength of 550 nm incident on the lens.

Solution: Using Equation 4-31 with $Z' = f$, the diameter of the airy disk is

$$2R = \frac{2.44\lambda f}{D} = \frac{(2.44)(550 \times 10^{-9} \text{ m})(0.15 \text{ m})}{0.04 \text{ m}}$$

$$2R = 5.03 \times 10^{-6} \text{ m}$$

Thus, the central bright spot (airy disk) in the diffraction pattern is **only** 5 micrometers in diameter. So, even though the focused spot is not a true point, it is small enough to be considered so in the world of large lenses, i.e., in the world of geometrical optics.

The previous discussion and example indicate that the size of the focal spot—structure and all—is limited by diffraction. No matter what we do, we can never make the airy disk smaller than that given by $2R = \frac{2.44 f \lambda}{D}$. That is the limit *set by diffraction*. So all optical systems are *limited* by diffraction in their ability to form true point images of point objects. We recognize this when we speak of *diffraction-limited optics*. An ideal optical system therefore can do no better than that permitted by diffraction theory. In fact, a real optical system—which contains imperfections in the optical lenses, variations in the index of refraction of optical components, scattering centers, and the existence of temperature gradients in the intervening atmosphere—will not achieve the quality limit permitted by diffraction theory. Real optical systems are therefore poorer than those limited by diffraction only. We often refer to real systems as *many-times diffraction limited* and sometimes attach a numerical figure such as “five-times diffraction-limited” to indicate the deviation in quality expected from the given system compared with an ideal “diffraction-limited” system.

V. POLARIZATION

We continue our discussion of the main concepts in *physical optics* with a brief look at *polarization*. Before we describe the polarization of light waves, let’s take a look at a simplistic—but helpful—analogy of “polarization” with *rope waves*.

A. Polarization—a simple analogy

Imagine a “magic” rope that you can whip up and down at one end, thereby sending a *transverse* “whipped pulse” (vibration) out along the rope. See Figure 4-24a. Imagine further that you can change the direction of the “whipped shape,” quickly and randomly at your end, so that a person looking back along the rope toward you, sees the “vibration” occurring in all directions—up and down, left to right, northeast to southwest, and so on, as shown in Figure 4-24b.

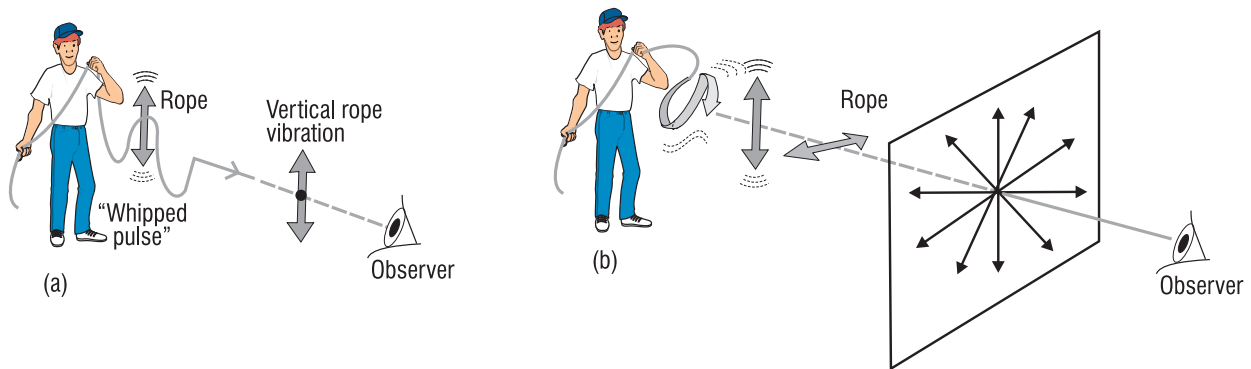


Figure 4-24 Rope waves and polarization

In Figure 4-24a, the rope wave is *linearly polarized*, that is, the rope vibrates in only one transverse direction—vertically in the sketch shown. In Figure 4-24b, the rope vibrations are in all transverse directions, so that the rope waves are said to be *unpolarized*.

Now imagine that the waves on the rope—representing all possible directions of vibration as shown in Figure 4-24b—are passed through a *picket fence*. Since the vertical slots of the fence pass only vertical vibrations, the many randomly oriented transverse vibrations incident on the picket fence emerge as only vertical vibrations, as depicted in Figure 4-25. In this example of transverse waves moving out along a rope, we see how we can—with the help of a polarizing device, the picket fence in this case—change unpolarized rope waves into polarized rope waves.

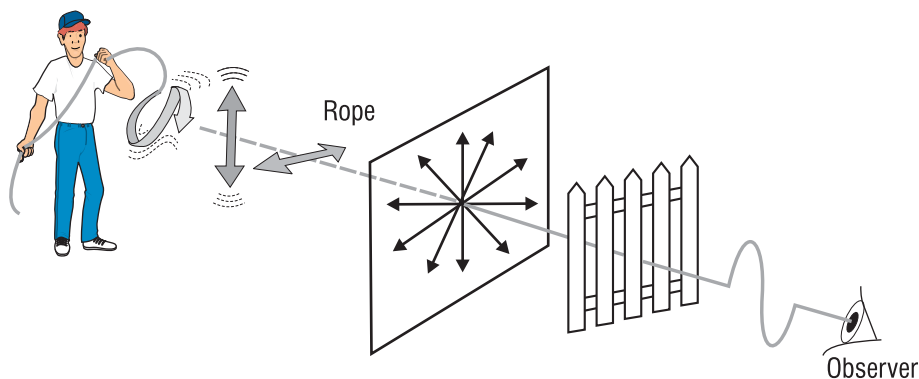


Figure 4-25 Polarization of rope waves by a picket fence

B. Polarization of light waves

The polarization of light waves refers to the *transverse* direction of vibration of the electric field vector of electromagnetic waves. (Refer back to Figure 4-3.) As described earlier, *transverse* means E-field vibrations *perpendicular* to the direction of wave propagation. If the electric field vector remains in a given direction in the transverse x-y plane—as shown in Figure 4-26—the light is said to be *linearly polarized*. (The “vibration” of the electric field referred to here is not the same as a *physical* displacement or movement in a rope. Rather, the vibration here refers to an increase and decrease of the electric field strength occurring in a particular transverse direction—at all given points along the propagation of the wave.) Figure 4-26 shows linearly polarized light propagating along the z-direction toward an observer at the left. The electric field E increases and decreases in strength, reversing itself as shown, always along a direction making an angle θ with the y-axis in the transverse plane. The E-field components $E_x = E \sin \theta$ and $E_y = E \cos \theta$ are shown also in the figure.

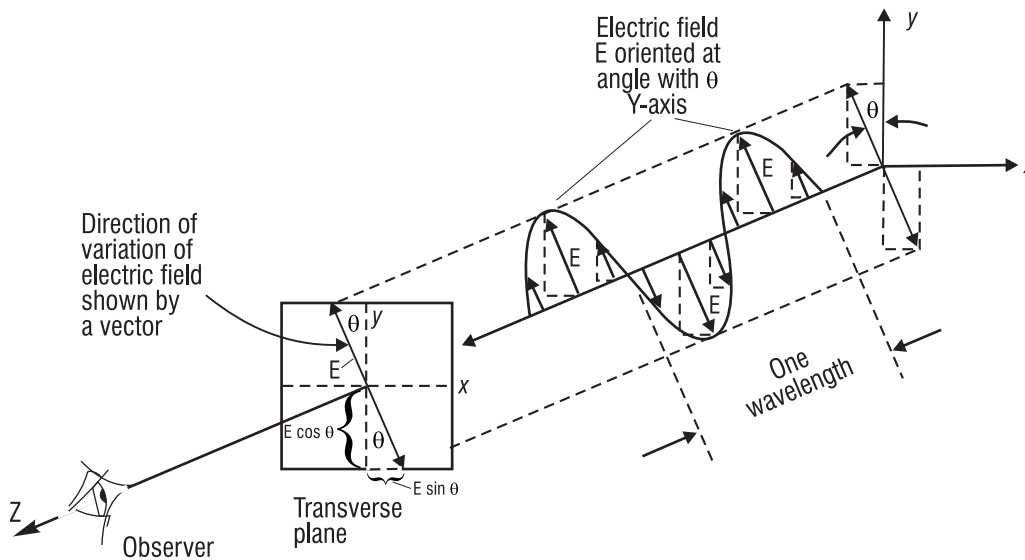
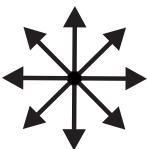
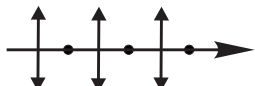
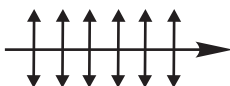


Figure 4-26 Linearly polarized light with transverse electric field E propagating along the z-axis

Table 1 lists the *symbols* used generally to indicate *unpolarized* light (E -vector vibrating randomly in all directions), *vertically polarized* light (E -vector vibrating in the vertical direction only), and *horizontally polarized* light (E -vector vibrating in the horizontal direction only). With reference to Figure 4-26, the vertical direction is along the y-axis, the horizontal direction along the x-axis.

Table 4-1 Standard Symbols for Polarized Light

Viewing Position	Unpolarized	Vertically Polarized	Horizontally Polarized
Viewed head-on; beam coming toward viewer			
Viewed from the side; beam moving from left to right			

Like the action of the picket fence described in Figure 4-25, a special optical filter—called either a *polarizer* or an *analyzer* depending on how it's used—transmits only the light wave vibrations of the E-vector that are lined up with the filter's *transmission axis*—like the slats in the picket fence. The combined action of a polarizer and an analyzer are shown in Figure 4-27. Unpolarized light, represented by the multiple arrows, is incident on a “polarizer” whose transmission axis (TA) is vertical. As a result, only *vertically polarized* light emerges from the polarizer. The vertically polarized light is then incident on an “analyzer” whose transmission axis is horizontal, at 90° to the direction of the vertically polarized light. As a result, **no** light is transmitted.

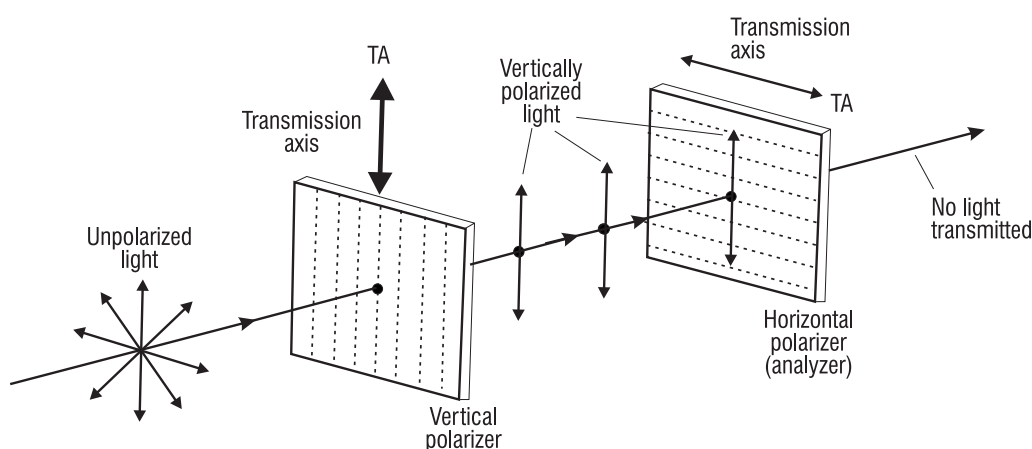


Figure 4-27 Effect of polarizers on unpolarized light

C. Law of Malus

When unpolarized light passes through a polarizer, the light intensity—proportional to the square of its electric field strength—is reduced, since only the E-field component along the transmission axis of the polarizer is passed. When linearly polarized light is directed through a polarizer and the direction of the E-field is at an angle θ to the transmission axis of the polarizer, the light intensity is likewise reduced. The reduction in intensity is expressed by the *law of Malus*, given in Equation 4-32.

$$I = I_0 \cos^2 \theta \quad (4-32)$$

where I = intensity of light that is passed through the polarizer

I_0 = intensity of light that is incident on the polarizer

θ = angle between the transmission axis of the polarizer and the direction of the E-field vibration

Application of the law of Malus is illustrated in Figure 4-28, where two polarizers are used to control the intensity of the transmitted light. The first polarizer changes the incident unpolarized light to linearly polarized light, represented by the vertical vector labeled E_0 . The second polarizer, whose TA is at an angle θ with E_0 , passes only the component $E_0 \cos \theta$, that is, the part of E_0 that lies along the direction of the transmission axis. Since the intensity goes as the square of the electric field, we see that I , the light intensity transmitted through polarizer 2, is equal to $(E_0 \cos \theta)^2$, or $I = E_0^2 \cos^2 \theta$. Since E_0^2 is equal to I_0 , we have demonstrated how the *law of Malus* ($I = I_0 \cos^2 \theta$) comes about.

We can see that, by rotating polarizer 2 to change θ , we can vary the amount of light passed. Thus, if $\theta = 90^\circ$ (TA of polarizer 1 is 90° to TA of polarizer 2) no light is passed, since $\cos 90^\circ = 0$. If $\theta = 0^\circ$ (TA of polarizer 1 is parallel to TA of polarizer 2) all of the light is passed, since $\cos 0^\circ = 1$. For any other θ between 0° and 90° , an amount $I_0 \cos^2 \theta$ is passed.

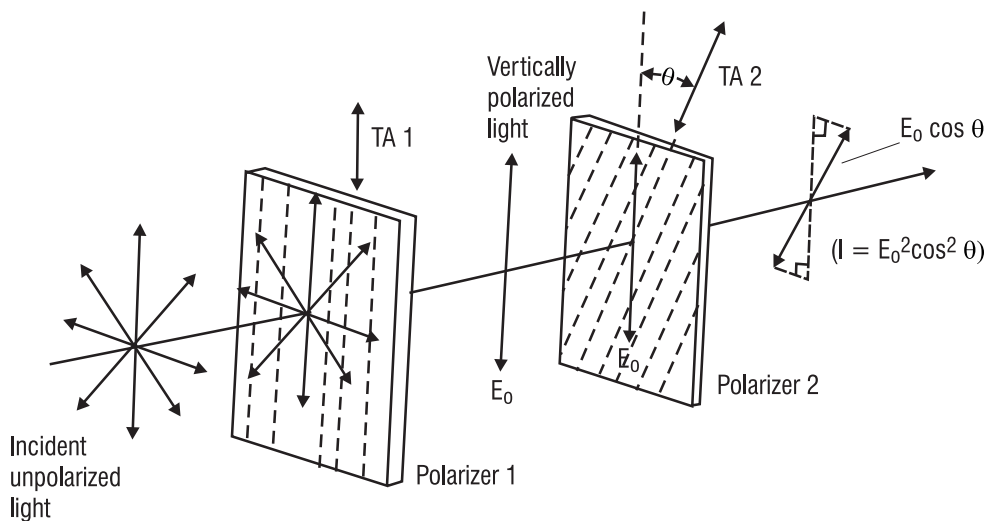


Figure 4-28 Controlling light intensity with a pair of polarizers

Example 11 shows how to use the law of Malus in a light-controlling experiment.

Example 11

Unpolarized light is incident on a pair of polarizers as shown in Figure 4-28.

- (a) Determine the angle θ required—between the transmission axes of polarizers 1 and 2—that will reduce the intensity of light I_0 incident on polarizer 2 by 50%.
- (b) For this same reduction, determine by how much the field E_0 incident on polarizer 2 has been reduced.

Solution:

- (a) Based on the statement of the problem, we see that $I = 0.5 I_0$. By applying the *law of Malus*, we have:

$$I = I_0 \cos^2 \theta$$

$$0.5 I_0 = I_0 \cos^2 \theta$$

$$\cos \theta = \sqrt{0.5} = 0.707$$

$$\theta = 45^\circ$$

So the two TAs should be at an angle of 45° with each other.

- (b) Knowing that the E-field passed by polarizer 2 is equal to $E_0 \cos \theta$, we have

$$E_2 = E_0 \cos \theta$$

$$E_2 = E_0 \cos 45^\circ$$

$$E_2 = 0.707 E_0 \cong 71\% E_0$$

Thus, the E-field incident on polarizer 2 has been reduced by about 29% after passing through polarizer 2.

D. Polarization by reflection and Brewster's angle

Unpolarized light—the light we normally see around us—can be polarized through several methods. The polarizers and analyzers we have introduced above polarize by *selective absorption*. That is, we can prepare materials—called *dichroic* polarizers—that selectively *absorb* components of E-field vibrations along a given direction and largely *transmit* the components of the E-field vibration perpendicular to the absorption direction. The perpendicular (transmitting) direction defines the TA of the material. This phenomena of selective absorption is what E. H. Land discovered in 1938 when he produced such a material—and called it *Polaroid*.

Polarization is produced also by the phenomenon of *scattering*. If light is incident on a collection of particles, as in a gas, the electrons in the particles absorb and reradiate the light. The light radiated in a direction perpendicular to the direction of propagation is partially polarized. For example, if you look into the north sky at dusk through a polarizer, the light being scattered toward the south—toward you—is partially polarized. You will see variations in the intensity of the light as you rotate the polarizer, confirming the state of partial polarization of the light coming toward you.

Another method of producing polarized light is by *reflection*. Figure 4-29 shows the *complete* polarization of the *reflected light* at a *particular angle of incidence* B , called the *Brewster angle*.

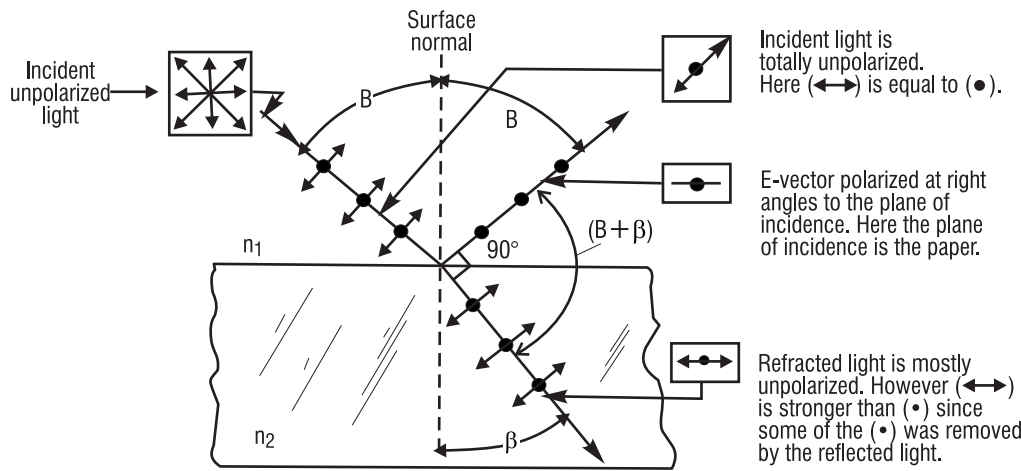


Figure 4-29 Polarization by reflection at Brewster's angle

The refracted light on the other hand becomes only partially polarized. Note that the symbols introduced in Table 4-1 are used to keep track of the different components of polarization. One of these is the dot ($\bullet$) which indicates E-field vibrations perpendicular to both the light ray and the plane of incidence, that is, in and out of the paper. The other is an arrow ($\leftrightarrow$) indicating E-field vibrations **in** the plane of incidence **and perpendicular** to the ray of light. The reflected E-field coming off at Brewster's angle is totally polarized in a direction in and out of the paper, perpendicular to the reflected ray. This happens only at Brewster's angle, that particular angle of incidence for which the angle between the reflected and refracted rays, $B + \beta$, is exactly 90° . At the angle of incidence B , the E-field component ($\leftrightarrow$) cannot exist, for if it did it would be **along** the reflected ray, violating the requirement that E-field vibrations must always be transverse—that is, perpendicular to the direction of propagation. Thus, only the E-field component perpendicular to the plane of incidence ($\bullet$) is reflected.

Referring to Figure 4-29 and Snell's law at the Brewster angle of incidence, we can write:

$$n_1 \sin B = n_2 \sin \beta$$

Since $\beta + B = 90^\circ$, $\beta = 90 - B$, which then allows us to write

$$n_1 \sin B = n_2 \sin (90 - B) = n_2 \cos B$$

or

$$\frac{\sin B}{\cos B} = \frac{n_2}{n_1}$$

and finally

$$\boxed{\tan B = \frac{n_2}{n_1}} \quad (4-33)$$

Equation 4-33 is an expression for Brewster's law. Knowing n_1 (the refractive index of the *incident* medium) and n_2 (the refractive index of the *refractive* medium), we can calculate the Brewster angle B . Shining light on a reflecting surface at this angle ensures complete polarization of the reflected ray. We make use of Equation 4-33 in Example 12.

Example 12

In one instance, unpolarized light in air is to be reflected off a glass surface ($n = 1.5$). In another instance, internal unpolarized light in a glass prism is to be reflected at the glass-air interface, where n for the prism is also 1.5. Determine the Brewster angle for each instance.

Solution:

- (a) Light going from air to glass. In this case, $n_1 = 1$ and $n_2 = 1.5$.

Using Equation 4-33

$$\tan B = \frac{n_2}{n_1} = \frac{1.5}{1}$$

$$B = \tan^{-1} 1.5 = 56.3^\circ$$

The Brewster angle is 56.3° .

- (b) Light going from glass to air: In this case, $n_1 = 1.5$ and $n_2 = 1.0$.

Then,
$$\tan B = \frac{n_2}{n_1} = \frac{1}{1.5} = 0.667$$

$$B = \tan^{-1} (0.667) = 33.7^\circ$$

The Brewster angle is 33.7° .

E. Brewster windows in a laser cavity

Brewster windows are used in laser cavities to ensure that the laser light—after bouncing back and forth between the cavity mirrors—emerges as linearly polarized light. Figure 4-30 shows the general arrangement of the windows—thin slabs of glass with parallel sides—mounted on the opposite edges of the gas laser tube—in this case a helium-neon gas laser.

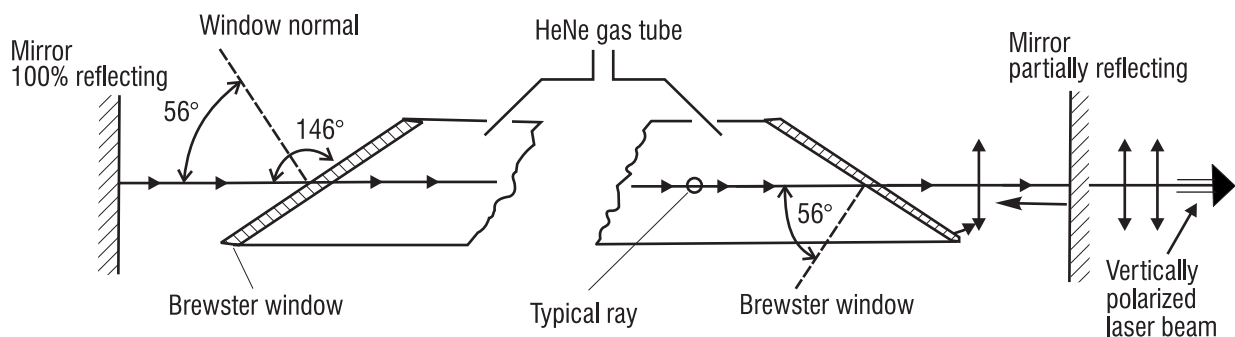


Figure 4-30 Brewster windows in a HeNe gas laser

As you can see, the light emerging is linearly polarized in a vertical direction. Why this is so is shown in detail in Figure 4-31. Based on Figure 4-29 and Example 12, Figure 4-31 shows that it is the refracted light—and not the reflected light—that is eventually linearly polarized.

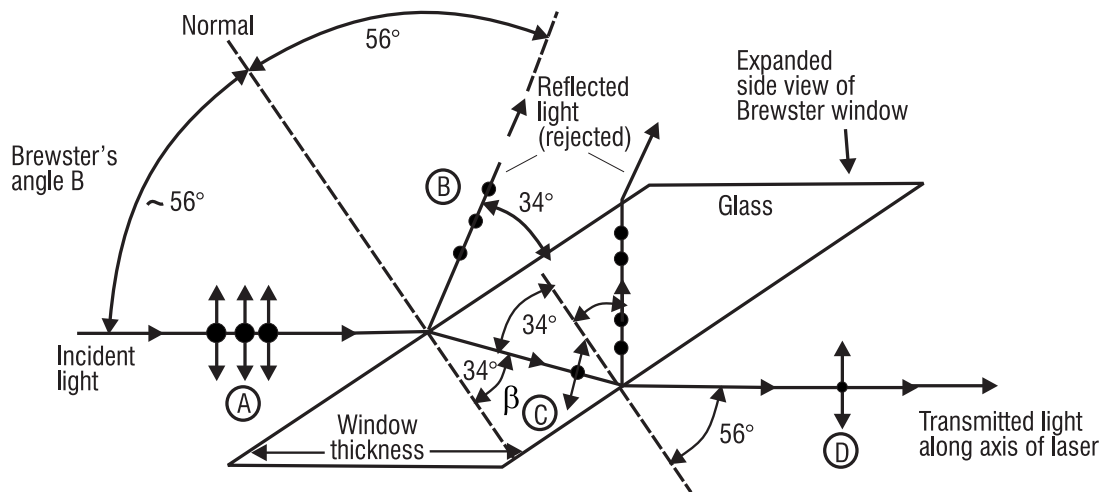


Figure 4-31 Unpolarized light passing through both faces at a Brewster angle

The unpolarized light at A is incident on the left face of the window—from air to glass—defining, as in Example 12, a Brewster angle of 56.3° . The reflected light at B is totally polarized and is *rejected*. The refracted (transmitted) light at C is now partially polarized since the reflected light has carried away part of the vibration perpendicular to the paper (shown by the dots). At the right face, the ray is incident again at a Brewster angle (34°) for a glass-to-air interface—as was shown in Example 12. Here again, the reflected light, totally polarized, is *rejected*. The light transmitted through the window, shown at D, now has even less of the vibration perpendicular to the paper. After hundreds of such passes back and forth through the Brewster windows, as the laser light bounces between the cavity mirrors, the transmitted light is left with only the vertical polarization, as shown exiting the laser in Figure 4-30. And since all of the reflected light is removed (50% of the initial incident light) we see that 50% of the initial incident light *remains* in the refracted light, hence in the laser beam.

Laboratory

In this laboratory you will complete the following experiments:

- Carry out a quantitative mapping of the intensity variation across a Fraunhofer airy-diffraction pattern.
- Determine the wavelength of light by using a machinist's rule as a reflection grating.
- Convert unpolarized light to polarized light by reflection at Brewster's angle.

Equipment List

The following equipment is needed to complete this laboratory.

- 1 HeNe laser (unpolarized, TEM₀₀ output, 1–3-mW range, 632.8 nm)
- 1 diode laser pointer (5 mW or less)
- 1 precision pinhole aperture (150- μ m diameter)
- 1 photomultiplier with fiber optic probe
- 1 linear translator capable of transverse motion in 0.1-mm increments
- 2 optical benches, calibrated, 2 meters long
- 3 bench mounts with vertical rods
- 2 laboratory jacks
- 1 neutral-density filter
- 1 632.8-nm filter
- 2 H-type Polaroid sheet mounts with TAs identified
- 1 diffuser (ground glass plate)
- 1 reflecting glass plate (microscope slide)
- 1 machinist's rule, marked off in 64ths of an inch

Procedure

A. *Quantitative mapping of airy diffraction pattern*

1. Set up the equipment as shown in Figure L-1. With the help of Equation 4-22, determine the pinhole-to-screen distance Z' to ensure the formation of a *Fraunhofer* airy-diffraction pattern of the 150- μ m hole at the "screen" (see Example 8). Set the pinhole-to-screen distance accordingly.

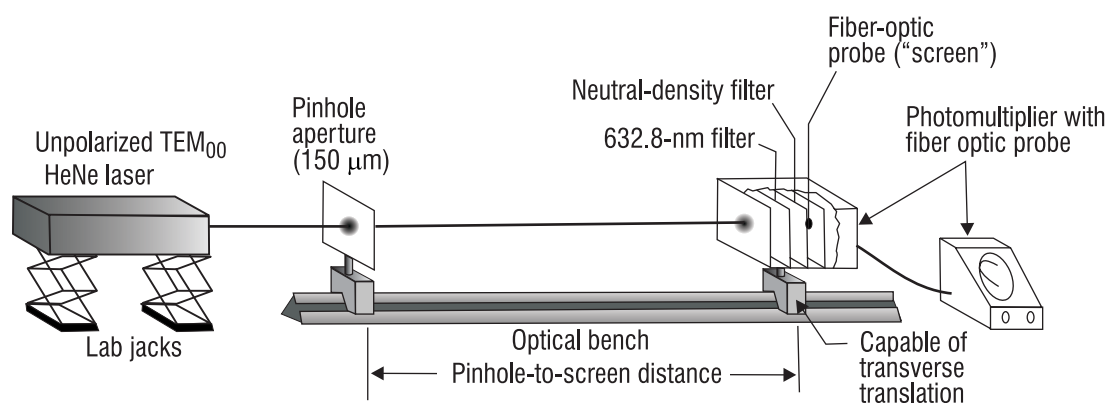


Figure L-1 Arrangement of apparatus for recording intensity distribution of Fraunhofer diffraction pattern from a circular pinhole

2. With room lights off, align the laser, 150-μ pinhole, and the tip of the fiber optic probe so that the laser beam becomes the axis of symmetry for each component. Use a 5" × 8" index card to observe the airy pattern in front of the fiber optic probe (the *virtual* location of the "screen"), ensuring that clear, sharp airy disk and set of concentric rings are formed. Adjust the positions of the laser pinhole and fiber optic tip relative to one another to obtain a maximum intensity reading at the center of the airy disk. (Be patient!) Note that a 632.8-nm filter is added near the fiber optic tip to let you work with room lights on. The neutral-density filter shown in Figure L₁ may be used—if necessary—as an additional intensity control, to permit scanning the entire airy pattern without a scale change on the photometer.
3. After the laser beam, 150-μm pinhole, and fiber optic tip have been carefully aligned and the 632.8-nm filter is in place, turn on the lights and take intensity readings. With the horizontal translator, move the fiber optic tip assembly back and forth, transversely across the optical bench (and the Fraunhofer diffraction pattern) several times. Watch the photometer to ensure that the alternate maxima and minima of intensity in the airy pattern are being detected.
4. During the trial runs, choose sensitivity and scale factor settings on the photometer so that the highest readings remain on scale while the lowest readings are still clearly recorded. When you get satisfactory variations in the photometer readings as the fiber optic tip is scanned across the airy pattern, you can begin to record readings of intensity versus position. Try to obtain, at the very least, intensity variations across the center disk and two of the adjoining rings. (The pinhole-to-screen distance may have to be reduced to around 100 cm to ensure that the translator scan encompasses the desired extent of the airy pattern. In that event, the pattern may be in the *near field* rather than the *far field* and the equation given in Figure 4-20 may not hold exactly—but it will be close enough.) Readings can be taken every 0.5 mm or so, beginning with the second ring, moving on through the central disk and on to the second ring on the opposite side. Record the photometer readings versus position and plot them on suitable graph paper.
5. Compare the intensity distribution with that shown qualitatively in Figure 4-20. Since the pinhole diameter, wavelength, and pinhole-to-screen distance are all known for the plot,

measure the radius of the central airy disk on the plot and compare this result with that predicted by Equation 4-26.

B. Determine the wavelength of light with a reflection grating

1. To perform this experiment you need only a diode laser pointer, a mount for the laser that allows it to tilt downward, a solid table on which to position the laser mount and the machinist's rule, and a wall (screen), five to fifteen feet from the rule. Figure L-2 shows the general setup. Choose an appropriate angle α to form a clear diffraction pattern on the wall, locating several orders $y_1, y_2 \dots y_p$, as shown in Figure L-2.

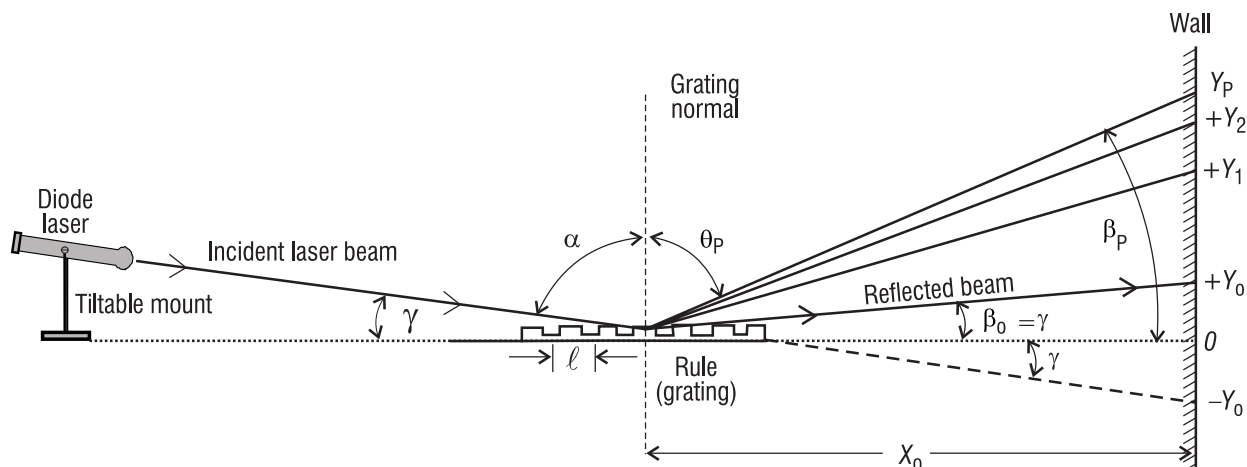


Figure L-2 Using the grooves on a machinist's rule as a reflection grating

2. In Figure L-2, the symbols shown are:

γ = slant angle laser beam makes with the grating (rule) surface

α = angle of incidence of laser beam with grating normal

θ_p = the direction angle to the p th diffraction order, measured relative to the normal

$\beta_0 = \gamma$ = angle of laser beam reflected from rule, relative to the surface

β_p = diffraction angle to the p th diffraction order, measured relative to the surface

ℓ = "grating" spacing between adjacent grooves on the rule

x_0 = distance from center of rule to the wall (or screen)

Locate on the wall the reflected beam (at $+y_0$) and the diffraction orders $y_1, y_2, y_3 \dots y_p$. The point $(-y_0)$ locates the spot formed by the laser beam if it had gone through the rule directly onto the wall. The point $+y_0$ locates the point of specular reflection of the laser beam off of the rule surface. The O position is the halfway point between $+y_0$ and $-y_0$.

3. If you begin with Equation 4-29 and adjust for sign conventions (since α and θ_p are on opposite sides of the normal for a reflection grating) you obtain the modified equation,

$$p\lambda = \ell (\sin \theta_p - \sin \alpha)$$

From the geometry in Figure L-2 and a series of substitutions and approximations for $\sin \theta_p$ and $\sin \alpha$, you arrive eventually at a useful working equation for λ that involves only ℓ , p , y_p , y_0 , and x_0 , each directly measurable, as seen in Figure L-2. This equation is

$$\lambda = \frac{\ell}{2p} \left[\frac{y_p^2 - y_0^2}{x_0^2} \right]$$

4. Obtain values for several measures of y_p and use the above equation for each measure to determine the wavelength λ of the diode laser. Take the average for your best value of λ . Knowing the true wavelength, determine how close your measured value comes. Express the deviation as a percent.

C. Conversion of unpolarized light to linearly polarized light

1. Using light from an unpolarized HeNe or diode laser, arrange your system as shown in Figure L-3. The incident unpolarized light passes through a diffuser—such as a ground glass plate—and on toward the reflecting surface (microscope slide). The light reflects off the glass surface and then passes through an analyzer on its way toward the observer. When the reflecting glass surface is rotated around a vertical axis so that the angle of incidence is equal to Brewster's angle—about 56° —the reflected light is found to be totally polarized with the E -vector perpendicular to the plane of incidence. (Recall that the plane of incidence is the plane that contains the incident ray and the normal to the reflecting surface. In Figure L-3, therefore, the plane of incidence is horizontal—parallel to the tabletop.

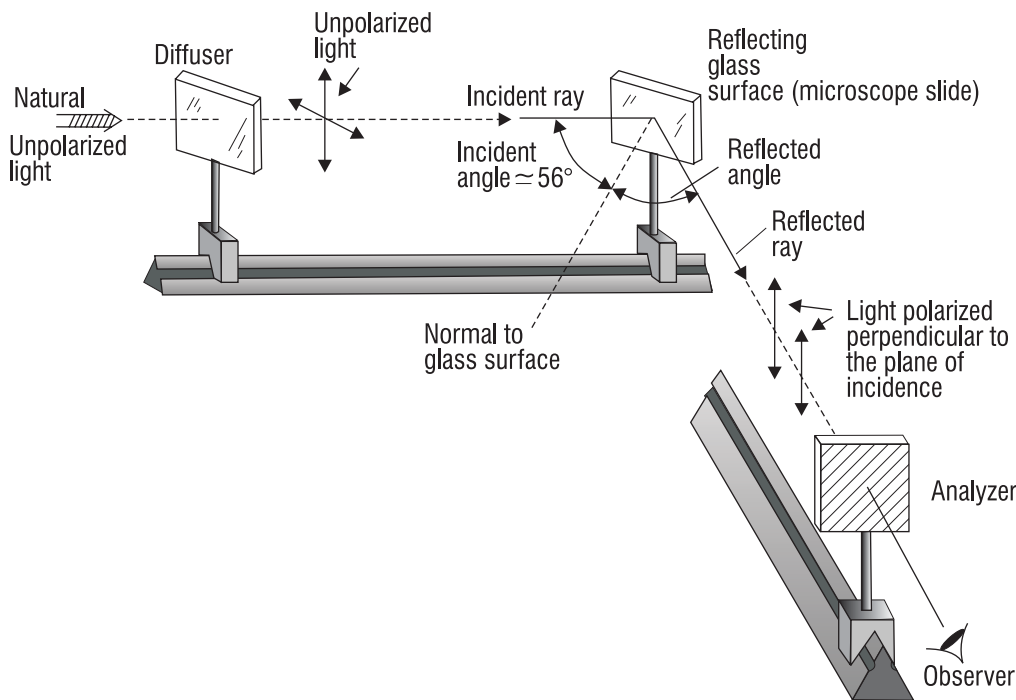


Figure L-3 Polarization by reflection at Brewster's angle

2. With the appropriate analyzer, whose transmission axis (TA) is known, verify that the light reflecting from the surface of the glass microscope slide is indeed vertically polarized, as indicated in Figure L-3. Explain your method of verification.

Other Resources

- The Education Council of the Optical Society of America (OSA) has prepared a *discovery kit* designed to introduce students to modern optical science and engineering. The kit includes two thin lenses, a Fresnel lens, a mirror, a hologram, an optical illusion slide, a diffraction grating, one meter of optical fiber, two polarizers, four color filters, and instructions for eleven detailed experiments. OSA offers teacher membership opportunities. Contact the Optical Society of America, 2010 Massachusetts Avenue, NW, Washington, D.C. 20036, 800-762-6960.
- Atneosen, R., and R. Feinberg. "Learning Optics with Optical Design Software," *American Journal of Physics*, Vol 59, March 1991: pp 242-47.
- "Teaching Optics With an O/H Projector," Douglas S. Goodman, Polaroid Corporation, 38 Henry Street, Cambridge, Massachusetts.
- K-12 Optics Outreach kit—available from SPIE, Bellingham, Washington.

References

Textbooks

- Ball, C. J. *An Introduction to the Theory of Diffraction*. New York, Pergamon Press, 1971.
- Beiser, Arthur. *Physics*, 3rd Edition, Menlo Park, California: The Benjamin/Cummings Publishing Company, 1982.
- Heaven, O. S. *Thin Film Physics*, New York, Barnes and Noble, 1970.
- Hecht, E., and A. Zajac. *Optics*, 2nd Edition, Reading, Massachusetts: Addison Wesley Publishing Company, 1987.
- Pedrotti, F., and L. Pedrotti. *Introduction to Optics*, 2nd Edition. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1993.
- Pedrotti, F., and L. Pedrotti. *Optics and Vision*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1998.
- Serway, R. A. *Principles of Physics*. Orlando, Florida: Saunders College Publishing, 1992.
- Shurcliff, N. A., and S. S. Ballard. *Polarized Light*. Princeton, New Jersey: D. Von Nostrand, 1964.
- Waldman, Gary. *Introduction to Light*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1983.

Articles

- Englert, B-G., M. O. Scully, and H. Walther. "The Duality in Matter and Light," *Scientific American* (December 1994): 86.

Mandoli, D. F., and W. R. Briggs. "Fiber Optics in Plants," *Scientific American* (August 1984): 90.

Wehner, R. "Polarized Light Navigation by Insects," *Scientific American* (July 1976): 106.

Weisskopf, Victor F. "How Light Interacts with Matter," *Lasers and Light*, Readings from *Scientific American*. W. H. Freeman and Company, 1969, pp 14-26.

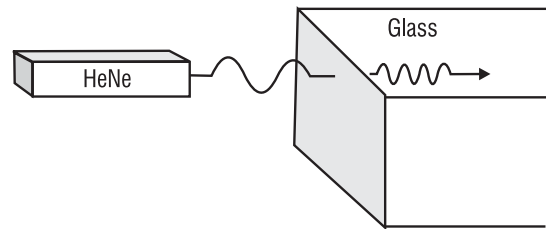
Optical Manufacturing Companies

Optics and Optical Instruments Catalog. Edmund Scientific, Industrial Optics Division, Barrington, New Jersey.

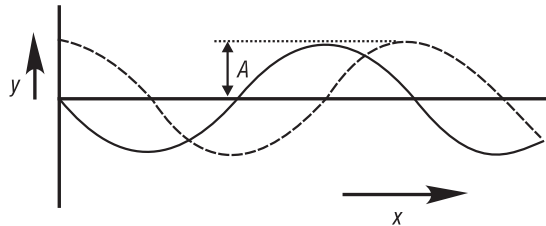
Melles-Griot Catalog. Melles-Griot Optical Systems, Rochester, New York, and Melles-Griot Photonics Components, Irvine, California.

Problem Exercises

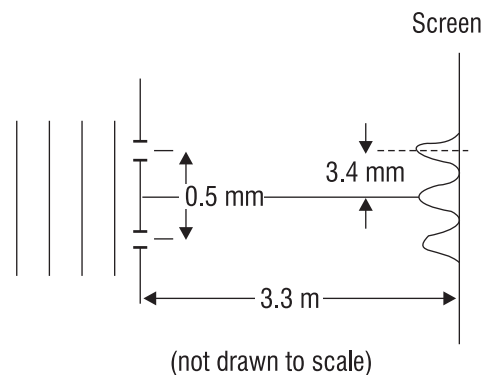
1. A HeNe laser in air emits light at 632.8 nanometers. The beam passes into a glass substance of index 1.35. What are the speed, wavelength, and frequency of the laser light in the glass substance?



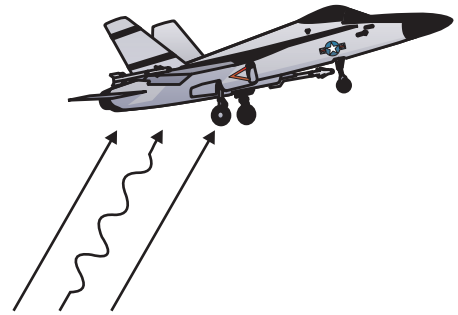
2. Use the principle of superposition to sketch a picture of the resultant waveform for the two waves shown at the right. Does the resultant wave ever reach an amplitude of $2A$?



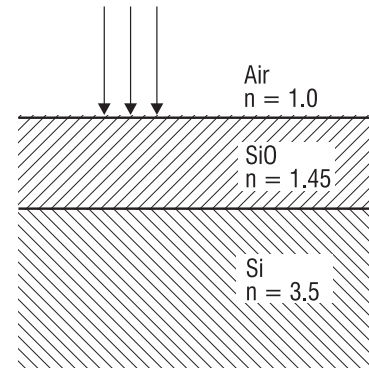
3. A Young's double-slit interference experiment is carried out with blue-green argon laser light. The interference pattern formed on a screen 3.3 m away contains the first-order ($m = 1$) bright fringe 3.4 mm from the center of the pattern. If the separation of the double slits is 0.50 mm, what is the wavelength of the argon laser light?



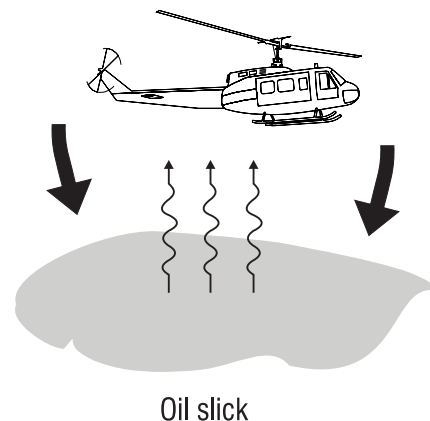
4. Suppose you are asked to design a nonreflecting surface for a Stealth aircraft. Your goal is to select an antireflective polymer of index 1.6 and of optimum thickness, so that radar waves of wavelength $\lambda = 3.5$ cm will not be reflected from the aircraft surface. What is the thinnest layer of polymer you can apply to achieve your goal?



5. Solar cells made of silicon (Si) are designed with nonreflecting, thin-film coatings such as silicon monoxide (SiO) to minimize the reflection of incident light. Determine the thickness of the thinnest film of SiO that will cause the least reflection of sunlight. Take the average wavelength of sunlight to be near 550 nanometers.

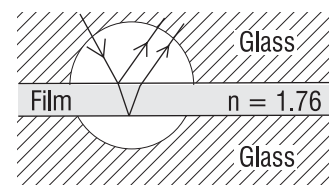


6. An oil spill on an ocean coastline near you produces an oil slick on the water for miles around. After a few days, you take a helicopter ride out over the water and notice—with the help of a handheld spectrometer—that the oil slick produces a first-order maximum of reflected light of wavelength 550 nanometers. What is the thickness of the oil slick at that time? Assume the oil has $n = 1.25$ and saltwater has $n = 1.34$.

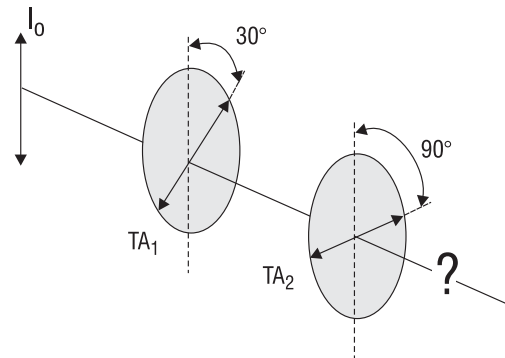


7. A laser beam of unknown wavelength is incident on a single slit of width 0.25 mm and forms a Fraunhofer diffraction pattern on a screen 2.0 m away. The width of the central bright fringe is measured to be about 7 mm. What is the wavelength of the laser light? What might the laser be?

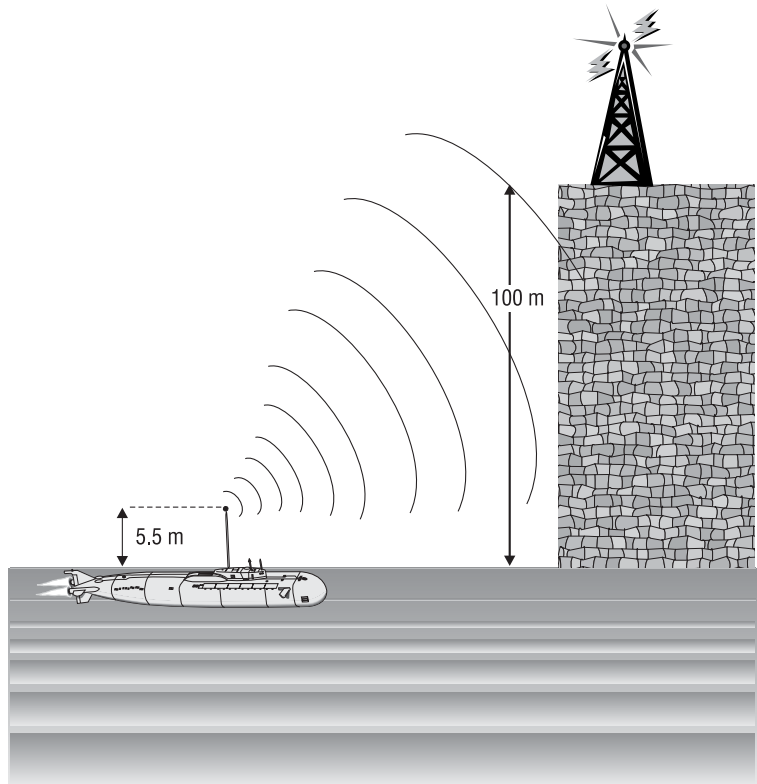
8. A thin layer of liquid (methylene iodide) is sandwiched between two glass microscope slides, as shown in the accompanying sketch. Light of wavelength $\lambda = 550$ nm is incident on the glass-liquid interface. The liquid has an index of refraction of $n = 1.76$. (a) Is there an automatic phase shift of $\lambda/2$ for the light reflected at the top face of the liquid film? (b) Is there an automatic phase shift of $\lambda/2$ for the light reflected at the bottom of the film? (c) What is the minimum thickness of liquid film required if the light incident *perpendicularly* on the sandwich is to be strongly reflected?



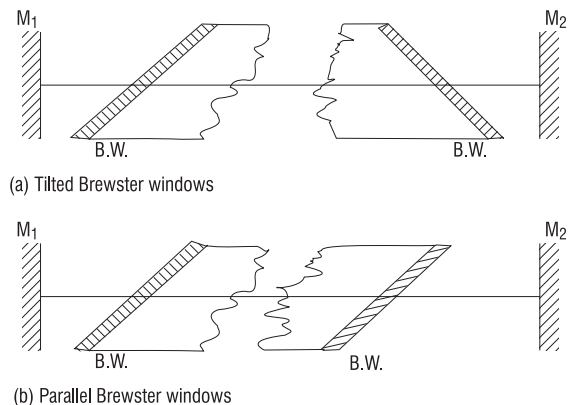
9. Nearly plane waves of CO_2 laser light of wavelength $\lambda = 10.6 \mu\text{m}$ emerge from a circular aperture 10 cm in diameter. If the CO_2 laser light is to be examined on a target in the *far field*, about how far from the aperture should this examination take place?
10. Refer to Figure 4-20, which shows a Fraunhofer diffraction pattern for a circular aperture. If the aperture is of diameter $150 \mu\text{m}$ and a helium-cadmium laser of wavelength $\lambda = 442 \text{ nm}$ is used to illuminate the aperture, determine (a) an approximate far-field distance from aperture to screen for a Fraunhofer pattern, (b) the half-angle beam divergence of the laser beam in the far field, and (c) the radius of the airy disk on the screen.
11. If one were to send a laser beam of wavelength 694 nm through a telescope aperture of diameter 2.5 meters on toward the moon, $3.84 \times 10^5 \text{ km}$ away, what would be the diameter of the laser beam when it hit the moon's surface?
12. What is the angular separation in second order between light of wavelength 400 nm and 700 nm when the light is incident normally (perpendicularly) on a diffraction grating of 5000 grooves/cm ?
13. Vertically polarized light of intensity I_0 is incident on a polarizer whose TA makes an angle of 30° with the vertical. The light then passes through a second polarizer whose TA makes an angle of 90° with the vertical, as shown in the sketch. (a) What is the intensity of the light, in terms of I_0 , that passes through the second polarizer? (b) What is its orientation?



14. A submarine floating on the ocean ($n = 1.34$) transmits a message to a receiver on a 100-m high cliff located at the coastline. The submarine's antenna extends 5.5 m above the ocean surface. (a) If the transmitted signal is completely polarized by reflection from the ocean surface, how far must the submarine's antenna be from the coastline? (b) Relative to the ocean surface, what is the direction of polarization of the polarized signal?



15. Figure 4-30 shows the two Brewster windows tilted toward one another, each at the appropriate angle with the horizontal. See sketch (a). What would be gained or lost by having the windows parallel to each other, as in sketch (b)? Hint: Use your knowledge of Snell's law to draw a center ray completely through both windows from M_1 to M_2 . Use Figure 4-31 to help with details at each window. What happens to the center ray for each sketch?



Lasers

William T. Silfvast
School of Optics/CREOL
University of Central Florida
Orlando, Florida

Lasers are devices that amplify or increase the intensity of light to produce a highly directional, high-intensity beam that typically has a very pure frequency or wavelength. They come in sizes ranging from approximately one-tenth the diameter of a human hair to that of a very large building. Lasers produce powers ranging from nanowatts to a billion trillion watts (10^{21} W) for very short bursts. They produce wavelengths or frequencies ranging from the microwave region and infrared to the visible, ultraviolet, vacuum ultraviolet, and into the soft-X-ray spectral regions. They generate the shortest bursts of light that man has yet produced, or approximately five million-billionths of a second (5×10^{-15} sec).

Lasers are a primary component of some of our most modern communication systems and are the probes that generate the audio signals from our compact disk players. They are used for cutting, heat treating, cleaning, and removing materials in both the industrial and medical worlds. They are the targeting element of laser-guided bombs and are the optical source in both supermarket checkout scanners and tools (steppers) that print our microchips.

Because of the special stimulated nature of the laser light source, and the apparatus needed to produce laser light, laser photons are generally not as cheap to produce or to operate as are other light sources of comparable power. We presently do not use them to light our rooms, as lamp bulbs for our flashlights, as headlights for our automobiles, or as street lamps. Lasers also don't generally provide "white light" but instead produce a specific "color" or wavelength, depending upon the laser used.

The word **LASER** is an acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation. Stimulated emission of radiation is a natural process first identified by Einstein. It occurs when a beam of light passes through a specially prepared medium and initiates or stimulates the atoms within that medium to emit light in exactly the same direction and exactly at the same wavelength as that of the original beam. A typical laser device (Figure 5-1) consists of an amplifying or gain medium, a pumping source to input energy into the device, and an optical cavity or mirror arrangement that reflects the beam of light back and forth through the gain

medium for further amplification. A useful laser beam is obtained by allowing a small portion of the light to escape by passing through one of the mirrors that is partially transmitting.

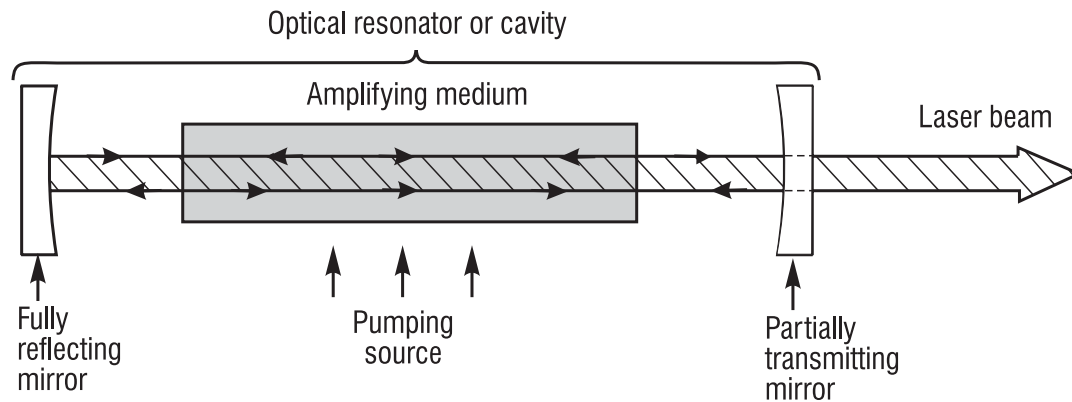


Figure 5-1 Basic laser components including gain medium, pumping source, and mirror cavity

Prerequisites

Before you begin working with this module, you should have completed Modules 1-1, *Nature and Properties of Light*; 1-2, *Light Sources and Laser Safety*; 1-3, *Basic Geometrical Optics*; and 1-4, *Basic Physical Optics*. In addition you will need a working knowledge of algebra, exponents, and logarithms.

Objectives

When you finish this module you will:

- understand how lasers operate
- understand how gain or amplification is produced
- know how various beam characteristics occur
- know about longitudinal and transverse modes
- design laser amplifiers
- design laser cavities or resonators
- understand unstable resonators
- be familiar with Q-switching
- understand mode locking
- be familiar with how a variety of laser types work and be familiar with their wavelengths, power capabilities, and beam properties
- know about the laser's unique properties (different from other light sources), which are essential in a variety of applications

Scenarios

Three types of job functions involving lasers are those in laser manufacturing relating to designing, assembling, and testing of lasers; those relating to using lasers in various types of applications; and those associated with field servicing of lasers.

Assembling and testing lasers—John is involved in designing, assembling, and installing a laser amplifier, cavity mirrors, and the associated optical elements into the laser assembly. He is also challenged by carrying out critical functions such as mirror alignment, using a reference laser to obtain a course alignment, and then doing a fine alignment by observing the beam quality and the output power. John might have to determine the optimum transmission of the laser output mirror to match the laser gain, and test it to obtain the maximum power from the laser. In this case, Equation 5-10 of this module might be a useful start to the optimization. Designing procedures for testing the quality and cleanliness of the optics as well as checking the beam quality with a commercial mode analyzer would also be important job functions.

Using lasers in various applications—Rod had a large number of opportunities when he sought a job in the area of laser applications. He found that the area covered such a wide range that he couldn't investigate all of them. He found the largest single application of lasers to be in materials interactions and materials processing. In one job possibility, Rod would be involved in setting up robotic systems to drill holes, heat treat metals, ablate materials, etc. Related job functions included designing and/or setting up beam-pointing systems, beam-focusing controls, beam-power and beam-quality measurements, and automatic feeding of materials. All are very necessary functions of a laser technician. Rod also investigated the area of medical lasers where he would be involved in arranging beam-delivery systems for laser surgery or setting up and operating laser diagnostic tools. Communication was another possibility. One particularly interesting job involved installing and testing fiber laser amplifiers both for undersea communication and for local area networks. Computers also intrigued Rod. In that area he found that he could work in optical memory storage and retrieval where he would be involved in establishing critical alignment and operation specifications for information storage. Military applications were also an exciting area. Intriguing projects in the military included working on the development of laser-guided weapons, laser range finders, and laser radar.

Field service—Donna decided she didn't want to have to work in only one location, so she investigated the possibility of being a field service technician, which would include installation and troubleshooting. In such a job, Donna would be able to travel around a given territory testing and repairing lasers and the associated equipment. The job would require significant troubleshooting skills, which appealed to her. Also she found that many of the jobs would involve extensive travel over a specific portion of the country. During job interviews, Donna was told that she would need a knowledge of not only the laser system but also the equipment that uses the laser.

Opening Demonstration

A laser is generally a very simple device. If you were standing in the center of a circle of people, each of them approximately one meter away from you, and you held a lighted match above your head, each person in the circle would see the match as having the *identical* brightness, that of a relatively weakly radiating yellowish-orange flame. This occurs because the light is radiating equally in all directions. Let's assume that you could take all the light radiating from the match and concentrate it into a single direction, say into the entrance pupil of one person's eyes. If you could do that, the intensity of the light would increase many orders of magnitude and could possibly damage the person's eye. That is how a laser is made, by redirecting the light that normally would be emitted in all directions from a material and concentrating that light into a single direction. It takes special properties of the radiating material to provide that concentration, and ordinary matches are not a material that can do that. The match was used only as an example to stress the simplicity of the concept of a laser.

Measure the diameter of the pupil of your partner's eye. You can then compute the partial solid angle $\Delta\Omega$ that would be intercepted by that pupil as a fraction of the total solid angle ($\Omega = 4\pi$ steradians) that the light is radiating in all directions. The partial solid angle $\Delta\Omega$ is defined as the area A of the intercepting detector, in this case the pupil of the eye, divided by the square of the distance R from the source (lighted match) to the pupil or

$$\Delta\Omega = \frac{A}{R^2} \quad (5-1)$$

as shown in Figure 5-2. In this case the area A would be $A = \pi r^2$ where r is the radius of the

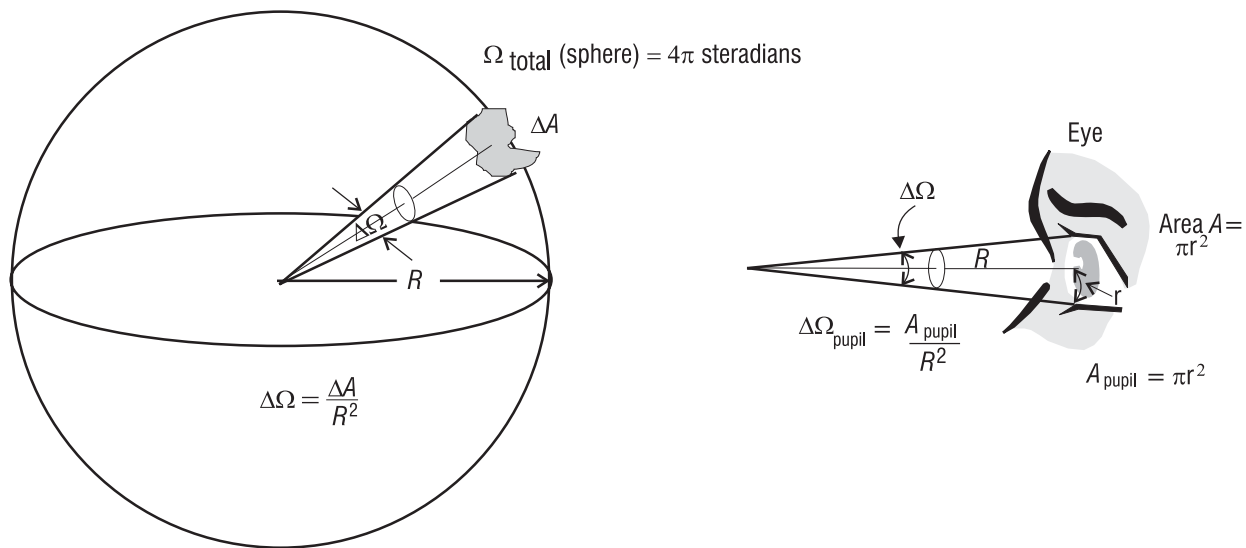


Figure 5-2 Diagram showing a partial solid angle defined by an area ΔA and by the pupil of an eye.

pupil of the eye. The amount of enhancement or gain G (dimensionless) the beam would achieve in undergoing this redirection is then the ratio of the total solid angle (light emitted in all directions) to that intercepted by the eye or

$$G = \frac{\Omega}{\Delta\Omega} \quad (5-2)$$

Based upon the value of G you obtain, you can then appreciate, with this simple demonstration, the capabilities of the laser and how such intense light beams might be produced.

Your instructor can demonstrate the purity of color of the laser beam and the low beam divergence. Your instructor can use either a prism or a transmission-diffraction grating to transmit both the beam from a flashlight and a laser beam, such as a helium-neon laser or a laser pointer, onto a screen in a darkened room. The light from the flashlight must be passed through a small aperture (approximately the diameter of the laser beam) before it is incident upon the prism to make a fair comparison with the laser. This will show how the flashlight is composed of a rainbow of colors, whereas the laser has a very discrete color or wavelength. It will also show the different beam divergences by demonstrating how little the laser beam expands (in the vertical direction) compared to the flashlight. (The comparison must be made in the vertical direction because the colors of the flashlight are expanded through the prism in the horizontal direction.)

Basic Concepts

I. REQUIREMENTS FOR A LASER

A. Laser gain medium

Nearly all lasers are produced as a result of electrons jumping from an excited energy level within a radiating species to a lower-lying energy level and, in the process, radiating light that contributes to the laser beam. Those radiating species can include:

atoms such as in the red helium-neon (HeNe) laser, the visible and ultraviolet argon ion and helium-cadmium (HeCd) lasers, and the green and yellow copper vapor lasers (CVL)

molecules such as in the infrared carbon dioxide (CO₂) laser, the ultraviolet excimer lasers such as ArF and KrF, and the pulsed N₂ laser

liquids such as those involving various organic dye molecules dilutely dissolved in various solvent solutions

dielectric solids such as those involving neodymium atoms doped in YAG or glass to make the crystalline Nd:YAG or Nd:glass lasers

semiconductor materials such as gallium arsenide or indium phosphide crystals or various mixtures of impurities blended with those and other semiconductor species

Each of the above species contains a lowest energy level referred to as the ground state in which the electrons predominantly reside at room temperature, as indicated by level 0 in Figure 5-3.

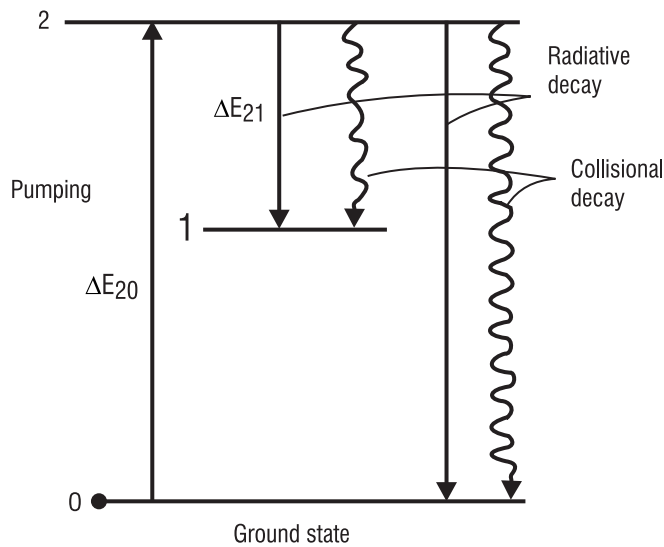


Figure 5-3 Simplified energy diagram of an atom showing excitation and emission processes

The electrons are moved to higher-lying (excited) levels such as 1 and 2 by means of various pumping processes that will be described in the next section. They then decay back to lower-lying levels within a period of time called the lifetime of the level, and eventually find their way back to the ground state when the pumping source is removed. There are three types of processes involving the interaction of light beams with atoms that have electrons residing in various energy levels. Examples of those are depicted in Figure 5-4. First an electron residing in level 2 can spontaneously jump to level 1, radiating a photon of light when it does so. That process is known as *spontaneous emission* as indicated in Figure 5-4a. Most excited energy

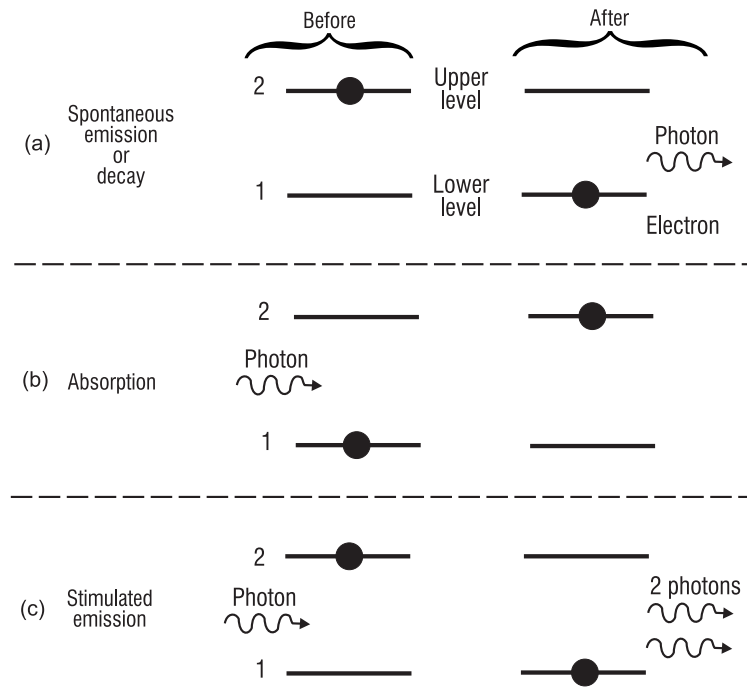


Figure 5-4 The three radiation processes that can occur when light interacts with matter (atoms)

levels undergo spontaneous emission. Each level has a specific lifetime τ over which it will remain in that level before decaying to a lower-lying level. That lifetime is determined by the interactions of the electron with the other electrons and nuclei of that atom. Typical lifetimes of electrons residing in specific levels that decay by radiating in the visible portion of the spectrum are of the order of 10–100 nsec. Of course the ground state cannot decay further and thus has infinite lifetime. The photon radiated during spontaneous emission has the exact wavelength λ_{21} and frequency ν_{21} corresponding to the difference in energy ΔE_{21} of the two involved energy

$$\Delta E_{21} = h\nu_{21} = \frac{hc}{\lambda_{21}} \quad (5-3)$$

levels (1 and 2 in this case) according to the relationship in which h is Planck's constant such that $h = 6.63 \times 10^{-34}$ joule-sec and c is the speed of light, $c = 3 \times 10^8$ m/sec. Also the wavelength λ_{21} is generally given in meters (often expressed in micrometers (μm) or nanometers (nm)). Because different materials have different energy-level arrangements, they radiate at different wavelengths and thus emit different colors or frequencies of light that are specific to the material. Nearly all the light we see originates from such transitions between energy levels of various kinds of matter.

The second process is *absorption*, shown in Figure 5-4b, which occurs if the atom has its electron in level 1 of Figure 5-3 and a photon of light of wavelength λ_{21} collides with the atom. During the collision, the photon is absorbed by the atom and the electron is moved up to the higher energy level 2. This process is the way light interacts with practically all of matter. It can happen from any energy level that is occupied (generally the ground state) and always boosts the atom to a higher-lying level while eliminating the photon. This often results in heating of the absorbing material.

The third process, shown in Figure 5-4c, is referred to as *stimulated emission*. It results when an electron is in a higher-lying level, such as level 2 in Figure 5-3, and a photon of light of wavelength λ_{21} collides with the atom. During the collision the photon stimulates the atom to radiate a second photon having exactly the same energy ΔE_{21} (and wavelength according to Equation 5-3) as that of the incident photon and traveling in exactly the same direction in order to satisfy the laws of conservation of energy and momentum. Hence, one photon leads to two identical photons, which, in effect, leads to an amplification process. A photon has been gained at the expense of the loss of energy stored within the atom.

When a large group of atoms is assembled and irradiated with light, most of those atoms are in the ground-state energy level (see Figure 5-3). If the photons of the impinging light have the appropriate energy ΔE_{20} for example, as indicated in Figure 5-3, the light will be absorbed according to the following expression for the variation of intensity I with the distance L into the material

$$I = I_0 e^{-\sigma_{20} N_0 L} \quad (5-4a)$$

in which I_0 is the intensity of the beam when it first reaches the atoms, σ_{20} is referred to as the cross section for absorption or emission of those two levels, and N_0 is the population density of atoms residing in level 0 (number of atoms per unit volume). If N_0 is in atoms/cm³ and L is in cm, the absorption cross section σ_{20} must be expressed in units of area or cm² (hence the name

cross section). Equation 5-4a indicates that the amount of beam absorption depends on both the number density of atoms residing in level 0 and the length L or thickness of the medium comprising those atoms as indicated in Figure 5-5. Also, the exponential factor suggests quite rapid absorption if the exponent is large. For example, $e^{-2} = 0.135$ and $e^{-4} = 0.018$. Hence, if either the length of the medium or the population is doubled, the beam intensity drops by nearly a factor of 8! Chemists have used this effect for many decades to measure the concentration of a material dissolved into a solvent. Equation 5-4a is known as Beer's law, and the product $\sigma_{20} N_0$ is referred to as the absorption coefficient. This absorption process is also one of the techniques used in pumping lasers in order to transfer population to level 2 from level 0, as will be discussed later.

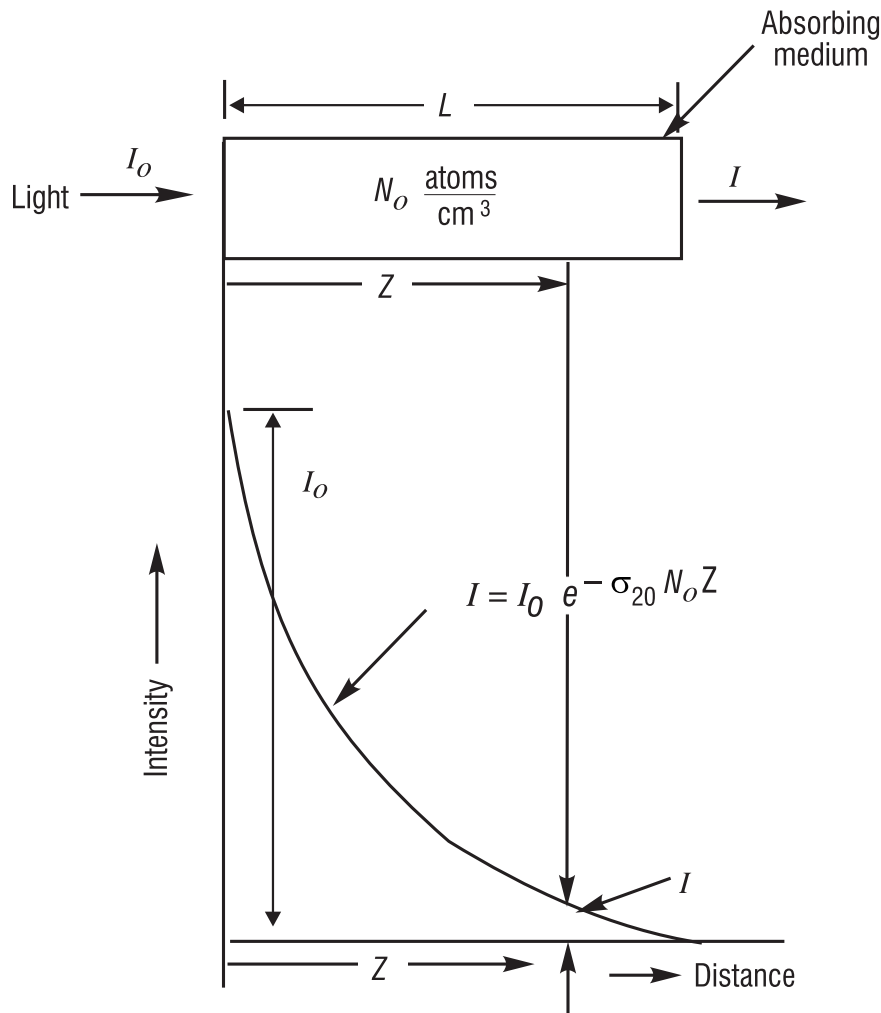


Figure 5-5 Intensity variation versus depth z into an absorbing sample

The absorption described above could have been equally applied if population initially existed in level 1, and light of energy ΔE_{21} and wavelength λ_{21} would be absorbed by the medium according to the following equation.

$$I = I_0 e^{-\sigma_{21} N_1 L} \quad (5-4b)$$

An alternative situation will now be considered. Suppose that we were able to “pump” (excite) a significant amount of population of the medium from level 0 to level 2 according to Equation 5-4a. Also, for the time being let us assume that there is no population in level 1. (This is an unlikely scenario but we will do this as a “thought” experiment for illustrative purposes.) Then again, let us consider having a beam of photons of energy ΔE_{21} and wavelength λ_{21} enter the medium. According to the earlier discussion, and considering the process described in Figure 5-4c, the only process that can occur is stimulated emission, and we would expect more photons to be generated as the beam progresses. That is exactly what happens! Since the absorption indicated in Figure 5-4b and also described in Equation 5-4a is a symmetrical process with the stimulated emission process of Figure 5-4c, it is not surprising that the beam evolves in a similar way to that of Equation 5-4a except that a sign reversal must be made in the exponent to reflect the process of photon production instead of photon absorption. This can be described mathematically in the equation below

$$I = I_0 e^{+\sigma_{21} N_2 L} \quad (5-5)$$

in which we now have the population density N_2 in the expression along with the appropriate cross section σ_{21} .

Now, if population is allowed to be in both level 1 and level 2, both absorption and stimulated emission will occur within the medium and therefore Equations 5-4 and 5-5 must be combined to give

$$I = I_0 e^{+\sigma_{21} (N_2 - N_1) L} \quad (5-6)$$

as indicated in Figure 5-6. Hence, if more population exists in level 2 than in level 1, N_2 will be greater than N_1 and the exponent of Equation 5-6 will be positive. The beam will grow and emerge from the medium with a greater intensity than when it entered. In other words, for amplification or gain to occur, the condition must be

$$\frac{N_2}{N_1} > 1 \quad (5-7)$$

Having N_2 be larger than N_1 is known as having a *population inversion*, which is not a normal, naturally occurring relationship. This would be the equivalent of having a mountain in which there is more dirt at higher levels than at lower levels. The mountain would taper inward toward the bottom rather than outward, which is generally an unstable situation. The only way to maintain such an “inversion” is to continually transfer or “pump” more dirt up to higher levels by a conveyor belt or some other process. The equivalent transfer to higher levels, or “pumping” is also required in lasers to maintain the population inversion of level 2 with respect to level 1 such that amplification can be produced.

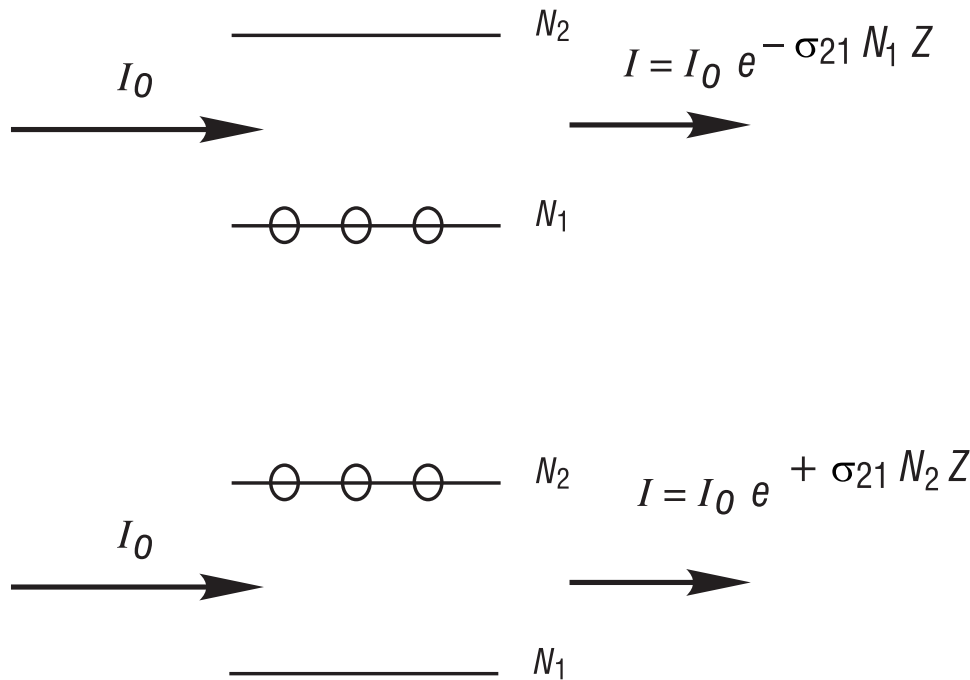


Figure 5-6 Absorption and stimulated emission effects combined in a laser gain medium

Population inversions in gases—Inversions in gases are generally produced by applying a voltage across a gas discharge tube that consists of a long, narrow glass or ceramic tube serving to confine the gain medium, with electrodes installed at each end of the tube. In its simplified form the electrodes, which are essentially electrical feedthroughs, are attached to each end of the tube to allow a voltage to be applied across the length of the tube. The tube is then filled with a low-pressure gas or gas mixture that includes the species that will serve as the gain medium. The applied voltage produces an electric field within the laser tube that accelerates the electrons within the gas. Those electrons collide with the gas atoms and excite the atoms to excited energy levels, some of which serve as upper laser levels. Lower-lying levels, those to which higher-lying levels can transition, typically decay to the ground state faster than the higher-lying levels, thereby establishing a population inversion between some of the higher and lower levels as indicated in Figure 5-7. This inversion can be envisioned by considering that, if the lower levels drain out faster than the upper levels, there will be less population left in those lower levels than in the higher-lying levels. The laser light then occurs when the higher-lying levels decay to the lower levels while radiating photons at the wavelengths corresponding to the energy separation between the levels. In many instances the excitation is a two-step process in which the electrons first excite a long-lived or metastable (storage) level or they ionize the atom, leaving an ion of that species and another electron. In either case, that level then transfers its stored energy to the upper laser level via a subsequent collision with the laser species. The laser transitions in gaseous laser media typically occur at relatively precise, discrete wavelengths that correspond to the energy difference of inherently narrow energy levels.

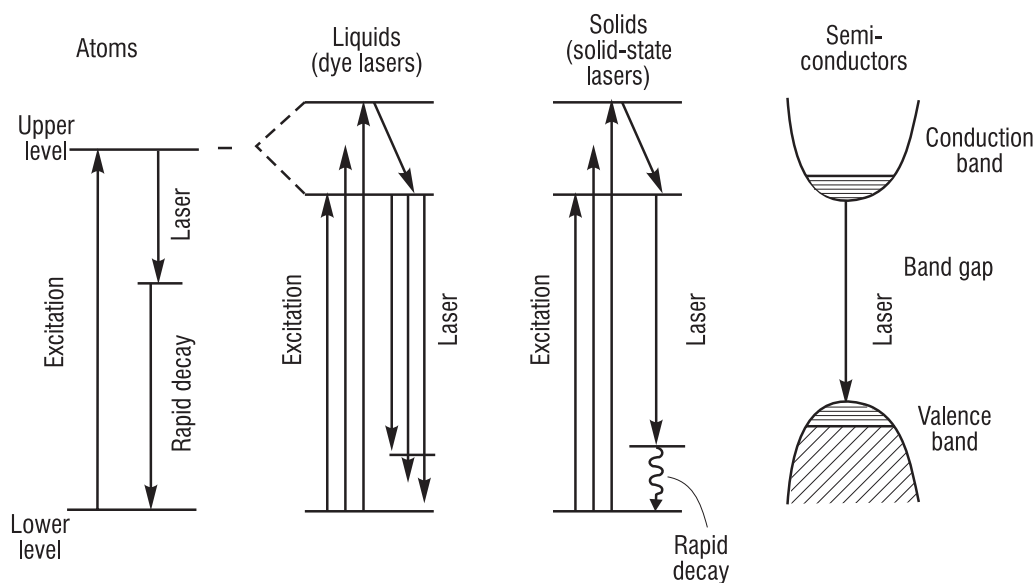


Figure 5-7 Inversion processes in gases, liquids, solids, and semiconductors

Population inversions in liquids—Most excited energy levels in liquids decay so rapidly due to collisions with the surrounding nearby atoms or molecules that they can't stay around long enough to participate in a lasing process. There are some molecules however, namely organic dye molecules, that do have a sufficiently long lifetime in an upper energy level (of the order of 1–5 nsec) so they can participate in the laser process by being excited to an upper laser level. These molecules also have the ability to radiate the energy from that level rather than lose the energy due to decay by collisions. Those molecules are the dyes that are used to color cloth and other objects that we use in our everyday life. When dissolved in a solvent such as alcohol or water, they can be concentrated in sufficient quantity to be used as a laser gain medium. In these dissolved dye solutions, electrons *cannot* be made to flow in the form of an electrical current within the liquid as they can in gases. Therefore the pumping of the upper laser levels must be carried out by optical means such as a flashlamp or another laser as shown in Figure 5-7. When the light is applied to the dye solution, it is absorbed at certain wavelengths by the dye as described by Equation 5-4a, placing the dye molecules in highly excited upper laser levels. A population inversion is then produced between those levels and a very broad range of lower-lying energy levels, thereby allowing the possibility for a *wide range of laser wavelengths* to be produced within the gain medium. Those lower levels are not initially pumped by the light and therefore are sufficiently empty to produce the inversion. Dye lasers thus allow the possibility of wide wavelength *tunability* and have been used extensively in doing a variety of spectroscopic studies in which very specific laser wavelengths are desired.

Population inversions in crystalline solids and glasses—As in the case of liquids, when energy levels in solids are excited, typically by irradiating those solids with light, the levels tend to decay much more rapidly via collisions with their surrounding neighbors rather than by radiating their energy in the form of light. In a few cases, however, specific types of atoms are embedded into a transparent host material (such as a specific crystalline solid or a glass) at concentrations of up to 1 part in 100, and the atoms radiate their energy rather than decay by collisions. These specific types of atoms, such as chromium or neodymium, consist of a radiating electron surrounded by a “screen” of other electrons that protect that radiating electron

from being bombarded by collisions from neighboring atoms. The consequence is that the atoms can absorb pump light that passes through the transparent host medium and can then subsequently radiate that energy. Gemstones such as rubies fall into that category. Ruby, a desired gemstone and also the material that comprised the gain medium for the first laser, consists of chromium atoms doped into a transparent sapphire (Al_2O_3) host crystal. The color of the ruby crystal is determined by the chromium atoms, which absorb light in the blue and green regions of the spectrum and radiate in the red.

When these types of laser crystals absorb light, the energy ends up in excited energy levels that serve as the upper laser level. These crystals have the property that the upper laser level has a very long lifetime before it decays by radiating when compared to all other types of laser gain media. The population inversion in most of these lasers occurs by the lower laser levels being rapidly depleted by collisions with the neighboring atoms (see Figure 5-7) since these levels are not screened or protected as are the upper laser levels. An exception to this is the ruby laser in which the lower laser level is the ground state. In this case the pumping power must be excessively high in order to pump more than half of the chromium atoms into the upper laser level to produce an inversion.

In these solid-state laser gain media, some of the doping atoms produce very broad excited energy levels and others have very narrow energy levels. The broad energy levels allow a broad wavelength region over which gain or amplification occurs and thus allow broad wavelength tunability of the lasers. The narrow energy levels produce lasers operating over a very narrow wavelength region or narrow bandwidth.

Population inversions in semiconductors—Inversions in semiconductors are produced when joining a *p*-doped semiconductor material with an *n*-doped semiconductor material in a similar way to that of producing a transistor to create a *pn* junction. The *n*-doped material contains an excess of electrons and the *p*-doped material has an excess of holes (a material with excess positive charge). When a voltage is applied across the junction, with the positive voltage on the *p* side, the electrons are pulled through the junction toward the positive electrode and the holes are attracted to the negative side, producing an electrical current flow across the junction. The electrons and holes meet within the junction and are attracted to each other because of opposite charges. When they meet, they recombine and emit radiation and also can produce a population inversion. This inversion occurs between energy levels located above and below the semiconductor bandgap (see Figure 5-7), the gap in energy below which the material is transparent. This energy typically corresponds to a wavelength in the infrared, and hence most semiconductors radiate in the infrared and are not transparent in the visible spectral region like glass is. However, semiconductor lasers are under development to operate in the green and blue regions of the spectrum. At very low currents, a population inversion does not occur even though recombination radiation is emitted. In fact, such nonlaser-like emission is the source of radiation from a light-emitting diode (LED). In comparison, to produce a population inversion, a very high current density is applied within the junction region. However, this high current density leads to excessive heat deposition in the material; therefore a significant part of the development of semiconductor lasers involves how to remove the heat, or to make smaller junctions so that less current is required. The material and its corresponding energy bandgap determine the laser wavelength.

Equation 5-6 describes the way in which a beam is amplified if a population inversion exists between two energy levels such as 1 and 2, as described above. An inversion is a necessary

condition for making a laser but not a sufficient condition. The exponential factor in Equation 5-6 must be high enough for the beam to grow and develop into the kind of narrow beam that is expected from a laser. For example, if the exponent turns out to have a value of only 0.00001, there will be no noticeable beam growth even though Equation 5-7 might be satisfied. The exponent of Equation 5-6 consists of a cross section σ_{21} that is characteristic of a specific material and also a specific radiative transition in that material. It is referred to as a cross section because it has dimensions of length² or area, as we mentioned earlier. Table 5-1 lists cross sections for some of the laser transitions described in this course. The population difference $N_2 - N_1$, which is sometimes expressed as ΔN_{21} , is a value determined by the power available from the pumping source. Values of ΔN_{21} are also shown in Table 5-1. Such pumping processes are described in the next section. The other factor that affects gain is the length of the gain medium. If σ_{21} and ΔN_{21} are not quite sufficient, the length L of the gain medium can be increased to increase the exponent of Equation 5-6. We will show later that the amplifier length L can be *effectively* increased by putting mirrors around it such that the beam will pass back and forth through it many times during the beam's growth process.

Table 5-1

Type of Laser	$\lambda_{21}(\text{nm})$	$\Delta\nu_{21}(\text{Hz})$	$\sigma_{21}(\text{cm}^2)$	$\Delta N_{21}(\text{cm}^{-3})$	$g_{21}(\text{cm}^{-1})$	$I_{\text{sat}}(\text{W}/\text{cm}^2)$
HeNe	632.8	2×10^9	3×10^{-13}	7×10^9	2×10^{-3}	6.2
Argon	488.0	2×10^9	2.5×10^{-12}	1×10^{15}	5×10^{-3}	16.3
HeCd	441.6	2×10^9	9×10^{-14}	4×10^{12}	3×10^{-3}	7.1
Copper Vapor	510.5	2×10^9	8×10^{-14}	6×10^{13}	5×10^{-2}	9.0
CO ₂	10,600	6×10^7	3×10^{-18}	5×10^{15}	8×10^{-3}	1.6×10^{-2}
Excimer	248.0	1×10^{13}	2.6×10^{-16}	1×10^{16}	2.6×10^{-2}	3.4×10^5
Dye (Rh6-G)	577	5×10^{13}	2×10^{-16}	2×10^{18}	2.4	3.4×10^9
Ruby	694.3	3×10^{11}	2.5×10^{-20}	4×10^{19}	1.0	3.8×10^7
Nd:YAG	1064.1	1.2×10^{11}	6.5×10^{-19}	3×10^{19}	2.0	1.2×10^7
Ti:Al ₂ O ₃	760	1.5×10^{14}	3.4×10^{-19}	3×10^{18}	1.0	2.0×10^9
Semiconductor	800	1×10^{14}	1×10^{-15}	1×10^{18}	10^3	2.5×10^9

It is useful to describe the product of σ_{21} and ΔN_{21} as the small-signal-gain coefficient g_{21} or

$$g_{21} = \sigma_{21} \Delta N_{21} \quad (5-8)$$

Hence, Equation 5-6 can be rewritten as

$$I = I_0 e^{g_{21} L} \quad (5-9)$$

By considering the units of both σ_{21} (length²) and ΔN_{21} (1/length³) we can see that g_{21} has the units of 1/length. Hence, if σ_{21} is given in units of cm² and ΔN_{21} is given in units of (1/cm³), g_{21} will be given in (1/cm), more commonly expressed as cm⁻¹. Values of the cross sections σ_{21} and ΔN_{21} , and the small-signal gain g_{21} , are listed in Table 5-1 for some of the lasers described in this module.

Bandwidth of laser gain medium—The bandwidth of the laser gain medium determines the range of wavelengths over which amplification can occur for any specific laser. This bandwidth is expressed in either a wavelength range $\Delta\lambda_G$ or a frequency range $\Delta\nu_G$. These two expressions are related by

$$\Delta\lambda_G = \frac{\lambda^2}{c} \Delta\nu_G \quad (5-10)$$

in which λ is the laser wavelength and c is the speed of light. The bandwidth of the gain medium is usually determined by the bandwidth over which the spontaneous emission occurs for a given laser transition. This bandwidth is determined by specific properties of the energy levels involved in the transitions, such as their lifetimes, how the atoms interact with other atoms, how closely the atoms are arranged, etc. Typically, atomic gas lasers have bandwidths of the order of 1 GHz (10^9 Hz). Molecular lasers have bandwidths that are sometimes a factor of 10 to 10,000 larger than that due to the closeness in wavelength of several molecular transitions that overlap in frequency. Solid-state lasers can have relatively narrow bandwidths of the order of 100 GHz in cases such as the Nd:YAG laser, or very wide bandwidths, of the order of 100 THz (10^{14} Hz) in the case of the titanium sapphire laser. Semiconductor lasers have bandwidths typically of 10^{13} Hz. Comparisons of the laser gain bandwidths for the HeNe, Nd:YAG, and Ti:Al₂O₃ lasers are shown in Figure 5-8. These various bandwidths are not the bandwidths of the laser beam that emerges from the amplifier but do indicate the range over which amplification can occur. Laser mirror cavity properties primarily determine the bandwidth of the emerging laser beam, as will be described later under laser beam properties.

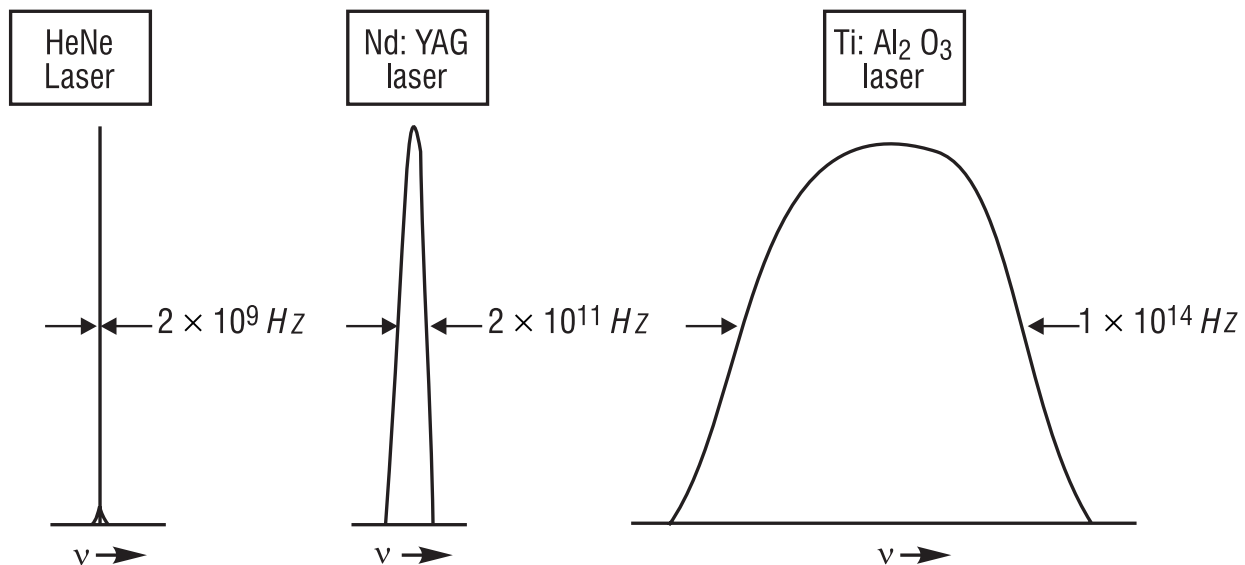


Figure 5-8 Laser gain bandwidths for the HeNe, Nd:YAG, and Ti:Al₂O₃ lasers

B. Laser pumping sources

Laser pumping sources are the means by which energy is transferred into the laser gain medium to produce the required population inversion ΔN_{21} . These pumping sources generally consist of either electrons flowing within the medium or light being absorbed by the medium.

Electron pumping—Electron pumping is used primarily in gaseous or semiconductor gain media. In gases, many electrons are produced when a few initial electrons within the gain medium are accelerated by an electric field within the medium and these many electrons then collide with neutral atoms, exciting those atoms to higher-lying energy levels and even ionizing some of the atoms (removing an electron). The freed electrons are also accelerated, producing an avalanche of electrons and therefore an electrical current within the medium. The electrons lose their energy by transferring it to the atoms during the collision process. Some of the lasers operate on a pulsed basis, applying a large amount of current for a short period of time. Others operate on a continuous (cw) basis, using a much smaller but continuous current.

In semiconductors, the electrons flow through the semiconducting material by applying a voltage across the *pn* junction with the positive voltage on the side of the *p*-type material. This leads to recombination radiation when the electrons combine with the holes in the junction. The heat loading of the semiconductor limits the current.

Optical pumping—Optical pumping of lasers generally applies to the pumping of liquid (dye) lasers and to dielectric solid-state lasers and is provided by either flashlamps or other lasers.

The most common types of flashlamps used for pumping lasers are narrow, cylindrical quartz tubes with metal electrodes mounted on the ends, filled with a gaseous species such as xenon that serves as the radiating material within the lamp. A voltage is applied across the electrodes of the flashlamp and current flows through the gas, populating excited levels of the atoms within the gas that radiate and produce intense light emission. The process is similar to that of electron excitation of lasers described above except that a population inversion is not produced and the radiating material of the lamp radiates via spontaneous emission, rather than by stimulated emission as in the case of a laser gain medium. The pumping wavelength of the flashlamp is determined by the gaseous medium inserted within the flashlamp tube. Xenon is the most common species because of both its radiating efficiency and its emission of a broad spectrum of wavelengths from which to choose in matching the lamp emission to the pumping absorption bands of the laser.

Examples of flashlamp configurations for pumping lasers are shown in Figure 5-9. Figure 5-9a shows the flashlamp in the form of a helix wrapped around the laser rod. Figures 5-9b and 5-9c show the flashlamp inserted into an elliptically-shaped or circularly-shaped elongated laser cavity. In Figure 5-9b the flashlamp is located at one focus of the ellipse and the laser rod to be pumped at the other focus of the ellipse. Figure 5-9d shows two flashlamps used in a double elliptical cavity, one of the most favorable arrangements, with the laser rod in the center. Sometimes the laser mirrors are coated onto the ends of the laser rod and sometimes they are mounted externally to the rod, along the longitudinal axis of the cavity.

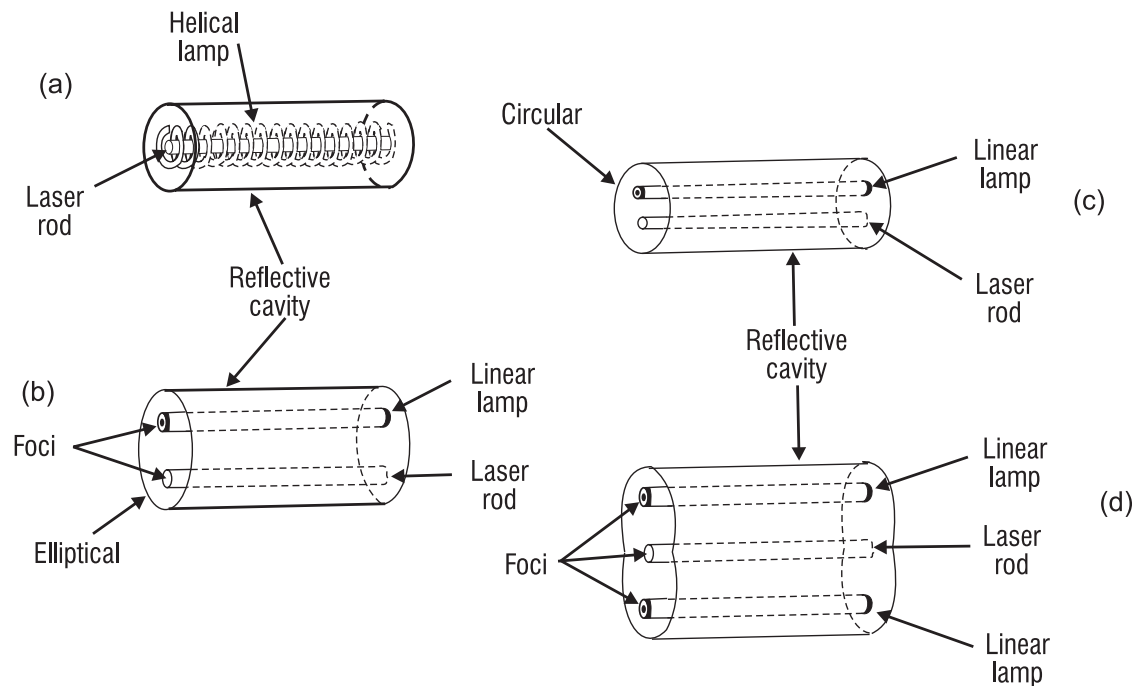


Figure 5-9 Flashlamp configurations for pumping lasers

Laser pumping is used in cases in which the pumping energy must be concentrated into a relatively small volume or for a very short time, or if the pumping wavelength must be provided over a fairly narrow-wavelength bandwidth. Pumping lasers include the argon ion or doubled Nd:YAG cw lasers for pumping titanium-sapphire lasers, excimer lasers for pumping dye lasers, and gallium arsenide semiconductor lasers for pumping Nd:YAG lasers. In most cases the laser is focused to a relatively small gain region, a line focus for dye lasers and a spot focus for the other lasers. Two examples of diode pumping of Nd:YAG lasers are shown in Figure 5-10.

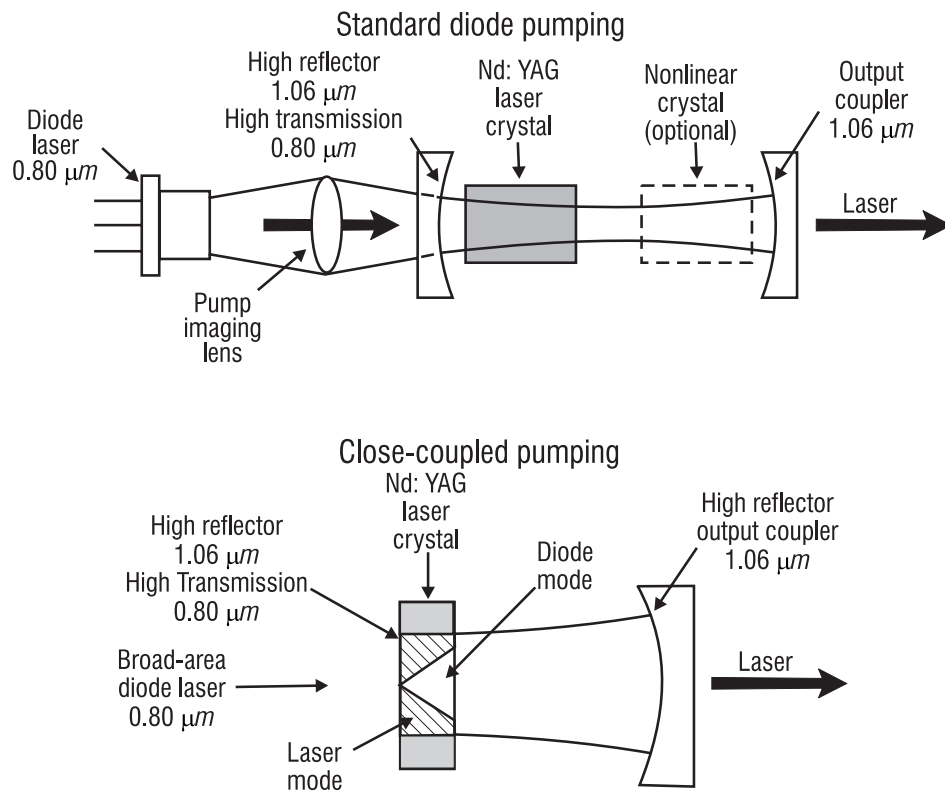


Figure 5-10 Examples of laser diode pumping of Nd:YAG lasers

C. Laser beam properties

Laser beam properties such as direction and divergence of the beam, the beam profile, and the wavelength and frequency characteristics of the laser within the wavelength region of the laser gain bandwidth are determined largely by the laser mirrors. The factors determining those properties include mirror curvature, surface quality, and reflectivity as well as separation and location, assuming that the structure holding the mirrors is a secure, vibration-free structure. The unique electromagnetic wave properties produced by the mirrors are referred to as modes. Before discussing these mirror properties, we must consider the shape of the gain medium and the beam growth to the point of beam saturation.

Shape of gain medium—The goal of constructing a laser is to capture most of the spontaneous emission that is normally emitted in all directions within the gain medium and redirect it into a single direction. This is done with the assistance of the gain or amplification that can be initiated within the medium. It is desirable to have the gain medium be of an elongated shape so that the gain, which is length dependent, will operate primarily in that one elongated direction. Hence, most laser gain media are long, narrow devices with mirrors located at the ends.

Growth of beam and saturation—If significant gain is provided along the length of the gain medium, the spontaneous emission emitted in the elongated direction will grow at a rate dependent upon the amount of gain available as it moves through the length of the medium. The emission that starts at one end and transits to the other end will have grown by a factor of between 0.02 (2%) and 10 (1,000%) in a single pass, depending upon the type of laser.

However, even the high factor-of-10 growth available in some lasers is not sufficient to produce a powerful unidirectional laser beam in one pass. Hence, mirrors are placed at both ends of the medium, forming a cavity to redirect the beam back and forth through the amplifier and thereby allow the beam to continue to grow until a point of beam saturation is achieved. At somewhere between 2 passes (dye lasers) and 500 passes (HeNe lasers), the beam will have become so intense within the laser cavity that there won't be sufficient atoms in the upper laser level within the gain medium to accommodate all of the impinging photons. This is the condition of beam saturation, and the intensity of the beam is known as the saturation intensity. Values of saturation intensity for a variety of lasers are given in Table 5-1.

The condition of saturation occurs when the exponent of Equation 5-6 ($\sigma_{21}\Delta N_{21}L$) reaches a value of approximately 12. In this saturation, the length L is the effective length of many passes through the amplifier so we will define that length as L_T . Hence, we have the condition to reach saturation as ($\sigma_{21}\Delta N_{21}L_T \cong 12$) or ($g_{21}L_T \cong 12$). At that point the beam will have grown by a factor of approximately $e^{12} = 1.6 \times 10^5$ according to Equation 5-6. When it reaches that intensity it will settle down to a stable value (as long as the pumping continues) in which the conversion from pump power to laser photons reaches an equilibrium balance. In the case of the HeNe laser described above, that requires 500 passes through the amplifier. If the amplifier length is 20 cm, the effective length in Equation 5-6 is 500×20 cm or 100 m. Hence, the beam travels through 100 m of amplifier length before it reaches saturation. If more pump power is applied, above the value where the saturation intensity is reached, more laser power will be produced and will be available through the output mirror of the laser.

The condition for a population inversion and thus amplification within the amplifier was given by Equation 5-7 ($N_2/N_1 > 1$). However, even though gain might exist within the amplifier, the laser still might not develop a beam if the gain is not sufficiently high to overcome losses within the laser cavity. The laser mirrors won't have 100% reflectivity and there might be additional losses such as scattering and reflective losses at windows and other optical elements placed within the cavity. Hence, a simple expression is used to determine the threshold condition for a laser beam to develop, based upon the laser cavity characteristics. For a laser in which the amplifier length has a value of L and the mirrors have identical reflectivities R , with no other losses in the cavity, the threshold condition for the gain coefficient g is given as

$$g = \frac{1}{2L} \ln \frac{1}{R^2} \quad (5-11)$$

which has dimensions of 1/length. Any value of g higher than that given by Equation 5-11 will produce a laser beam within the cavity. For a more complex laser cavity in which the mirrors have different reflectivities R_1 and R_2 , and a_1 and a_2 represent other losses within the cavity (beyond the amplifier), the expression for the threshold gain g is given as

$$g = \frac{1}{2L} \ln \left[\frac{1}{R_1 R_2 (1 - a_1)(1 - a_2)} \right] + \alpha \quad (5-12)$$

The term α represents a potential absorption loss within the amplifier itself, which is present in only a few types of lasers. It is a distributed loss expressed in the same units as g or (1/length). For example, in solid-state lasers it is termed excited state absorption.

Example 1

Consider a HeNe laser in which the mirror reflectivities might be $R_1 = 0.999$ (99.9%) and $R_2 = 0.990$ (99%) and the cavity losses are $a_1 = a_2 = 0.002$ (0.2%) and $\alpha = 0$. For that situation calculate the gain per pass that would be necessary to operate the laser at threshold.

Using Equation 5-12 for the gain coefficient, we can obtain

$$g = \frac{1}{2 \times 20} \ln \left[\frac{1}{(0.999)(0.990)(1 - 0.002)(1 - 0.002)} \right] = 0.00038/\text{cm} = 0.038\%/\text{cm}$$

Hence the increase over a 20-cm-length amplifier would be

$$gL = 20 \text{ cm} \times (0.00038)/\text{cm} = 0.0076 \text{ or } 0.76\% \text{ per pass.}$$

The useful power from the laser is obtained by locating a partially transmitting “output” mirror at one end of the amplifier so that part of the beam “leaks out” of the mirror cavity as shown in Figure 5-1. The initial gain in the amplifier must be greater than the loss of the transmitting mirror (plus other mirror and cavity losses) or the beam will not develop as described in Equations 5-8 and 5-9. A simple expression for the optimum mirror transmission T_{opt} in terms of the small-signal-gain coefficient g , the actual amplifier length L , and the absorption loss a (averaged over a single pass from one mirror to the other) can be expressed as

$$T_{opt} = (gL a)^{1/2} - a \quad (5-13)$$

Example 2

For the HeNe laser given in Example 1, assume that the gain is 10 times the threshold value or $g = 10 \times 0.00038 = 0.0038$ and $L = 20$ cm. Also assume that the absorption loss a is an average of a_1 and a_2 as defined above, or $a = 1/2(a_1 + a_2) = 0.002$. Compute the optimum mirror transmission for that situation.

Using Equation 5-13 we have for T_{opt}

$$T_{opt} = [(0.0038)(20)(0.002)]^{1/2} - (0.002) = 1.04\%$$

A mirror reflectivity of 98.96% or approximately 99% would be the appropriate reflectivity of the output mirror. This was in fact the transmission used in Example 1 for R_2 .

The laser beam output intensity I_{max} emitted from the output mirror can also be estimated in terms of the saturation intensity I_{sat} , T_{opt} , and the average absorption per pass a in the following expression

$$I_{max} = \left(\frac{T_{opt}^2}{2a} \right) I_{sat} \quad (5-14)$$

Example 3

For the HeNe example given above, estimate the power output from the laser. Using Equation 5-14 and the conditions described in the example above, $T_{opt} = 0.0104$, $a = 0.002$, as well as the value of the saturation intensity from Table 5-1 of $I_{sa} = 6.2 \text{ W/cm}^2$, we find that the maximum output power is $I_{max} = 167 \text{ mW}$. This is on the high end of the power spectrum for a HeNe laser.

Longitudinal cavity modes—When the beam is developing within the mirror cavity, traveling back and forth, certain wavelengths within the gain bandwidth of the laser tend to be more enhanced than others. These are wavelengths (or frequencies) in which the light beam in the cavity forms a *standing wave*. Such an effect occurs when an exact number of half-wavelengths of the light fit within the separation distance between the mirrors. Typically there will be several hundred thousand wave peaks for each standing wave that occurs within the cavity. Hence, each standing wave must have a wavelength such that an integral number of oscillating waves fits in the space separating the mirrors. If more than one standing wave is present, each standing wave (longitudinal mode) will be separated in frequency from the next one by a fixed exact amount that depends upon the laser cavity length d . That frequency separation $\Delta\nu$ between longitudinal modes can be obtained by dividing the speed of light c by twice the cavity length or

$$\Delta\nu = \frac{c}{2d} \quad (5-15)$$

In Figure 5-11, several of these modes are shown occurring within the frequency bandwidth of a typical gas laser. Typically, the separation in frequency is of the order of 500 MHz ($5 \times 10^8 \text{ Hz}$) whereas the laser frequency itself is of the order of 500,000,000,000,000 Hz ($5 \times 10^{14} \text{ Hz}$). For example, two of these discrete standing waves might have frequencies of 500,000,000,000,000 Hz and 500,000,500,000,000 Hz, separated in frequency by 500 MHz.

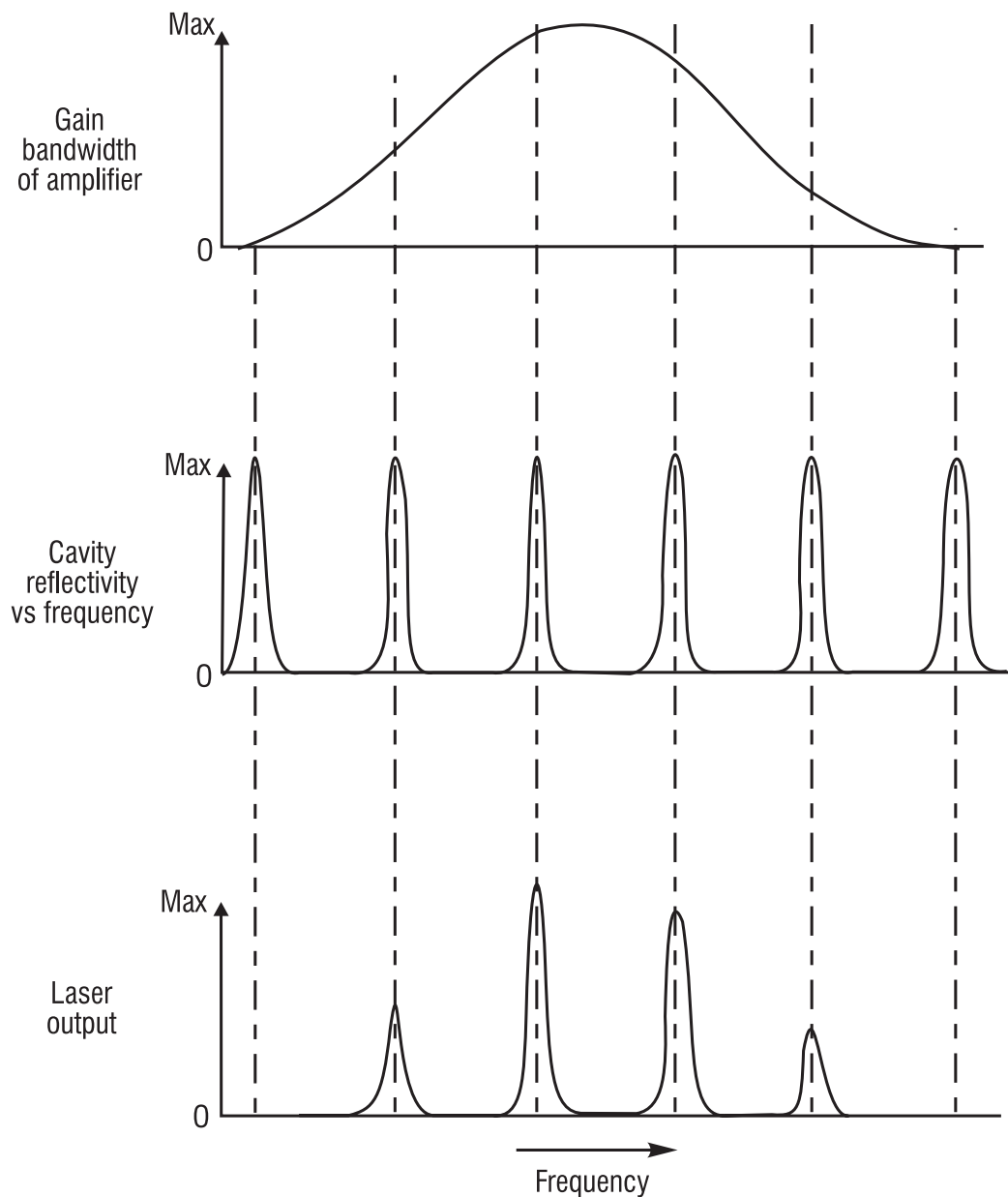


Figure 5-11 Several longitudinal modes are shown occurring within the gain bandwidth of a typical gas laser.

Each discrete standing wave is referred to as a longitudinal mode associated with the laser cavity. Figure 5-12 shows two such modes within a cavity. There will always be at least one longitudinal mode and there could be many more, depending on the frequency or wavelength bandwidth of the laser gain medium. If more than one longitudinal mode is being generated, they will be indistinguishable unless a spectrum analyzer is used to analyze the beam. They all travel in the same direction, and their color will be indistinguishable because their wavelengths (frequencies) are so similar, as indicated above.

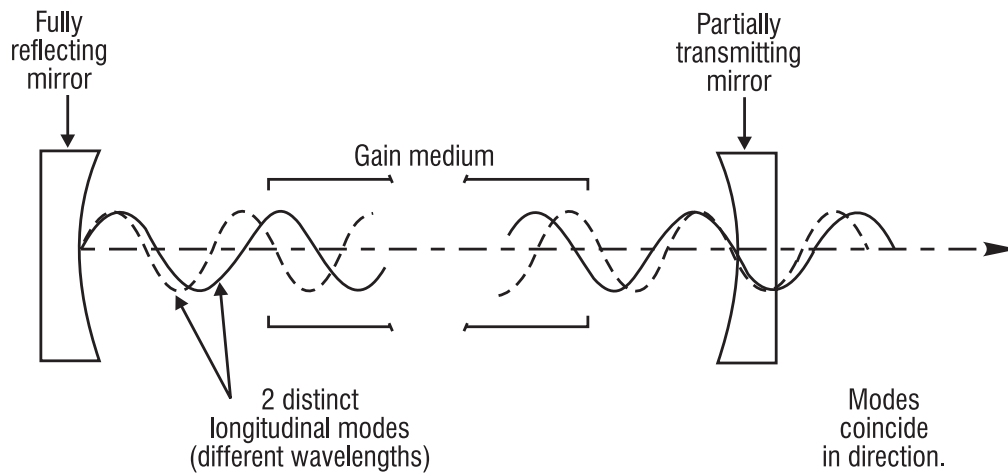


Figure 5-12 Two distinct longitudinal modes operating simultaneously in the same laser cavity

The frequency width of a single longitudinal mode can be very narrow, typically in the range of 10^6 to 10^8 Hz, determined by the mirror reflectivity (higher-reflecting mirrors produce narrower bandwidths) and by the cavity stability (free of vibrations).

Transverse modes—The presence of more than one longitudinal mode involves many light beams traveling exactly the same path through the amplifier but differing in wavelength depending upon the total number of wave cycles that fit between the mirrors. Contrary to this, different transverse modes involve slightly different optical paths through the amplifier and thus have slightly different directions when they emerge from the laser as shown in Figure 5-13. Because of the different optical path lengths, they also have slightly different frequencies. Each of these stable modes evolves because the light traveling that particular pathway recurs exactly from one round trip of the beam to the next, therefore developing into a steady beam. Each transverse mode traveling over its unique path might also consist of several longitudinal modes separated in frequency according to Equation 5-15.

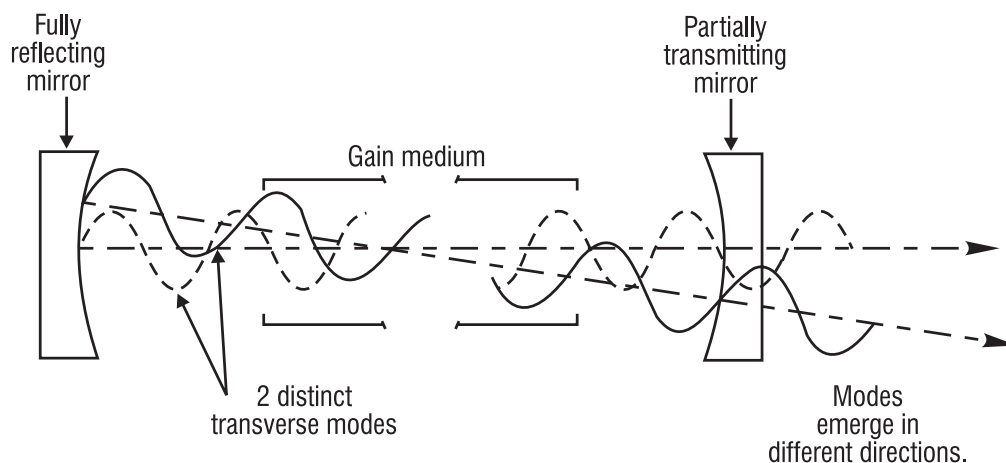


Figure 5-13 Two transverse modes occurring simultaneously within a laser cavity. The on-axis mode is the TEM_{00} mode. The angled mode is actually rotationally symmetric and would produce a doughnut spot on the wall.

The lowest-order transverse mode, known as the TEM_{00} mode, travels down the central axis of the laser gain medium. Higher-order modes have slightly diverging beams as shown in Figure 5-13. The TEM_{11} mode, for example, if it were the only mode present, would appear as a doughnut-shaped beam when projected onto a screen. Complex patterns can be present if several transverse modes are operating. In most cases, closely located transverse modes differ in frequency by a smaller value than do adjacent longitudinal modes that follow the same path through the amplifier.

The TEM_{00} mode has a beam-intensity profile in the direction transverse to the direction of propagation that is described by a Gaussian function as given by the following expression.

$$I = I_0 e^{-\frac{2r^2}{w^2}} \quad (5-16)$$

where I_0 is the intensity on the beam axis at any location, r is the perpendicular distance from the beam axis, and w is defined as the beam waist. This beam profile is shown on the left side of Figure 5-14. The beam waist, varying along the axis of the laser, is defined such that the intensity has fallen to $1/e^2$ of the intensity on axis. It turns out that 86.5% of the energy is contained within the beam radius in which $r = w$. The TEM_{00} mode is often the desired mode because it propagates with the least beam divergence and can be focused to the tightest spot. It can generally be obtained by placing an adjustable aperture within the laser cavity and decreasing the aperture diameter until only the TEM_{00} mode remains.

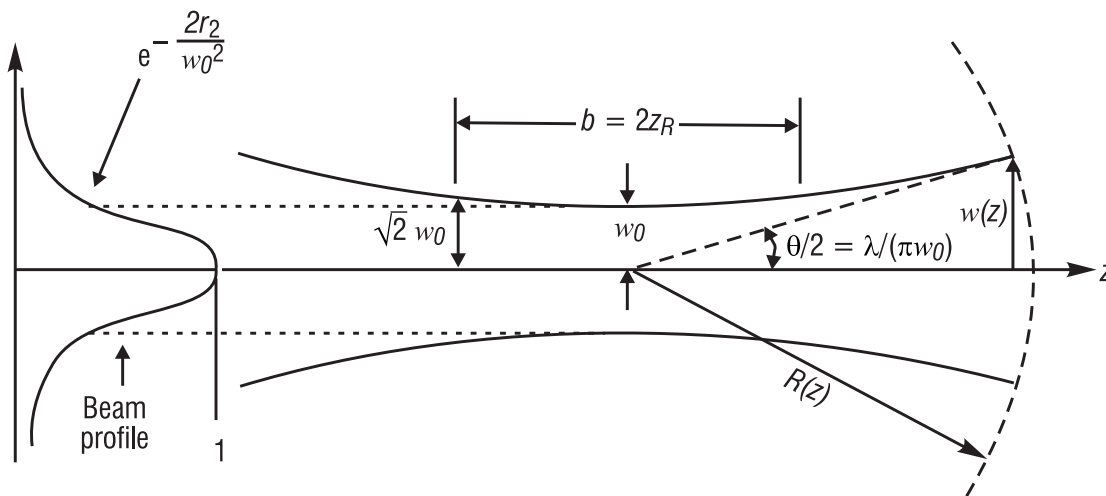


Figure 5-14 A diagram showing some of the parameters of a Gaussian laser beam (TEM_{00} mode)

Gaussian beams have a minimum beam waist w_0 that usually occurs somewhere between the laser mirrors. The beam then gradually expands from that location. If the laser mirrors have the same radius of curvature, the minimum beam waist occurs exactly halfway between the mirrors. If the minimum beam waist is known, the beam waist $w(z)$ at any distance z from where the minimum occurs can be determined from the following equation (provided the beam does not interact with any type of optical element that would change the beam in any way).

$$w(z) = w_0 \left[1 + \left(\frac{\lambda z}{\pi w_0^2} \right)^2 \right]^{1/2} \quad (5-17)$$

where λ is the wavelength of the beam. This is diagrammed in Figure 5-14. The expanding beam has a curved wavefront with a radius of curvature $R(z)$ given by

$$R(z) = z \left[1 + \left(\frac{\pi w_0^2}{\lambda z} \right)^2 \right] \quad (5-18)$$

The beam angular spread Θ in radians at distances well beyond the laser mirrors can be expressed as

$$\Theta = \frac{2\lambda}{\pi w_0} \quad (5-19)$$

as shown in Figure 5-14. This angular divergence Θ can be approximately determined by measuring the beam diameter at a known distance from the laser (by projecting the beam onto a screen) and then finding the ratio of the beam radius (half the diameter) to the distance from the laser. From Equation 5-19, it can be seen that a larger w_0 and/or a shorter wavelength λ gives a smaller angular beam divergence.

Also shown in Figure 5-14 is the value of $z = z_R$ such that

$$z_R = \frac{\pi w_0^2}{\lambda} \quad (5-20)$$

This value z_R is referred to as the Rayleigh range and is considered the depth of focus for a Gaussian beam. Twice the value of z_R or $b = 2 z_R$ is known as the confocal parameter. At a distance z_R in either direction from the location of the minimum beam waist w_0 , the value of the beam waist has increased to $\sqrt{2} w_0$ as shown in Figure 5-15. When a confocal laser cavity is set up (as described later) the *diameter* of the beam at the mirrors will thus have the value of $2\sqrt{2} w_0$.

For laser cavities in which the mirrors have the same radii of curvature, the expression for w_0 can be given as

$$w_0 = \left\{ \frac{\lambda}{2\pi} [d(2R-d)]^{1/2} \right\}^{1/2} \quad (5-21)$$

If the radius of curvature is equal to the separation between mirrors, or $R = d$, the cavity is known as a confocal resonator. For this case the expression for w_0 then simplifies even further to

$$w_0 = \left(\frac{\lambda d}{2\pi} \right)^{1/2} \quad (5-22)$$

Another very common laser cavity arrangement is known as the semi-confocal resonator cavity. This case uses a curved mirror and a flat mirror (infinite radius of curvature) with the mirror separation being *half* the radius of curvature. This, in effect, is one-half of the confocal resonator as described above because it folds half of the resonator back on itself by using a flat mirror at the halfway location. This is a common resonator to obtain a parallel beam at the output of the laser, if the output mirror is the flat mirror.

Stable laser cavities or resonators—A variety of laser mirror arrangements might be considered in constructing a laser cavity, as shown in Figure 5-15. In that figure, the mirror radii of curvature are defined in terms of the separation distance d between the mirrors. However, not all of these cavities allow for stable laser operation. Such stability is obtained if the laser beam is reflected back toward the axis of the amplifier rather than sending the beam in a diverging direction after each mirror reflection. Of those cavities shown, the stable ones are the large-radius mirrors (Figure 5-15b), the confocal cavity (Figure 5-15c), the semi-confocal cavity (Figure 5-15d), and the concave-convex cavity of Figure 5-15e.

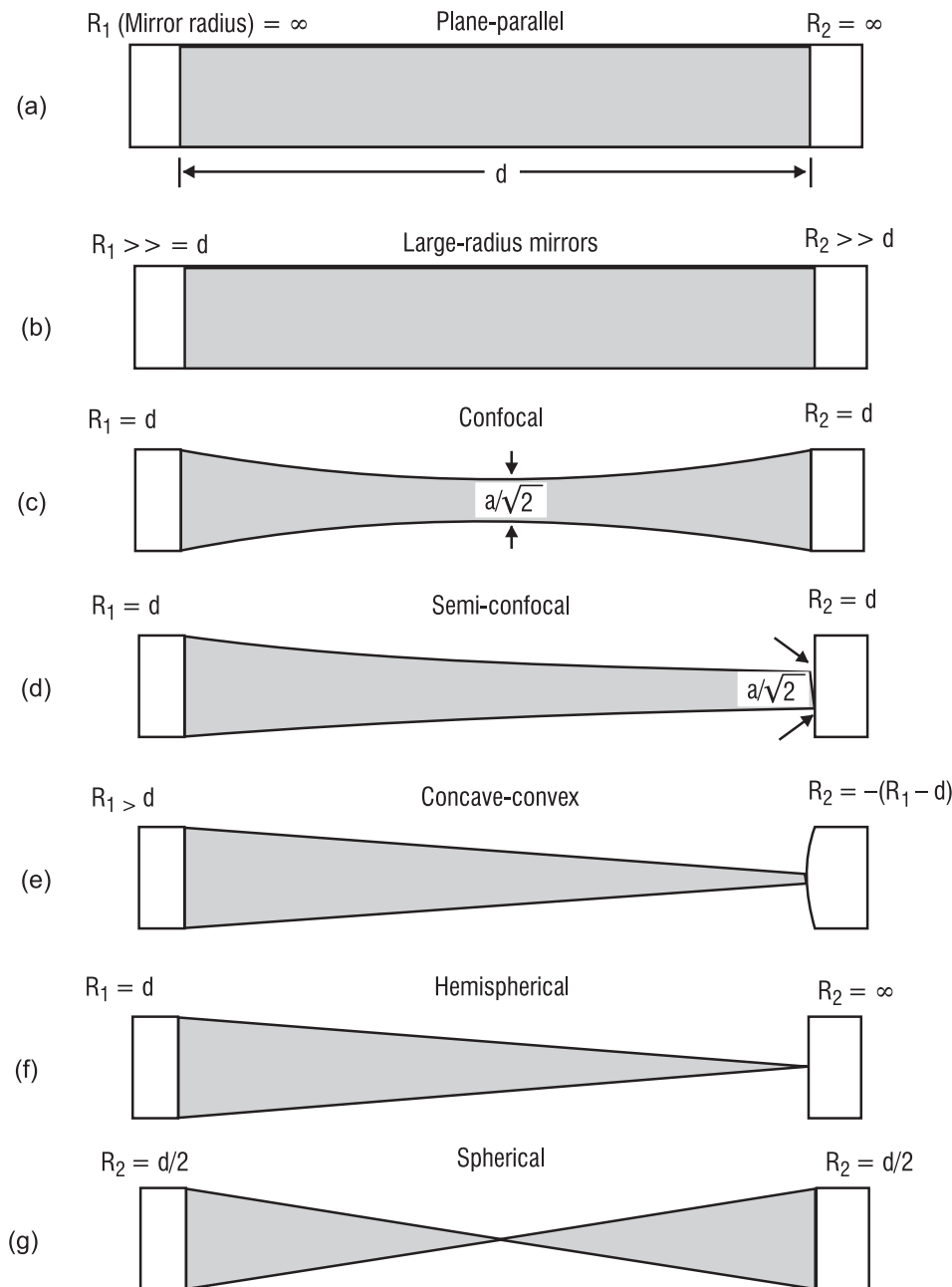


Figure 5-15 Various possible laser cavity configurations

With curved mirrors, a general way of defining a stable cavity is with one of the two following arrangements:

- (1) Both mirrors are concave. Either the center of curvature of each mirror lies beyond the other mirror or the center of curvature of each mirror lies between the other mirror and the center of curvature of the other mirror.
- (2) One mirror is convex and one is concave. The center of curvature of the concave mirror lies between the convex mirror and the center of curvature of the convex mirror.

A common cavity arrangement used to obtain the smallest output-beam diameter from the laser is the semi-confocal cavity as shown in Figure 5-15d. It consists of one curved and one flat mirror, with the flat mirror used as the output mirror. This is the equivalent of taking a stable cavity having two equal radii-of-curvature mirrors and replacing one of the curved mirrors with a flat mirror located at the halfway point between the two curved mirrors. For this arrangement the minimum beam waist w_0 occurs at the flat mirror.

Unstable resonators—A laser operating in the TEM_{00} mode typically has an inherently very narrow beam within the laser cavity. It thus cannot take advantage of a relatively wide gain region that might be available within the laser amplifier. Accessing the wide gain region could contribute significantly more energy to the laser output beam than is possible in the TEM_{00} mode. A special type of resonator cavity, referred to as an *unstable resonator*, has been developed to take advantage of such wide gain regions. A diagram of such a resonator is shown in Figure 5-16 with the gain region located between the concave and convex mirrors and having a transverse dimension the size of the diameter of the larger (concave) mirror. Such a

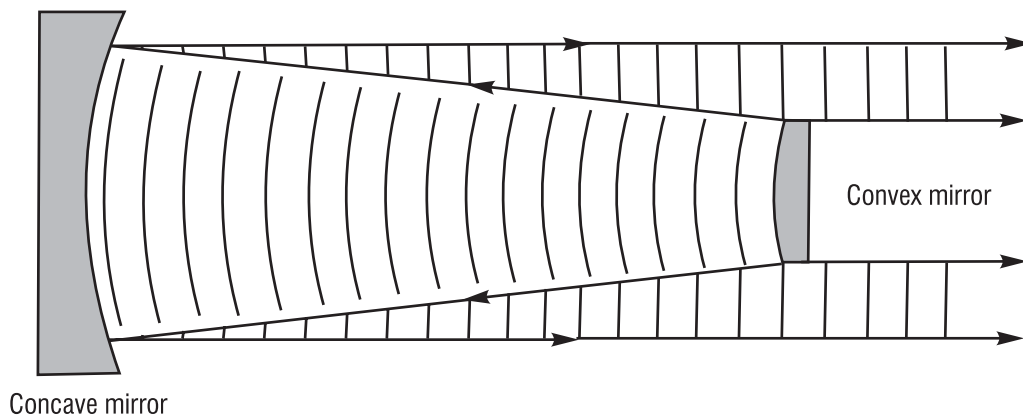


Figure 5-16 Diagram of an unstable resonator

resonator operates on a pulsed basis and requires a laser with a very high gain medium of the order of a factor of 10 or more per pass. In this resonator arrangement, the beam develops initially on the longitudinal axis between the two mirrors and then expands toward the convex mirror, reflecting and then further expanding to the larger concave mirror, and then passes out of the cavity in a straight line past the smaller (convex) mirror. The beam emerges from the laser in the shape of a doughnut, the center of the beam being obscured by the smaller mirror. The observed laser beam shape just beyond the mirror has the expected hole in the middle, however, farther away from the laser the hole fills in and a relatively high-quality, nearly Gaussian shaped beam evolves. This resonator configuration can produce significant amounts of pulsed energy that is extracted from the entire laser gain medium with a reasonably good beam quality.

Q-switching—Q-switching is a technique to produce short-duration high-energy laser pulses. It is achieved by pumping the laser gain medium with no cavity in place and then rapidly switching the mirror cavity into the system, thereby producing a giant pulse. Normally when a laser is “turned on” and pumping of the gain medium begins, as soon as the gain exceeds the mirror losses, laser energy begins to be extracted from the medium and an equilibrium balance is achieved with the expected power output from the laser. However, if the mirrors are not in

place at the ends of the gain medium, energy is not extracted and the medium will continue to accumulate population in the upper laser level for the duration of the lifetime of the level. Some laser gain media have upper laser levels with relatively long storage lifetimes before the electrons decay spontaneously to lower levels. These gain media are typically the solid-state lasers with storage times significantly longer than a microsecond, whereas most laser gain media have upper-level lifetimes of the order of one tenth to one thousandth of a microsecond. The longer storage lifetimes allow a very large population to build up in such levels when the laser medium is pumped with no mirrors present. In actual implementation of this process, the mirrors are situated at the appropriate locations in the cavity but an additional “shutter” is located between one of the mirrors and the end of the gain medium. This shutter can be an electro-optic shutter such as a Pockels cell or a Kerr cell, an acousto-optic shutter, or a saturable absorber. Opening the shutter, in effect, switches the mirrors into position for lasing to begin.

Mode locking—It is possible in a laser that has a very wide gain bandwidth to obtain thousands of longitudinal modes operating simultaneously. If those modes are all locked in phase so that they all oscillate upward or downward together, they will combine in a way that produces a chain of very short pulses separated in time by Δt_{sep} . This is the time it takes for a pulse to travel round trip within the cavity or

$$\Delta t_{sep} = \frac{1}{\Delta \nu} = \frac{2d}{c} \quad (5-23)$$

where $\Delta \nu$ is from Equation 5-15. This separation time is typically of the order of 10–20 nsec.

The duration of the ultrashort pulse Δt_p generated in such a process is given by the inverse of the frequency bandwidth spanned by all of the longitudinal modes. If N is the number of modes operating, the frequency bandwidth over which the modes are generated is just $N\Delta \nu$. Hence, Δt_p can be expressed as

$$\Delta t_p = \frac{1}{N\Delta \nu} \quad (5-24)$$

The maximum possible gain bandwidth $N\Delta \nu$ is the full gain bandwidth of the laser transition. Lasers with very wide gain bandwidths $\Delta \nu_{21}$, such as dye lasers and $\text{Ti:Al}_2\text{O}_3$ lasers (see Table 5-1), can produce the shortest pulses, light pulses as short as 5 femtoseconds (0.000000000000005 second or 5×10^{-15} sec) have been produced with this technique. While producing such short pulses is quite difficult, it is relatively easy nowadays to produce 100-fsec-duration pulses.

Ultrashort pulses, using mode-locking techniques, can be produced only if the laser is operating continuously (cw) or quasi cw (stable for periods of at least several cavity round-trip times as given by Equation 5-23). The principal method of locking the longitudinal modes in phase is to place a shutter within the laser cavity near one mirror. This shutter is opened when the short pulse arrives, allowing it to pass through to the mirror and back toward the amplifier, and then quickly closes until the next arrival of that pulse, when the shutter again opens. The shutter can be either an active shutter, such as an acousto-optic device driven with RF power, or a passive nonlinear optical shutter that opens when an intense pulse arrives. The active shutter is controlled with external electronics, whereas a passive shutter operates nonlinearly in a passive way, by rapidly opening for a high-intensity pulse but remaining closed for a low-intensity

pulse. This prevents the formation and growth of weaker pulses that could effectively compete with the original short pulse, and thereby lengthen it beyond its desired duration.

II. LASER PROPERTIES RELATED TO APPLICATIONS

Most of the properties described below are interrelated. Often, in carrying out a specific laser design to provide one of these properties, many of the others will also be obtained with no extra effort. Nevertheless, each effect will be described separately, since only one of the properties is typically sought for a specific application.

A. Collimation

Collimated light is light in which all of the light rays or waves are traveling in a specific direction and hence they are all parallel to each other. Lasers produce the most collimated light of any type of light source. Such collimated light is used for reference beams in construction, leveling and grading land, alignment of pipe such as sewer pipe, and sending light over long distances without suffering significant divergence, and in laser pointers. Producing the most collimated light, in other words the least divergent light, is determined by the cavity mirror properties including the radii of curvature of the mirrors and the separation between mirrors as indicated in Equations 5-17, 5-18, and 5-19. For the smallest beam divergence, w_0 must be large, as you can see from Equation 5-19. Also, the rays of the laser beam are the most parallel when the beam is at the location of the minimum beam waist w_0 as described in Equations 5-17, 5-20, and 5-21. This parallelism of the beam can be realized by using a semi-confocal resonator arrangement for the laser cavity as shown in Figure 5-15d and described after Equation 5-21, with the flat mirror as the output mirror.

B. Monochromaticity

Monochromaticity refers to how pure in color (frequency or wavelength) the laser beam is or, in other words, how narrow the laser beam frequency bandwidth is. Note that this is essentially $N\Delta\nu$ as described above in the mode-locking discussion if the laser is lasing with more than one longitudinal mode. If the laser is operating in a single longitudinal mode, as most solid-state and semiconductor lasers do, the actual laser linewidth can be significantly narrower, the width of a single longitudinal mode beam. For most applications requiring a single narrow wavelength, most lasers would normally provide a sufficiently narrow frequency output bandwidth, of the order of 10^9 – 10^{11} Hz. This would represent a bandwidth that is less than 0.1% of the frequency or wavelength of the beam itself (or even smaller in most instances). However, in some applications, such as in remote sensing or in establishing a new frequency standard, a much narrower linewidth is required. Linewidths of the order of 1 MHz (10^6 Hz) or less can be obtained by operating with a single longitudinal and single transverse mode (TEM_{00}). The narrowing is enhanced by choosing highly reflecting mirrors, constructing a very stable mirror cavity in conjunction with the amplifier by eliminating vibrations of the mirrors and other cavity elements, and providing temperature stability.

C. Coherence

Coherence refers to the how much *in step* or *in phase* various portions of a single laser beam are. The closeness in phase of various portions of the laser frequency bandwidth is referred to as temporal or longitudinal coherence. The closeness in phase of different spatial portions of the beam after the beam has propagated a certain distance is referred to as spatial or transverse coherence. This phased relationship determines how readily the various portions of the beam can interfere with each other, after the beam has propagated a specific distance, to produce such effects as diffraction (bending) of light and related applications such as holography. Typically, applications involve dividing the beam into two or more beams that travel different lengths or pathways and are then recombined. When they are recombined they will interfere with each other, producing the desired effect if those parts are still in phase (coherent); if they are no longer in phase, the effect will not occur. The coherence length is used to describe the beam propagation distance over which the beams stay in phase.

For longitudinal or temporal coherence, the coherence length ℓ_c is related to the wavelength λ and the total frequency bandwidth of the laser $\Delta\nu_L$ by

$$\ell_c = \frac{\lambda^2}{\Delta\lambda} \quad (5-25)$$

Note: $\Delta\lambda$ is the actual bandwidth of the laser beam given in wavelength units.

For transverse or spatial coherence, the transverse coherence length ℓ_t is related to the laser wavelength λ , the laser source diameter at its origin s , and the distance r the beam has propagated from its origin, by the following relationship.

$$\ell_t = \frac{r\lambda}{s} \quad (5-26)$$

D. Intensity and Radiance

Intensity or irradiance is the power of the laser beam divided by the cross-sectional area of the beam. It is thus typically given in watts per square centimeter (W/cm^2). It is a measure of the amount of energy that can be applied to a specific region within a given amount of time. It is one of the two most important parameters in using the laser for materials processing applications such as welding, cutting, heat treating, ablating, and drilling, or for laser surgery. The other important parameter is the laser wavelength, since the amount of absorption of all materials, including biological materials, is dependent upon the wavelength of the light. In some instances a deep penetration of the beam is desired, for example in doing processes that must be carried out quickly. In that situation, a laser wavelength in which the material has a relatively low absorption would be selected. Other applications might require a shallow penetration in order to control the quality of the edge to be left after the process is completed, such as in some surgical processes or in drilling very small holes. Thus, a wavelength region of high absorption would be chosen for the laser. A general rule is that absorption is very high for most materials at ultraviolet wavelengths and decreasing at longer wavelengths. However, this does not hold true for all materials or for all wavelengths. Many materials have high absorption peaks at specific

wavelengths that could occur anywhere in the spectrum, so one must be careful to obtain the absorption versus wavelength curves for the desired material before choosing the specific laser.

In all instances where high beam intensity is desired, the availability of the laser with sufficient intensity at that wavelength must be considered. Not all wavelengths with such intensity are possible. There are, however, lasers such as the CO₂ laser in the middle infrared (10.6 μm); the Q-switched Nd:YAG laser in the near infrared (1.06 μm) and frequency-doubled Nd:YAG to the green (530 nm); the copper vapor laser in the green (510 nm) and yellow (578 nm); and the ultraviolet excimer lasers including the XeF laser (351 nm), XeCl laser (308 nm), KrF laser (at 248 nm), ArF laser (193 nm), and F₂ laser (157 nm). For various aspects of materials processing, the beam must have an intensity in the range of 10^8 – 10^9 W/cm².

Radiance is a parameter that includes the beam intensity (W/cm²) and takes into account the beam divergence angle. The divergence angle is generally given in steradians (see Equation 5-2 and Figure 5-2), which is a three-dimensional angular measurement as opposed to the term degrees, which describes angular spread in only two dimensions (in a plane). As noted in Figure 5-2a, a complete sphere contains 4π steradians. Hence, radiance is given in units of watts per unit area per unit solid angle or (W/cm²-sr). Laser beam divergence is usually given in milliradians (mr) because of the very low divergence of most lasers. The approximate beam divergence in radians can be obtained by measuring the laser beam diameter at a specific, relatively long distance from the laser and dividing it by the square of the distance to where the measurement is made. To obtain the beam divergence in steradians, Equation 5-1 can be used.

Radiance becomes useful when a beam must be propagated over a reasonable distance before it is used or where the divergence can affect the focusing ability of the beam. Since most materials applications do not involve the tightest focusing possible for a given beam, intensity is usually the more important parameter.

E. Focusability

Many applications of lasers involve their ability to be focused to a very small spot size. Perhaps one of the most demanding applications is in focusing the small diode laser in a compact disk player. To store as much information as possible on each disk, that information must be included in the smallest grooves possible on the disk. The width of the grooves is determined by the ability of a laser beam to access a single groove without overlapping into adjacent grooves. Hence, the diameter of the spot size to which the laser beam can be focused becomes a very important parameter.

The smallest diameter that can be obtained with a focused laser, assuming that a single TEM₀₀ mode can be obtained from the laser, is approximately the dimension of the wavelength of the laser and is given by the following expression.

$$d_{\min} \cong \frac{4 \lambda (f/\#)}{\pi} \quad (5-27)$$

in which the $f/\#$ is the focal length of the lens used for the focusing divided by the useful diameter of the lens, the same notation as on camera lenses (see Module 1-3, *Basic Geometrical Optics*). If the laser beam is less than the actual lens diameter, the beam diameter is used instead of the lens diameter in determining the $f/\#$. In other words, a laser operating in the visible

spectral region with a wavelength of the order of 500 nm could be focused to a size of less than one hundredth the width of a human hair! The effective $f/\#$ focusing lens (ratio of focal length to laser beam diameter intercepted by the lens) must be of the order of unity to obtain such a small focus. Most lasers, however, can be focused relatively easily to spot diameters of the order of 0.1–0.2 mm. Extra care must be taken in terms of beam quality (mode quality) and lens focal length to obtain smaller spot diameters.

III. EXAMPLES OF COMMON LASERS

A. HeNe

The helium-neon laser (HeNe) was the first gas laser. The most widely used laser wavelength is the red wavelength (632.8 nm) with a cw power output ranging from 1 to 100 mW and laser lengths varying from 10 to 100 cm. HeNe lasers can also be operated at the 543.5-nm green wavelength and several infrared wavelengths. Initiation of a relatively low electrical current through a low-pressure gas discharge tube containing a mixture of helium and neon gases produces the population inversion. With this gas mixture, helium metastable atoms are first excited by electron collisions with helium ground-state atoms. This energy is then transferred to the desired neon excited energy levels, thereby producing the required population inversion with lower-lying helium energy levels.

B. Argon ion and Krypton ion

The argon ion laser and the krypton ion laser provide a wide range of visible and ultraviolet laser wavelengths. They produce cw output at wavelengths ranging from 275 to 686 nm and powers of up to tens of watts. Running a dc current through a long, narrow-bore plasma discharge tube filled with a low-pressure (0.1 torr) argon or krypton gas produces the population inversion. The argon atoms must be ionized to the first through third ionization stages to reach the appropriate energy levels for laser action. As a result, these lasers are relatively inefficient but still extremely useful for certain applications because of their short wavelengths.

C. HeCd

The helium cadmium laser (HeCd) operates continuously (cw) in the blue (441.6 nm) and ultraviolet (354 and 325 nm) portions of the spectrum with powers ranging from 20 to 200 mW and laser lengths of 40–100 cm. The population inversion in the amplifier region is produced by heating metallic cadmium to a vaporized state in a narrow-bore quartz discharge tube, mixing it with helium gas, and running an electrical discharge current of up to 100 mA through the tube. The excitation mechanisms include *Penning ionization* (helium metastable atoms colliding with neutral Cd atoms and exchanging energy), electron collisional ionization, and photoionization from strong, short-wavelength radiation originating within the helium atoms. The laser uses an effect known as *cataphoresis* to transport the cadmium atoms through the discharge and thereby provide the necessary uniform cadmium distribution within the gain region.

D. Copper vapor

The pulsed copper vapor laser (CVL) provides high average powers of up to 100 W at green (510 nm) and yellow (578 nm) wavelengths at very high repetition rates of up to 40 kHz and pulse durations of 10–50 nsec. The copper is heated within the laser gain region to temperatures of up to 1600°C in 2–10-cm-diameter ceramic tubes typically of 100–150 cm in length. The lasers are self-heated such that most of the energy provided by the discharge current provides heat to bring the plasma tube to the necessary temperature. Excitation occurs by electrons colliding with neutral copper atoms to excite them to the relevant laser-related energy levels.

E. CO₂

The carbon dioxide laser (CO₂), operating primarily in the middle infrared spectral region around 10.6 μm , is one of the world's most powerful lasers, producing cw powers of over 100 kW and pulsed energies of up to 10 kJ. It is also available in smaller versions with powers of up to 100 W from lasers the size of a shoe box. These lasers operate in a gas discharge in mixtures of helium, nitrogen, and CO₂ gases. Electron collisions with the nitrogen molecules within the discharge produce metastable energy levels. The energy contained in those levels is subsequently transferred by collisions to the CO₂ molecule, where the population inversion is produced. This is one of the most efficient lasers, with conversion of input electrical power to laser output power of up to 30%.

F. Excimer

Excimer lasers consist of mixtures of noble gas atoms such as argon, krypton, and xenon with reactive gases such as fluorine or chlorine operating in a special type of high-pressure gaseous discharge. They are therefore also known as rare gas-halide lasers. The actual laser species is an excited-state molecule containing a combination of the two types of atoms, such as ArF, KrF, XeF, and XeCl. The term excimer results from a contraction of the words “excited state dimer,” which indicates the excited-state nature of the lasing molecule. The lasers operate primarily in the ultraviolet spectral region with wavelengths at 193 nm (ArF), 248 nm (KrF), 308 nm (XeCl), and 351 nm (XeF). The laser output consists of 10–50-nsec pulses typically of 0.2 to 1 J/pulse at repetition rates of up to 1 kHz. These lasers are relatively efficient (1–5%) and are of a size that can fit on a desktop. The excitation occurs with electrons within the discharge colliding with and ionizing the rare gas atoms while at the same time dissociating the halide molecules into either F or Cl atoms to form negative halogen ions F[−] and Cl[−]. The negative halogen ions and positive rare gas ions readily combine in an excited state to form the laser species since they are of opposite charge and hence attract each other.

G. Organic dye

Dye lasers have the advantage of wide tunability in wavelength. When changing from one dye to another, the total wavelength region that can be covered ranges from 320 to 1500 nm. The gain medium for dye lasers consists of a solvent such as alcohol or water within which an organic dye is dissolved in a typical concentration of 1 part in ten thousand. If a diffraction grating or prism is used to replace one of the laser mirrors, the grating can be rotated to tune the laser wavelength over the spectrum of the dye. Each dye has a tunable gain bandwidth of

approximately 30–40 nm with a linewidth as narrow as 10 GHz or less. Dye lasers are available in either pulsed (up to 50–100 mJ/pulse) or cw (up to a few W) in tabletop systems that are pumped either by flashlamps or by other lasers such as frequency-doubled or -tripled YAG lasers or argon ion lasers. Most dye lasers are arranged to have the dye mixture circulated by a mechanical pump into the gain region from a much larger reservoir because the dyes degrade at a slow rate due to the excitation (optical pumping) process. Dye lasers, with their broad gain spectrum, are particularly attractive for producing ultrashort light pulses by a mode-locking process as described earlier. The shortest pulses ever generated are 5 thousandths of a trillionth of a second (5×10^{-15} sec).

H. Ruby

The ruby laser, which lases at the extreme red end of the visible spectrum at 694 nm, was the very first laser demonstrated. It consists of a crystalline sapphire (Al_2O_3) host material into which chromium atoms (the lasing species) are infused (doped) at a concentration of up to 0.05% by weight. The energy-level arrangement of the ruby laser, with the ground state as the lower laser energy level, makes for a very ineffective pumping process, in which very large amounts of pump light are required before gain is achieved. Therefore, this laser is not as efficient as other solid-state lasers and is not used much anymore.

I. Nd:YAG and Nd:glass

Neodymium is a very effective laser species when doped into either yttrium-aluminum-garnet (YAG) or glass host materials. The primary laser wavelength is at 1.06 μm for YAG and 1.05 μm for glass. The lasers are optically pumped either by flashlamps or by other lasers (especially GaAs semiconductor diode lasers). YAG is a very attractive host material for Nd because it has a high thermal conductivity and is a robust material, and hence the laser can produce high average-power output without having the crystal break. The Nd:YAG laser produces cw powers of up to 250 W and pulsed energies of up to 1 J/pulse.

J. $\text{Ti:Al}_2\text{O}_3$

Another class of solid-state lasers is the broad-bandwidth, tunable lasers. The most well-known and used laser in this category is the titanium-sapphire laser, $\text{Ti:Al}_2\text{O}_3$, consisting of titanium atoms doped into a sapphire (Al_2O_3) host at a concentration of up to 0.1% by weight. This laser operates over a wavelength ranging from 660 nm to 1,180 nm (1.08 μm), which gives it a gain bandwidth of 520 nm, the largest of any laser except perhaps the free-electron laser (a laser generated by an oscillating beam of high-energy electrons). This large bandwidth allows for large tunability as well as very short pulse production via mode-locking. It is also the one solid-state laser that has a much shorter upper-level lifetime than most other solid-state lasers, just under 4 μsec , which makes it difficult to pump with flashlamps and also difficult to Q-switch. Hence, this laser is typically pumped with other lasers such as the argon ion laser or the doubled Nd:YAG laser.

$\text{Ti:Al}_2\text{O}_3$ lasers are used in infrared spectroscopy of semiconductors, laser radar, rangefinders, and remote sensing and in medical applications such as photodynamic therapy. They are also

used to produce short pulses of X rays by focusing high-intensity pulses onto a solid target from which a tiny, high-temperature, high-density plasma is produced that radiates large quantities of X rays.

Other lasers in this broadband, tunable category include the alexandrite laser operating at 700 nm to 820 nm and the chrome-doped LiSAF laser, operating at 780 nm to 1010 nm.

K. Erbium fiber

Fiber lasers were first operated in Nd-doped glass fibers, but the fiber laser of major current interest is the erbium-doped fiber laser operating at 1.4 to 1.6 μm . This fiber laser can be spliced into a normal transmitting optical fiber, and, when optically pumped with a semiconductor laser at either 980 nm or 1.48 μm , it provides amplification to a communication signal that is being transmitted through the fiber. The pump light is fed into the fiber line—with a beam-combining device—thereby merging with the signal. This laser amplifier is especially useful in undersea optical fiber cables transmitting phone and data information between continents. More recently, it has also been used in local area networks and other shorter-distance communication applications. Several of these amplifiers must be installed equally spaced within a fiber when transmitting a message from America to Europe, for example. The amplifier bandwidth allows for a technique referred to as WDM (wavelength division multiplexing), which involves sending many separate signal beams at slightly different wavelengths through the same fiber, each of which is amplified by the single erbium-doped fiber amplifier. Hence, the data volume is increased significantly without having to install more fiber communication lines.

L. Semiconductor lasers (solid state)

Semiconductor lasers are small, very efficient lasers with dimensions of less than a millimeter (see semiconductor gain media). The wavelengths of commercial lasers range from approximately 600 nm in the red to 1.6 μm in the near infrared. Lasers in the blue and green are also under advanced development, but very few are available commercially. These lasers consist of small semiconductor crystals grown such that they can be cleaved in short segments of approximately 0.5 mm in length. The cleaving is done in a direction perpendicular to the laser axis, leaving a surface (facet) at each end of the gain medium that serves as a mirror. No reflective coatings are generally required; the cleaved surface itself serves to provide a reflectivity of the order of 35% or more from each of the two mirror surfaces, which is ample due to the very high gain or amplification present in the laser. Because of the short separation between mirrors, it is generally possible to obtain laser operation on only one longitudinal mode. Also, because of the short cavity length, the laser operates in a highly multi-transverse-mode output with a high angular divergence beam. To obtain single TEM_{00} -mode operation, it is necessary to coat the two end facets of the laser with an antireflection coating at the laser wavelength. Then an external mirror cavity can be installed with the appropriate mirror radii and reflectivity, as well as a suitable aperture on axis to restrict beam spread.

Semiconductor lasers are mass produced by depositing various layers of *p*- and *n*-doped material, insulating material, and metal contact layers by lithographic techniques. The most common semiconductor lasers are heterostructure lasers in which additional layers of different materials of similar electronic configurations are grown adjacent to the *pn* junction. This helps

confine the electrical current flow to the junction region in the direction perpendicular to the layers, thus minimizing the required current, energy deposition, and heat loading in the devices. The laser mode in the transverse direction is controlled either by (1) gain guiding, in which the gain is produced over a specific narrow lateral extent determined by fabrication techniques, or by (2) index guiding, in which the index of refraction in the transverse direction is varied to provide total internal reflection of the guided mode. Quantum-well lasers are semiconductor lasers that have a very thin gain region in the direction perpendicular to the layers, of the order of 5–10 nm. The laser gain bandwidth is significantly reduced due to a quantum effect. The gain coefficient remains the same, but the volume over which the energy deposition occurs is confined to an even smaller region. This significantly reduces the threshold current without significantly sacrificing laser output power. Because of their low threshold current and their low power consumption, quantum-well lasers are presently the most commonly used semiconductor lasers.

The largest applications of semiconductor lasers are in communication, in which the laser provides the signal, and in compact disk players, in which the laser is focused into the disk grooves and reflected to detect the digitally coded information. They are also used in high-speed printing systems and laser pointers and as pump sources for solid-state lasers (mainly Nd:YAG).

IV. OPERATION OF A HE NE LASER

The helium-neon laser is the most commonly used gas laser. Hence it serves as a good example in describing the detailed operation of a gas laser.

A. Laser structure

The He-Ne laser operates in a narrow-bore glass or quartz discharge tube with laser mirrors mounted at both ends of the tube. The tube can be anywhere from 10 to 100 cm in length, with most having lengths of 10–20 cm. The tube is filled with a low-pressure (2–3 torr) gas mixture of helium and neon in a ratio of approximately 5:1. At each end of the tube are electrical connections, referred to as electrodes, through which a high-voltage power supply is connected to the tube. The positive high voltage is connected to the anode electrode. The electrode at the other end of the tube, known as the cathode, is at ground potential. The anode electrode consists of a metal pin inserted through the glass so that one end extends into the gas region of the tube and the other is connected to the power supply. The cathode electrode typically consists of a cylindrical aluminum canister a few cm in diameter located inside a glass cylinder that concentrically surrounds the discharge tube, as shown in Figure 5-17. The large surface area of the canister cathode serves to minimize sputtering of the electrode material from the cathode when the discharge current is flowing in the tube. Excessive sputtering would contaminate the gas within the tube and reduce or eliminate laser operation.

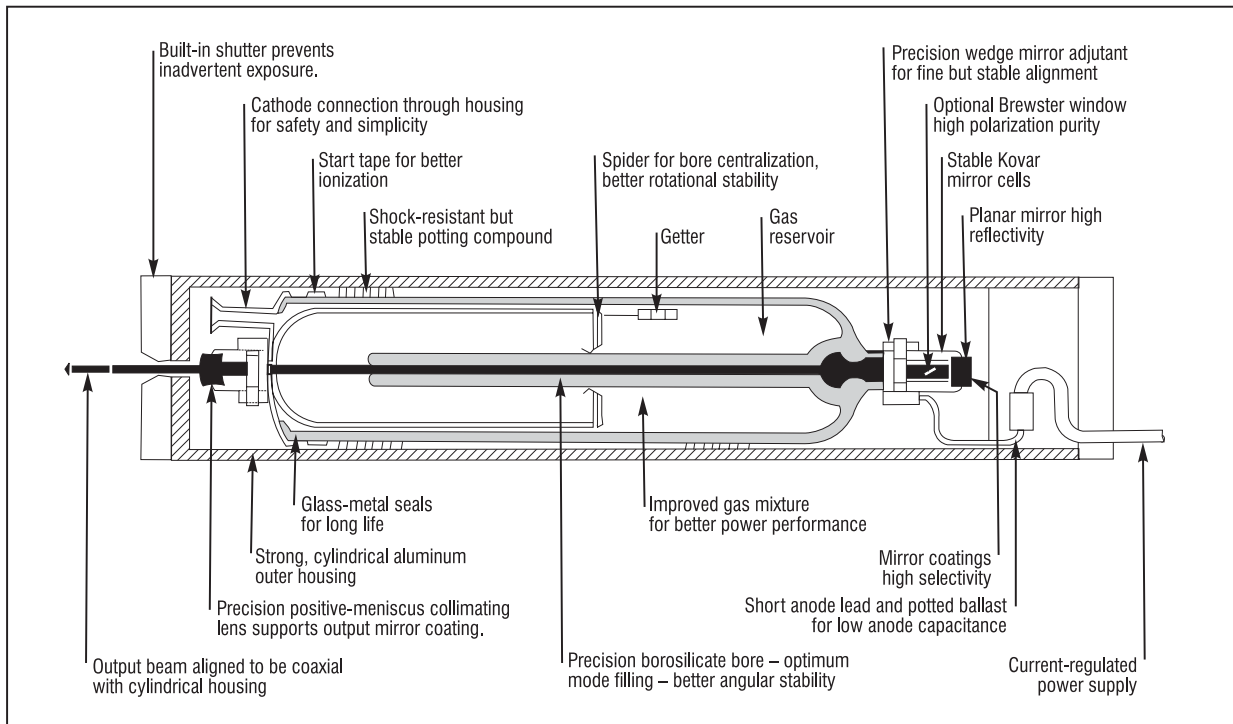


Figure 5-17 *HeNe laser diagram*

B. Laser operation

When a voltage is applied between the electrodes, a discharge current develops within the tube with the discharge electrons flowing from the cathode toward the anode. As the discharge current develops, the tube begins to emit light in a similar fashion to that of a neon sign. The light is produced by the electrons flowing within the gas and colliding with both helium and neon atoms, “knocking” them to excited energy levels from which they can radiate light via spontaneous decay. A partial diagram of the electronic energy levels of both neutral neon and neutral helium is shown in Figure 5-18.

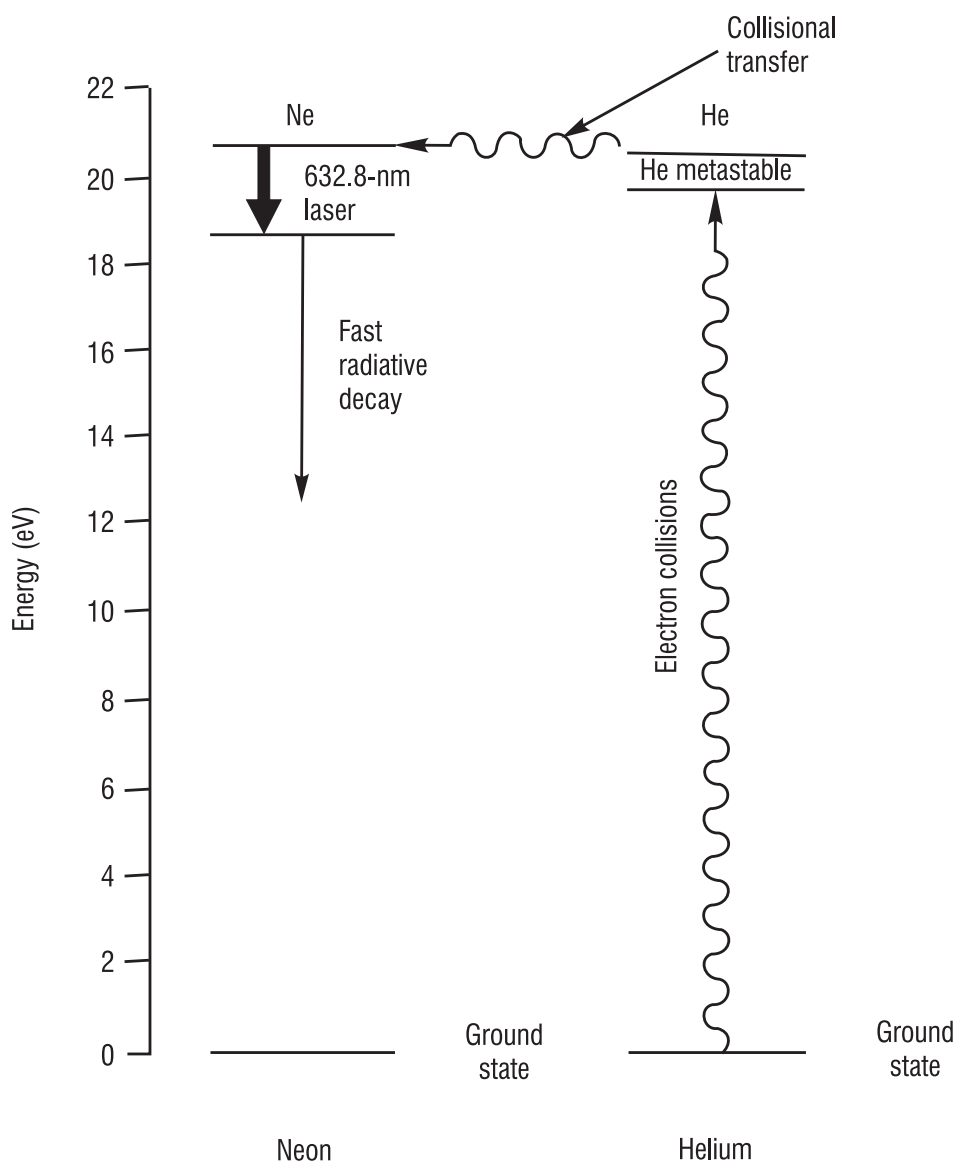


Figure 5-18 Relevant energy levels in helium and neon gases indicating energy-transfer processes

This diagram shows only one of the upper laser energy levels for neon so as not to complicate the description of the excitation process. All the laser transitions in the HeNe laser originate from within the neon atom. The presence of helium serves to provide the excitation or pumping of the neon laser levels in a two-step process. In helium, there are two excited energy levels known as metastable levels, which cannot emit light and thus do not easily decay. They serve as storage reservoirs since energy is collected in these levels. The energies of these metastable levels nearly exactly coincide with some of the excited energy levels of neon. Hence, when the helium metastable levels collide in the discharge with neon atoms in their ground state, the energy from the helium metastables is transferred to the neon atoms, effectively “pumping” them directly into their upper laser-energy levels. This establishes a population inversion in the neon atoms between those levels and lower-lying neon levels, and the necessary gain or amplification process is initiated.

C. Laser beam development

Some of the neon atoms that radiate from those higher-lying energy levels send light in the direction of the elongated bore of the discharge tube. That light is amplified as it travels down the tube as described by Equation 5-6. The light reaches a mirror and (if the mirror is appropriately aligned) is reflected back through the tube, where it is again amplified. It reaches the mirror at the other end and is again sent back through the tube to be amplified further. This process continues until a strong beam develops after about 500 passes, typically within a time of less than a millionth of a second. At that point the growth of the beam saturates, since only a finite number of atoms residing in the upper laser level can contribute to the population inversion and thus the gain or amplification within the tube. Thus, the laser beam very rapidly reaches an equilibrium situation in which the loss of a portion of the light transmitted through one of the mirrors (the output mirror) balances the energy pumped into the tube through the discharge current according to Equation 5-11 or 5-12. The light emitted from the partially transmitting mirror becomes the useful laser light. The optimum transmission can be determined by the use of Equation 5-13.

When the discharge current is turned off, the laser beam decays in approximately the same time duration as that of the buildup time (when the laser is turned on), or approximately 1 μsec . This is due to the time required for the laser beam to leak out of the output mirror while bouncing back and forth between the mirrors with no amplification available.

D. Longitudinal frequency modes

The laser-gain bandwidth, the frequency spectrum over which gain occurs at the laser wavelength, is approximately 2 GHz. When the discharge current is initiated and the laser beam begins reflecting back and forth between the mirrors, amplification occurs over this entire frequency width. However, certain longitudinal cavity modes develop that further refine the frequency characteristics of the beam in the form of the development of longitudinal modes. These modes typically consist of several discrete narrow-frequency beams of several to tens of MHz in width (see Figure 5-12) having equally spaced frequencies in the region of the laser wavelength of approximately 250–750 MHz (determined by the laser mirror separation). These separate longitudinal modes all travel in exactly the same direction, with their beam profile determined by the transverse or spatial mode characteristics described in the next paragraph.

E. Transverse spatial beam modes

The laser spatial or transverse mode characteristics are determined by the amount of gain in the amplifier as well as the size and location of the restricting aperture of the laser cavity. This aperture could be the laser discharge tube itself or a smaller aperture built into the cavity to limit the lateral spatial extent of the laser beam. When the limiting aperture is sufficiently small but still allows the laser beam to develop, the beam is said to be in its lowest-order mode or “single mode.” This is a very pure mode with a Gaussian-shaped beam profile. This very smooth, rounded transverse profile is the beam profile used for very precise beam propagation as well as focusing to a small spot size. When larger apertures are used, the lateral extent of the beam profile is enlarged. The resulting beam size also expands more rapidly to a larger diameter (greater beam divergence) than does the single-mode beam. More total power is available when the laser is operating in this multimode configuration. A laser with a beam consisting of only a

single spatial mode is referred to as a single-mode laser, and that specific mode is designated as a TEM_{00} mode, using the nomenclature developed for microwave beams.

F. Laser cavity properties

The amount of power provided by the laser is determined by the length and diameter of the discharge tube as well as the value of the discharge current. As the current is increased, more laser power is available up to a certain maximum current, depending on the discharge-tube bore diameter. Typically, the maximum current is up to a few tens of mA. The laser mirror cavity must be held very rigid to maintain the alignment of the beam within the laser gain region. If misalignment develops, the beam power will be reduced and the quality of the transverse modes will deteriorate. In small HeNe lasers the rigidity of the mirror alignment is maintained by the glass structure of the tube onto which the mirrors are cemented. In longer lasers, an external mechanical cavity is used to keep the mirrors rigid. This cavity often consists of rods made of Invar, a very sturdy metal with very low thermal expansion, that span the length of the laser and attach to the mirror mounts at each end. In longer lasers, Brewster-angle windows are attached to each end of the discharge tube so that the beam will not suffer loss or reduction in power when it passes through those windows. When Brewster-angle windows are used, the beam is automatically polarized in a single transverse direction. Consequently, if a polarizing filter is placed in the laser output beam and rotated, the beam will reach a maximum power in one location and will be reduced to essentially zero power when the filter is rotated ninety degrees with respect to the direction of maximum transmission.

G. HeNe laser wavelengths

The most often used laser transition in neon is the red-colored laser operating at 632.8 nm. It is the easiest visible HeNe laser wavelength to produce since it has the highest gain. Additional visible wavelengths are the green laser at 543.5 nm, the yellow laser at 594 nm, and the orange laser at 612 nm. In the near infrared there are several laser transitions in the 1.15- μm region and one at 1.523 μm , and a strong transition at 3.39 μm .

Laboratory

This laboratory is designed to demonstrate several different aspects of the operation of a laser.

Experiments to be carried out

- Alignment of laser mirror cavity
- Laser beam sensitivity to mirror alignment
- Operation of the laser at 632.8 nm with several different laser mirror configurations
- Single-mode and multimode (transverse or spatial mode) operation

LASERS

- Operation of the laser in the green at 543.5 nm
- Variation of laser power with discharge current

Equipment List

HeNe laser discharge tube with Brewster windows

Power supply with variable current capability to operate discharge tube

Stable optical rail or optical bench upon which to mount laser and related optical components

Mirror mounts with interchangeable mirror capability

Separate HeNe laser for alignment

Adjustable aperture to be placed on axis within the laser optical cavity

Rotating beam polarizer

Laser power meter

The following sets of laser mirrors with diameters of 0.5–1.0 inch:

- 2 high-reflecting laser mirrors at 632.8 nm with 1-m radius of curvature
- 1 99%-reflecting laser mirror with 1-m radius of curvature
- 1 99%-reflecting laser mirror at 632.8 nm with infinite radius of curvature (flat)
- 2 high-reflecting laser mirrors at 543.5 nm with 1-m radius of curvature

Procedure

1. Set up alignment laser on optical rail or optical bench and turn the laser on.
2. Place HeNe laser tube on optical rail and adjust the tube to be collinear and concentric with the beam of the alignment laser.
3. Attach high-voltage connections of power supply to laser electrodes. DO NOT TURN ON AT THIS TIME!
4. Place mirror mounts at each end of laser tube with mirror centers concentric with laser tube.
5. Place aperture at one end of laser tube between tube and mirror and adjust so that aperture is concentric with laser tube.
6. Install a high-reflecting 632.8-nm, 1-m radius-of-curvature laser mirror in the mirror mount farthest from the alignment laser and adjust the reflected spot to be centered with the alignment laser output beam.
7. Install the other high-reflecting mirror at the end of the laser cavity nearer the alignment laser and adjust until reflected spot is centered with alignment laser beam.
8. Turn laser power supply ON and initiate starter pulse from power supply (laser discharge should be visible at this time).

9. Observe laser mirrors and Brewster-angle windows to see if laser beam has developed. If laser beam is apparent, install power meter at one end of laser, beyond the laser mirror, to detect output beam.
10. Adjust both the horizontal and vertical position adjusters of both laser mirror mounts while observing laser power meter, to maximize laser power output by optimizing laser mirror alignment.
11. Remove mirror from laser mirror mount located closer to the alignment laser.
12. Install 99%-reflecting 1-m radius-of-curvature 632.8-nm mirror in the laser mirror mount and adjust the reflected spot from the alignment laser to be centered on the alignment laser beam. When the reflected spot is centered, a strong beam should become apparent emerging from the 99%-mirror end of the test laser. Install the laser power meter in front of the beam and optimize the mirror alignment by observing the laser power meter.
13. When mirror alignment is optimized, place a turning mirror in front of the beam to project the beam onto a screen approximately 2 or more meters from the laser mirror, to observe the quality of the beam. While observing the beam projected on the screen, reduce the diameter of the aperture within the laser cavity until the laser beam is no longer apparent. Then slowly increase the aperture diameter until the beam just begins to appear. Note the quality of the laser spot projected upon the screen. This is the single transverse laser mode known as the TEM_{00} laser mode. It is the highest-quality laser mode available in terms of focusing and propagation capabilities.
14. Slowly increase the aperture and see the beam enlarge on the screen. This shows the development of higher-order transverse laser modes. Commercial lasers with this cavity arrangement are known as multimode lasers.
15. Remove the 99% 1-m radius-of-curvature output mirror and replace it with the flat 99%-reflecting output mirror. Adjust the alignment until the laser beam appears. Note how much smaller the output beam diameter has now become. This is the semi-confocal cavity arrangement for many small commercial lasers to give a high-quality, small-diameter beam.
16. Remove the laser mirrors and install and align the 543.5-nm mirrors as was done in (6), (7), and (9) above. Observe the green laser beam within the laser cavity. Adjust the power on the meter to produce maximum power output.
17. Reinstall the mirrors of (6) and (12) above and adjust mirrors for maximum laser output.
18. Install polarizer in front of laser output beam with power meter located beyond polarizer. Rotate polarizer and graph the power output as a function of polarizer angle, rotating the polarizer through a full 360-degree rotation. Note the two maxima and two minima. Compare the maxima and minima with the orientation of the Brewster-angle windows.

Problem Exercises

1. You are provided with a 10-cm-long Nd:YAG laser rod to be used as a laser amplifier, with no mirrors attached at the ends of the rod. The rod is to be flashlamp-pumped in order to produce gain. You transmit a separate beam from another Nd:YAG laser through the amplifier rod and measure the emerging beam to increase by a factor of 10 from when it entered the rod. Calculate the value of g for the amplifier rod for such a situation, using Equation 5-9. If the rod is then placed within a mirror cavity and operated as a laser with the beam bouncing back and forth through the laser rod, determine how many passes the beam will have to make to reach the saturation intensity. (Hint: Remember that $gL_T = 12$. Hence, if we let $L_T = ML$ where M is the number of passes and L is the length of the gain medium, we can solve for M .)
2. Estimate the beam waist of a laser at a distance of 25 cm from the output mirror. This is a HeNe laser operating at 632.8 nm with a flat mirror at the output end and a curved mirror of radius 50 cm at the other end. Separation between the two mirrors is 25 cm. How would this compare to the beam waist at the location of the curved laser mirror?
3. Murray positions a ruby laser rod halfway between two identical mirrors of 100-cm radius of curvature with a mirror separation of 20 cm. The laser is operated at 694.3-nm wavelength. It has a restrictive aperture that forces it to operate in a single transverse TEM₀₀ Gaussian mode. Murray aims the beam toward a satellite installed in a geosynchronous orbit 20,000 miles above the earth. Calculate the diameter of the beam when it reaches the satellite.
4. Lin adjusts a mode-locked titanium sapphire laser to produce very short mode-locked pulses. She measures the pulses to have a duration of 100 fsec. A distance of 200 cm separates the cavity end mirrors. From this information, determine the spacing in frequency between the longitudinal modes and estimate how many modes must be lasing to produce the 100-fsec pulses. Also determine the spacing in time between successive pulses.
5. The single-pass gain of an argon ion laser is 100% per pass such that the beam will double in intensity after passing once through the amplifier. The single-pass losses, not including the mirror-transmission losses but including scattering losses, are 0.2% per pass. If one mirror is a highly reflecting mirror and the other is used for an output mirror, determine the optimum reflectivity of the output mirror for coupling out the maximum power.
6. With the information obtained in Exercise 5, estimate the output intensity of the argon laser when using the optimum-output mirror determined above.
7. Wai-Min measures a 632.8-nm HeNe laser to have a spot diameter of 1 mm as it emerges from the laser. When the beam is propagated a certain distance away, he notices that it has expanded to a size of 1-cm diameter. He is interested in generating a hologram. Using Wai-Min's information, calculate the maximum distance the beam can propagate and still have a coherent interaction. (Hint: The outer edges of the beam, separated by the 1-cm diameter, are the portions of the beam farthest from each other. They are thus the most susceptible to becoming out of phase with each other.)
8. Eric carries out an investigation to use a pulsed ArF excimer laser operating at 193 nm for drilling 500-nm-diameter holes in ultrathin steel plates. He sets the laser up to operate in

the TEM₀₀ mode to obtain the best focus and realizes that he must order a lens for this purpose. Eric's laser beam has a transverse diameter of 2 cm, and he wants to obtain a 500-nm beam-spot size. What is the focal length of the lens that Eric needs?

9. Maria sets up a HeCd laser (operating at 441.6 nm) with two identical mirrors of 99% reflectivity and 2-m radii of curvature. What threshold gain coefficient would Maria have to achieve in the electrical discharge mixture of helium and cadmium vapor for a gain length of 30 cm? For this calculation, assume that the window losses and other scattering losses are negligible.
10. Brian sets up a laser mirror cavity to operate a carbon dioxide laser at a wavelength of 10.6 μm . He uses two identical mirrors of 2.0-m radius of curvature. Brian is restricted to placing the mirrors 1.0 m apart and to having the mirror curvatures exactly match the beam curvature in the cavity. Determine the minimum beam waist w_0 produced with this arrangement.

Student Project

Locate a flashlight and two different types of lasers, preferably a helium-neon laser and a laser pointer (a semiconductor diode laser), each of which is designed to be eye-safe. Install the three light sources in a rigid mount and project them into a relatively long room or hallway to measure the beam divergence. Measure the beam size and compute the angular divergence $\Delta\Omega$ in *steradians* for all three according to Equation 5-1. The helium-neon laser will most likely be operating in the TEM₀₀ mode and will thus have the smallest beam divergence. Also calculate the half-width angular divergence in *radians* by dividing the beam radius by the distance from the laser to where the spot size is measured. Knowing this angular divergence and wavelength, calculate the minimum beam waist of the HeNe laser from Equation 5-19.

Bibliography

1. Silfvast, William T. "Lasers," *Encyclopedia of Physical Science and Technology*, Volume 7, Academic Press, 1987. A nonmathematical overview of lasers for a general audience.
2. Silfvast, William T. "Lasers," *Handbook of Optics*, 2nd Edition, Edited by Mike Bass, McGraw Hill and Optical Society of America, 1995. A brief overview of lasers with some algebraic mathematical equations and formulas.
3. Silfvast, William T. *Laser Fundamentals*. New York: Cambridge University Press, 1996. A comprehensive, calculus-based treatment of lasers suitable for a senior-level or first-year graduate-college engineering or science student.
4. Siegman, A. E. *Lasers*. Mill Valley, California: Interscience Publishers, 1986. A very comprehensive, advanced treatment of lasers.

Optical Detectors and Human Vision

Jack Ready

***Consultant, formerly with Honeywell Technology Center
Edina, Minnesota***

Many photonics applications require the use of optical radiation detectors. Examples are optical radar, monitoring of laser power levels for materials processing, and laser metrology. Different types of optical detectors are available, covering the ultraviolet, visible, and infrared portions of the electromagnetic spectrum. Optical detectors convert incoming optical energy into electrical signals. The two main types of optical detectors are photon detectors and thermal detectors. Photon detectors produce one electron for each incoming photon of optical energy. The electron is then detected by the electronic circuitry. Thermal detectors convert the optical energy to heat energy, which then generates an electrical signal.

The detector circuit often employs a bias voltage and there is a load resistor in series with the detector. The incident light changes the characteristics of the detector and changes the current flowing in the circuit. The output signal is then the change in voltage drop across the load resistor.

In this module, we will describe some common optical detectors and their important characteristics. We shall not attempt to cover the entire field of light detection, which is very broad. Instead, we shall emphasize those detectors that are most commonly encountered in photonics applications.

Prerequisites

You should have the ability to solve algebraic equations, should understand basic trigonometric functions, and should have knowledge of laser safety procedures.

The following modules should have been completed previously or should be studied concurrently with this module:

Module 1-1 *Nature and Properties of Light*

Module 1-2 *Light Sources and Laser Safety*

Module 1-3 *Basic Geometric Optics*

Module 1-4 *Basic Physical Optics*

Module 1-5 *Lasers*

Objectives

When you finish this module, you will be able to:

1. Define important detector response characteristics, including responsivity, noise equivalent power, quantum efficiency, detectivity, rise time, and cutoff wavelength for a photon detector.
2. Define sources of detector noise, including shot noise, Johnson noise, $1/f$ noise, and photon noise. Explain methods employed to reduce the effect of these noise sources.
3. Describe and explain the operation of important types of photodetectors, including photon detectors, thermal detectors, photoemissive detectors, photoconductive detectors, photovoltaic detectors, and photomultiplier detectors. Describe the spectral response of each type.
4. Draw and explain a typical circuit for a photovoltaic detector.
5. Draw and explain a typical circuit for a photoconductive detector.
6. Describe important concepts related to human vision, including structure of the eye, the formation of images by the eye, and common defects of vision.
7. Given the necessary information, calculate the noise equivalent power of a detector.
8. Given the necessary information, calculate the detectivity of a detector.
9. Given the necessary information, calculate the quantum efficiency of a detector.
10. Given the necessary information, calculate the power reaching a detector after a laser beam is reflected from a Lambertian reflector.
11. Fabricate a circuit for operation of a photodiode. Use the circuit for detection of light in both photoconductive and photovoltaic modes of operation.
12. Determine the relative response of the detector circuit as a function of wavelength for several wavelengths in the visible spectrum.

Workplace Scenario

Maria is a senior photonics technician who uses detectors for infrared, visible, and ultraviolet light in many applications. Maria works in the advanced research laboratory of a large industrial company and has many years of photonics experience. She employs detectors for monitoring the output of lasers as she adjusts their mirrors. Working under the direction of a scientist, Maria has assembled equipment containing detectors for detecting the return signal in environmental monitoring applications and for controlling the progress of materials-processing applications. Her specific duties have included calibrating, cleaning, maintaining, testing, aligning, mounting, installing, operating, and demonstrating detectors for light.

Opening Demonstration**Materials**

Helium-neon laser (few-milliwatt output)

Power meter (Coherent Fieldmaster LM-2 or equivalent)

Procedures

In this introductory demonstration, you will use a power meter to measure the output of a HeNe laser.

Assemble the equipment as shown in Figure 6-1. Turn on the HeNe laser, making sure to observe proper laser safety procedures. Direct the beam into the aperture of the power meter head. Observe the reading on the display.

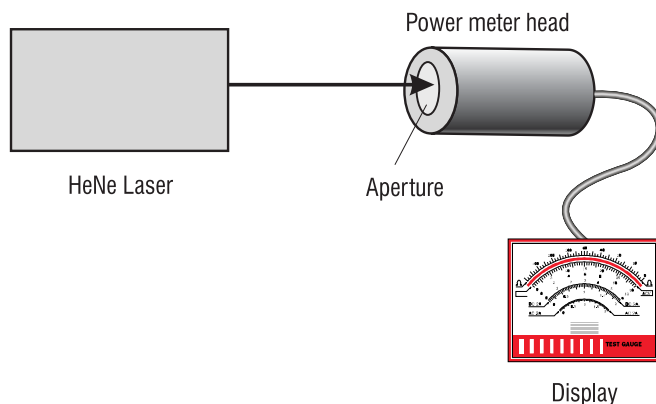


Figure 6-1 Arrangement of equipment for introductory laboratory measurement of the output power of a HeNe laser

Adjust the control for the output of the laser. Observe the changes in the power meter reading as the power of the laser is varied.

Basic Concepts

I. BASIC INFORMATION ON LIGHT DETECTORS

When light strikes special types of materials, a voltage may be generated, a change in electrical resistance may occur, or electrons may be ejected from the material surface. As long as the light is present, the condition continues. It ceases when the light is turned off. Any of the above conditions may be used to change the flow of current or the voltage in an external circuit and thus may be used to monitor the presence of the light and to measure its intensity.

A. Role of an optical detector

Many photonics applications require the use of optical detectors to measure optical power or energy. In laser-based fiber optic communication, a detector is employed in the receiver. In laser materials processing, a detector monitors the laser output to ensure reproducible conditions. In applications involving interferometry, detectors are used to measure the position and motion of interference fringes. In most applications of light, one uses an optical detector to measure the output of the laser or other light source. Thus, good optical detectors for measuring optical power and energy are essential in most applications of photonics technology.

Optical detectors respond to the power in the optical beam, which is proportional to the square of the electric field associated with the light wave. Optical detectors therefore are called “square-law detectors.” This is in contrast to the case of microwave detectors, which can measure the electric field intensity directly. All the optical detectors that we will describe have square-law responses.

Detection and measurement of optical and infrared radiation is a well-established area of technology. This technology has been applied to photonics applications. Detectors particularly suitable for use with lasers have been developed. Some detectors are packaged in the format of power or energy meters. Such a device is a complete system for measuring the output of a specific class of lasers, and includes a detector, housing, amplification if necessary, and a readout device.

B. Types of Optical Detectors

Optical detectors are usually divided into two broad classes: photon detectors and thermal detectors. In photon detectors, quanta of light energy interact with electrons in the detector material and generate free electrons. To produce free electrons, the quanta must have sufficient energy to free an electron from its atomic binding forces. The wavelength response of photon detectors shows a long-wavelength cutoff. If the wavelength is longer than the cutoff wavelength, the photon energy is too small to produce a free electron and the response of the photon detector drops to zero.

OPTICAL DETECTORS AND HUMAN VISION

Thermal detectors respond to the heat energy delivered by light. These detectors use some temperature-dependent effect, like a change of electrical resistance. Because thermal detectors rely on only the total amount of heat energy reaching the detector, their response is independent of wavelength.

The output of photon detectors and thermal detectors as a function of wavelength is shown schematically in Figure 6-2. This figure shows the typical spectral dependence of the output of photon detectors, which increases with increasing wavelength at wavelengths shorter than the cutoff wavelength. At that point, the response drops rapidly to zero. The figure also shows how the output of thermal detectors is independent of wavelength, and extends to longer wavelengths than the response of photon detectors.

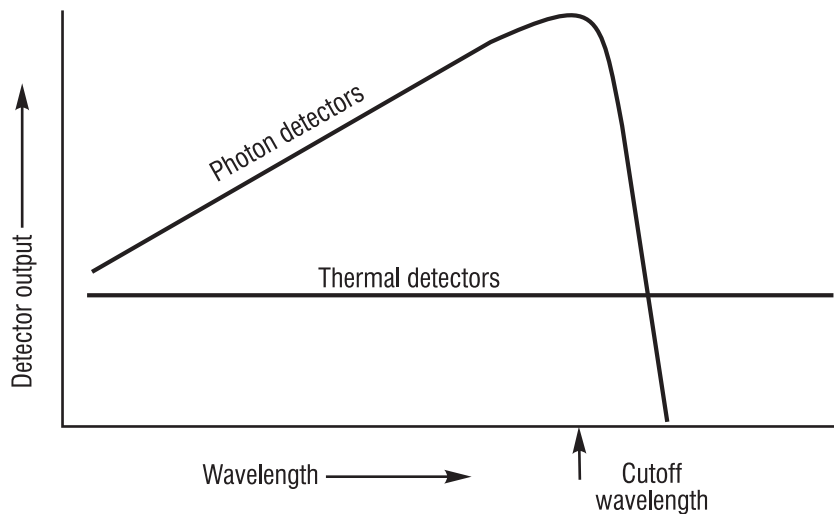


Figure 6-2 Schematic drawing of the relative output per unit input for photon detectors and thermal detectors as a function of wavelength. The position of the long-wavelength cutoff for photon detectors is indicated.

Figure 6-2 is intended to show only the relative shape of the output curves for these two types of detectors and is not intended to show quantitative values. Quantitative values will be presented in later figures for some specific detectors.

Photon detectors may be further subdivided according to the physical effect that produces the detector response. Some important classes of photon detectors are listed below.

- **Photoconductive.** The incoming light produces free electrons which can carry electrical current so that the electrical conductivity of the detector material changes as a function of the intensity of the incident light. Photoconductive detectors are fabricated from semiconductor materials such as silicon.
- **Photovoltaic.** Such a detector contains a junction in a semiconductor material between a region where the conductivity is due to electrons and a region where the conductivity is due to holes (a so-called pn junction). A voltage is generated when optical energy strikes the device.

- *Photoemissive*. These detectors are based on the photoelectric effect, in which incident photons release electrons from the surface of the detector material. The free electrons are then collected in an external circuit.

Photoconductive and photovoltaic detectors are commonly used in circuits in which there is a load resistance in series with the detector. The output is read as a change in the voltage drop across the resistor.

We shall discuss each of these effects in more detail later.

C. Detector characteristics

The performance of optical detectors is commonly characterized by a number of different parameters. It is important to define these parameters, sometimes called *figures of merit*, because manufacturers usually describe the performance of their detectors in these terms.

The figures of merit were developed to describe the performance of detectors responding to a small signal in the presence of noise. Thus, some of the figures of merit may not be highly relevant to the detection of laser light. For many laser applications, like laser metalworking, there is no question of detection of a small signal in a background of noise. The laser signal is far larger than any noise source that may be present. In other photonics applications, like laser communication, infrared thermal imaging systems, and detection of backscattered light in laser remote sensing, the signals are small and noise considerations are important.

Responsivity

The first term that we will define is *responsivity*. This is the *detector output per unit of input power*. The units of responsivity are either amperes/watt (alternatively milliamperes/milliwatt or microamperes/microwatt, which are numerically the same) or volts/watt, depending on whether the output is an electric current or a voltage.

The responsivity is an important parameter that is *usually specified by the manufacturer*. Knowledge of the responsivity allows the user to determine how much detector signal will be available for a specific application.

Noise Equivalent Power

A second figure of merit, which depends on noise characteristics, is the *noise equivalent power (NEP)*. This is defined as the *optical power that produces a signal voltage (or current) equal to the noise voltage (or current) of the detector*. The noise is dependent on the bandwidth of the measurement, so that bandwidth must be specified. Frequently it is taken as 1 Hz. The equation defining *NEP* is

$$NEP = \frac{H A V_N}{V_S (\Delta f)^{1/2}} \quad (6-1)$$

where H is the irradiance incident on the detector of area A ,
 V_N is the root mean square noise voltage within the measurement bandwidth Δf , and
 V_S is the root mean square signal voltage.

The *NEP* has units of $\text{watts}/(\text{Hz})^{1/2}$, usually called “watts per root hertz.” From the definition, it is apparent that the lower the value of the *NEP*, the better are the characteristics of the detector for detecting a small signal in the presence of noise.

Example 1

The noise equivalent power of a detector with area 1 cm^2 is measured to be $2 \times 10^{-8} \text{ watts}/(\text{Hz})^{1/2}$ with a bandwidth of 1 Hz. What power is incident on the detector if the ratio of the noise voltage to the signal voltage is 10^{-6} ?

Solution:

According to Equation 6-1, the irradiance H at the detector must be

$$H = \frac{NEP}{A \left(\frac{V_N}{V_S} \right) (\Delta f)^{1/2}} = 2 \times 10^{-8} / \{ (1) \times (10^{-6}) \times (1) \} = 0.02 \text{ W/cm}^2$$

Because the area of the detector was 1 cm^2 , the power is 0.02 W.

Detectivity

The *NEP* of a detector is dependent on the area of the detector. To provide a figure of merit that is dependent on the intrinsic properties of the detector, not on how large it happens to be, a term called detectivity is defined. Detectivity is represented by the symbol D^* , which is pronounced as D-star. It is defined as the *square root of the detector area per unit value of NEP*.

$$D^* = A^{1/2}/NEP \quad (6-2)$$

Since many detectors have *NEP* proportional to the square root of their areas, D^* is independent of the area of the detector. The detectivity thus gives a measure of the intrinsic quality of the detector material itself.

When a value of D^* for an optical detector is measured, it is usually measured in a system in which the incident light is modulated or chopped at a frequency f so as to produce an AC signal, which is then amplified with an amplification bandwidth Δf . These quantities must also be specified. The dependence of D^* on the wavelength λ , the frequency f at which the measurement is made, and the bandwidth Δf are expressed by the notation $D^*(\lambda, f, \Delta f)$. The reference bandwidth is often 1 Hertz. The units of $D^*(\lambda, f, \Delta f)$ are $\text{cm-Hz}^{1/2}/\text{watt}$. A high value of $D^*(\lambda, f, \Delta f)$ means that the detector is suitable for detecting weak signals in the presence of noise. Later, in the discussion of noise, we will describe the effect of modulation frequency and bandwidth on the noise characteristics.

Example 2

A detector has a noise equivalent power of 3×10^{-9} watts/(Hz)^{1/2} and an area of 0.4 cm². What is its value of D^* ?

Solution:

According to equation 6-2, D^* is $(0.4 \text{ cm}^2)^{1/2}/3 \times 10^{-9} \text{ watts/(Hz)}^{1/2}$

$$= 0.632 \text{ cm} \times 0.333 \times 10^9 \text{ Hz}^{1/2}/\text{watt}$$

$$= 2.11 \times 10^8 \text{ cm-Hz}^{1/2}/\text{watt}$$

Quantum efficiency

Another common figure of merit for optical detectors is the *quantum efficiency*. Quantum efficiency is defined as the *ratio of countable events produced by photons incident on the detector to the number of incident photons*. If the detector is a photoemissive detector that emits free electrons from its surface when light strikes it, the quantum efficiency is the number of free electrons divided by the number of incident photons. If the detector is a semiconductor pn-junction device, in which hole-electron pairs are produced, the quantum efficiency is the number of hole-electron pairs divided by the number of incident photons. If, over a period of time, 100,000 photons are incident on the detector and 10,000 hole-electron pairs are produced, the quantum efficiency is 10%.

The quantum efficiency is basically another way of expressing the *effectiveness of the incident optical energy for producing an output of electrical current*. The quantum efficiency Q (in percent) may be related to the responsivity by the equation:

$$Q = 100 \times R_d \times (1.2395/\lambda) \quad (6-3)$$

where R_d is the responsivity (in amperes per watt) of the detector at wavelength λ (in micrometers).

Example 3

A detector has a quantum efficiency of 10% at a wavelength of 500 nm. At a wavelength of 750 nm, the responsivity is twice the responsivity at 500 nm. What is the quantum efficiency at 750 nm?

Solution:

From Equation 6-3, we see that the increase in responsivity from 500 to 750 nm will increase the quantum efficiency Q by a factor of 2, but the increase in wavelength will decrease the quantum efficiency Q by a factor of 2/3, so that the net change in quantum efficiency will be an overall increase by a factor of 4/3, from 10% to 13.33%.

Detector response time

Another useful detector characteristic is the *speed of the detector response* to changes in light intensity. If a light source is instantaneously turned on and irradiates an optical detector, it takes a finite time for current to appear at the output of the device and for the current to reach a steady value. If the source is turned off instantaneously, it takes a finite time for the current to decay back to zero. The term *response time* refers to the *time it takes the detector current to rise to a value equal to 63.2% of the steady-state value* which is reached after a relatively long period of time. (This value is numerically equal to $1 - 1/e$, where e is the base of the natural logarithm system.) The *recovery time* is the *time it takes for the photocurrent to fall to 36.8% of the steady-state value* when the light is turned off instantaneously.

Because optical detectors often are used for detection of fast pulses, another important term, called *rise time*, is often used to describe the speed of the detector response. *Rise time* is defined as the *time difference between the point at which the detector has reached 10% of its peak output and the point at which it has reached 90% of its peak response*, when it is irradiated by a very short pulse of light. The *fall time* is defined as the *time between the 90% point and the 10% point on the trailing edge of the pulse waveform*. This is also called the decay time. We should note that the fall time may be different numerically from the rise time.

Of course, light sources are not turned on or off instantaneously. To make accurate measurements of rise time and fall time, the source used for the measurement should have a rise time much less than the rise time of the detector being tested. Generally, one should use a source whose rise time is less than 10% of the rise time of the detector being tested.

The intrinsic response time of an optical detector arises from the transit time of photogenerated charge carriers within the detector material and from the inherent capacitance and resistance associated with the device. The measured value of response time is also affected by the value of the load resistance that is used with the detector, and may be longer than the inherent response time. There is a tradeoff in the selection of a load resistance between speed of response and high sensitivity. It is not possible to achieve both simultaneously. Fast response requires a low load resistance (generally 50 ohms or less), whereas high sensitivity requires a high value of load resistance. It is also important to keep any capacitance associated with the circuitry, the electrical cables, and the display devices as low as possible. This will help keep the RC (resistance $\times$ capacitance) time constant low. Manufacturers often quote nominal values for the rise times of their detectors. These should be interpreted as minimum values, which may be achieved only with careful circuit design and avoidance of excess capacitance and resistance in the circuitry.

Linearity

Yet another important characteristic of optical detectors is their *linearity*. Detectors are characterized by a response in which the *output is linear with incident intensity*. The response may be linear over a broad range, perhaps many orders of magnitude. If the output of the detector is plotted versus the input power, there should be no change in the slope of the curve. Noise will determine the lowest level of incident light that is detectable. The upper limit of the input/output linearity is determined by the maximum current that the detector can produce without becoming saturated. *Saturation* is a condition in which there is *no further increase in detector response as the input light intensity is increased*. When the detector becomes saturated,

one can no longer rely on its output to represent the input faithfully. The user should ensure that the detector is operating in the range in which it is linear.

Manufacturers of optical detectors often specify maximum allowable continuous light level. Light levels in excess of this maximum may cause saturation, hysteresis effects, and irreversible damage to the detectors. If the light occurs in the form of a very short pulse, it may be possible to exceed the continuous rating by some factor (perhaps as much as 10 times) without damage or noticeable changes in linearity.

Spectral response

The *spectral response* defines how the *performance of a detector (responsivity or detectivity) varies with wavelength*. The spectral response is defined by curves such as shown in Figure 6-2, which presents generalized curves showing relative spectral response as a function of wavelength for photon detectors and thermal detectors. The exact shape of the spectral response and the numerical values depend on the detector type and the material from which the detector is fabricated. Many different types of detectors are available, with responses maximized in the ultraviolet, visible, or infrared spectral regions. Again, the manufacturer usually specifies the spectral response curve. One should choose a detector that responds well in the spectral region of importance for the particular application.

D. Noise considerations

Noise in optical detectors is a complex subject. In this module we will do no more than present some of the most basic ideas. *Noise is defined as any undesired signal*. It masks the signal that is to be detected.

Noise can be external and internal. *External noise* involves disturbances that appear in the detection system because of factors outside the system. Examples of external noise could be pickup of hum induced by 60-Hz electrical power lines and static caused by electrical storms. *Internal noise* includes all noise generated within the detection system itself. Every electronic device has internal sources of noise, which represent an ever-present limit to the smallest signal that may be detected by the system.

Noise cannot be described in the same manner as usual electric currents or voltages. We think of currents or voltages as functions of time, such as constant direct currents or sine-wave alternating voltages. The noise output of an electrical circuit as a function of time is completely erratic. We cannot predict what the output will be at any instant. There will be no indication of regularity in the waveform. The output is said to be random.

Now we will consider some of the sources of noise encountered in optical detector applications. A complete description of all types of noise would be very long. We will describe four noise sources often encountered in connection with optical detectors.

- Johnson noise
- Shot noise
- $1/f$ noise
- Photon noise

Johnson noise

Johnson noise is generated by thermal fluctuations in conducting materials. It is sometimes called thermal noise. *It results from the random motion of electrons in a conductor.* The electrons are in constant motion, colliding with each other and with the atoms of the material. Each motion of an electron between collisions represents a tiny current. The sum of all these currents taken over a long period of time is zero, but their random fluctuations over short intervals constitute Johnson noise.

To reduce the magnitude of Johnson noise, one may cool the system, especially the load resistor. One should reduce the value of the load resistance, although this is done at the price of reducing the available signal. One should keep the bandwidth of the amplification small; one Hz is a commonly employed value.

Shot noise

The term *shot noise* is derived from *fluctuations in the stream of electrons in a vacuum tube*. These variations create noise because of the random fluctuations in the arrival of electrons at the anode. The shot noise name arises from the similarity to the noise of a hail of shots striking a target.

In semiconductors, the major source of shot noise is random variations in the rate at which charge carriers are generated and recombine. This noise, called generation-recombination or *gr noise*, is the semiconductor manifestation of shot noise.

Shot noise may be minimized by keeping any DC component to the current small, especially the dark current, and by keeping the bandwidth of the amplification system small.

1/f noise

The term *1/f noise* (pronounced *one over f*) is used to describe a number of types of noise that are present when the modulation frequency f is low. This type of noise is also called excess noise because it exceeds shot noise at frequencies below a few hundred Hertz.

The mechanisms that produce $1/f$ noise are poorly understood. The noise power is inversely proportional to f , the modulation frequency. This dependence of the noise power on modulation frequency leads to the name for this type of noise.

To reduce $1/f$ noise, an optical detector should be operated at a reasonably high frequency, often as high as 1000 Hz. This is a high enough value to reduce the contribution of $1/f$ noise to a small amount.

Photon noise

Even if all the previously discussed sources of noise could be eliminated, there would still be some noise in the output of an optical detector because of the random arrival rate of photons in the light being measured and from the background. This contribution to the noise is called *photon noise*; it is a noise source external to the detector. It imposes the ultimate fundamental limit to the detectivity of a photodetector.

The photon noise associated with the fluctuations in the arrival rate of photons in the desired signal is not something that can be reduced. The contribution of fluctuations in the arrival of photons from the background, a contribution that is called background noise, can be reduced.

The background noise increases with the field of view of the detector and with the temperature of the background. In some cases it is possible to reduce the field of view of the detector so as to view only the source of interest. In other cases it is possible to cool the background. Both these measures may be used to reduce the background noise contribution to photon noise.

The types of noise described here, or a combination of them, will set an upper limit to the detectivity of an optical detector system.

II. TYPES OF DETECTORS

We now return to the discussion of different types of detectors and present more detail on the various available optical detectors.

A. Photon detectors

We have defined photon detectors and thermal detectors briefly. We begin a more detailed discussion of detector types with photon detectors. In photon detectors, quanta of light energy produce free electrons. The photon must have sufficient energy to exceed some threshold. In other words, the wavelength must be shorter than the cutoff wavelength. We will consider three types of photoeffects that are often used for detectors. These are the *photovoltaic effect*, the *photoemissive effect*, and the *photoconductive effect*.

Photovoltaic effect

The *photovoltaic effect* occurs at a junction in a semiconductor. The junction is the boundary between a region where the conductivity is due to electrons and a region where the conductivity is due to holes (the absence of electrons). This is called a pn junction. At the junction, an electric field is present internally because there is a change in the level of the conduction and valence bands. This change leads to the familiar electrical rectification effect produced by such junctions. The photovoltaic effect is the generation of a voltage when light strikes a semiconductor pn junction.

The photovoltaic effect is measured using a high-impedance voltage-measuring device, which essentially measures the open-circuit voltage produced at the junction.

In the dark, no open circuit voltage is present. When light falls on the junction, the light is absorbed and, if the photon energy is large enough, it produces free hole-electron pairs. The electric field at the junction separates the pair and moves the electron into the n-type region and the hole into the p-type region. This leads to an open circuit voltage that can be measured externally. This process is the origin of the photovoltaic effect. We note that the open-circuit voltage generated in the photovoltaic effect may be detected directly and that no bias voltage nor load resistor is required.

If the junction is short-circuited by an external conductor, current will flow in the circuit when the junction is illuminated. One may measure either the open-circuit voltage or the short-circuit current. Both these quantities will give measures of the light falling on the junction.

Photoemissive effect

Now we turn to the *photoemissive effect*. The photoemissive effect involves the emission of electrons from a surface irradiated by quanta of light energy. A photoemissive detector has a cathode coated with a material that emits electrons when light of wavelength shorter than the cutoff wavelength falls on the surface. The electrons emitted from the surface are accelerated by a voltage to an anode, where they produce a current in an external circuit. The detectors are enclosed in a vacuum environment to allow a free flow of electrons from anode to cathode. These detectors are available commercially from a number of manufacturers. They represent an important class of detectors for many applications.

Some spectral response curves for photoemissive cathodes are shown in Figure 6-3. The cathodes are often mixtures containing alkali metals, such as sodium and potassium, from which electrons can easily be emitted. The responsivity in mA/watt of these devices is shown in the figure from the ultraviolet to the near infrared. At wavelengths longer than about 1000 nm, no photoemissive response is available. The short-wavelength end of the response curve is set by the nature of the window material used in the tube that contains the detector. The user can select a device that has a cathode with maximum response in a selected wavelength region. An important variation of the photoemissive detector is the photomultiplier, which will be described later.

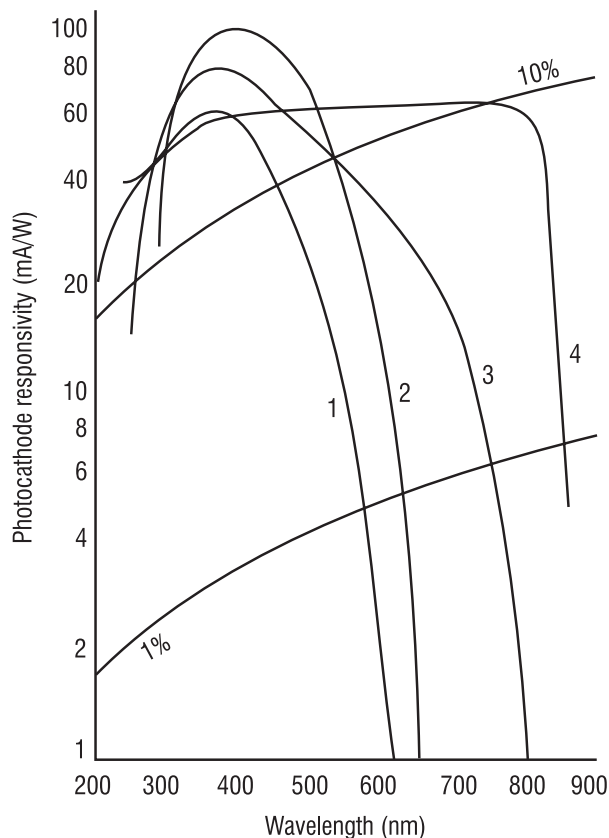


Figure 6-3 Response as a function of wavelength for a number of photoemissive surfaces. Curve 1 is the response of a bi-alkali type of cathode with a sapphire window; curve 2 is for a different bi-alkali cathode with a lime glass window; curve 3 is for a multi-alkali cathode with a lime glass window; and curve 4 is for a GaAs cathode with a 9741 glass window. The curves labeled 1% and 10% denote what the response would be at the indicated value of quantum efficiency.

Photoconductivity

A third phenomenon used in optical detectors is *photoconductivity*. A semiconductor in thermal equilibrium contains free electrons and holes. The concentration of electrons and holes is changed if light is absorbed by the semiconductor. The light must have photon energy large enough to produce free electrons within the material. The increased number of charge carriers leads to an increase in the electrical conductivity of the semiconductor. The device is used in a circuit with a bias voltage and a load resistor in series with it. The change in electrical conductivity leads to an increase in the current flowing in the circuit, and hence to a measurable change in the voltage drop across the load resistor.

Photoconductive detectors are most widely used in the infrared region, at wavelengths where photoemissive detectors are not available. Many different materials are used as infrared photoconductive detectors. Some typical values of detectivity (in $\text{cm}\cdot\text{Hz}^{1/2}/\text{watt}$) as a function of wavelength for some devices operating in the infrared are shown in Figure 6-4, along with values of detectivity for some other detectors to be discussed later. The photoconductive detectors are denoted PC. The exact value of detectivity for a specific photoconductor depends on the operating temperature and on the field of view of the detector. Most infrared photoconductive detectors operate at a cryogenic temperature (frequently liquid nitrogen temperature, 77 K) which may involve some inconvenience in practical applications.

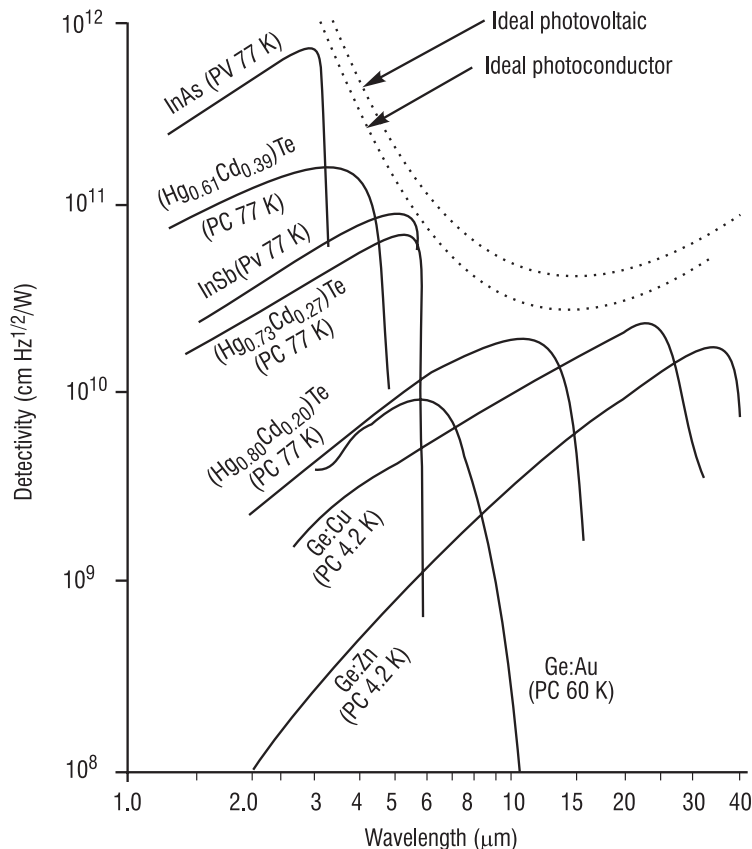


Figure 6-4 Detectivity as a function of wavelength for a number of different types of photodetectors operating in the infrared spectrum. The temperature of operation is indicated. Photovoltaic detectors are denoted PV; photoconductive detectors are denoted PC. The curves for ideal photodetectors assume a 2π steradian field of view and a 295 K background temperature.

A photoconductive detector uses a crystal of semiconductor material that has low conductance in the dark and an increased value of conductance when it is illuminated. It is commonly used in a series circuit with a battery and a load resistor. The semiconductor element has its conductance increased by light. The presence of light leads to increased current in the circuit and to increased voltage drop across the load resistor.

We now consider two specific types of photon detectors that are especially useful in photonics, the photodiode and the photomultiplier.

Photodiodes

We have discussed the photovoltaic effect, for which no bias voltage is required. It is also possible to use a pn junction to detect light if one does apply a bias voltage in the reverse direction. The reverse direction is the direction of low current flow, that is, with the positive voltage applied to the n-type material. A pn junction detector with bias voltage is termed a photodiode.

Figure 6-5 shows the current-voltage characteristics of a photodiode. The curve marked dark shows the current-voltage relation in the absence of light. It shows the familiar rectification characteristics of a pn semiconductor diode. The other curves represent the current-voltage characteristics when the device is illuminated at different light levels. A photovoltaic detector, with zero applied voltage, is represented by the intersections of the different curves with the vertical axis. Figure 6-5 is intended to show qualitatively how a photodiode operates. No quantitative values are shown for the axes in this figure; these values will vary from one material to another.

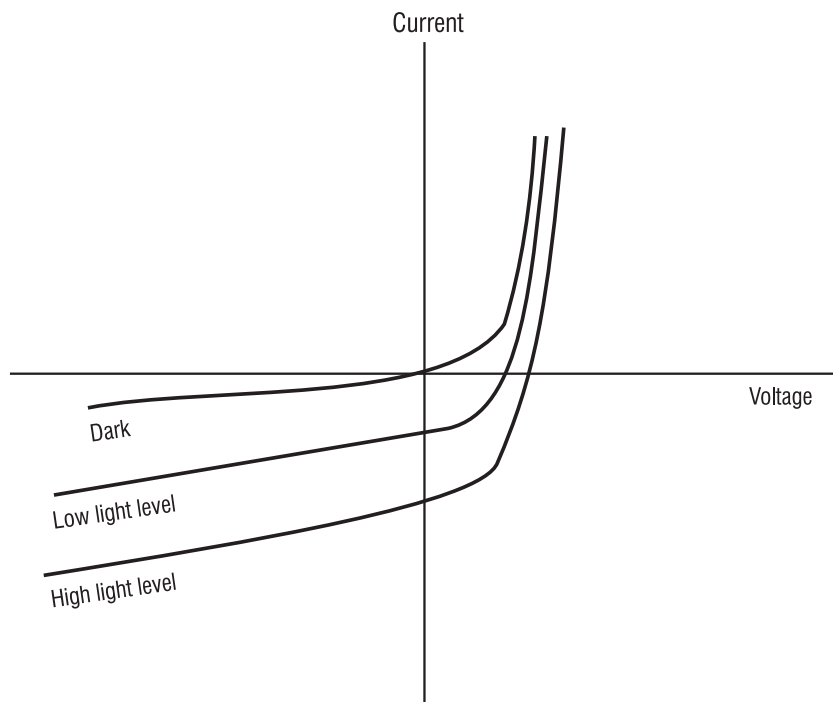


Figure 6-5 *Current-voltage characteristic for a photodiode*

A *photodiode detector* is operated in the lower left quadrant of this figure, where the current that may be drawn through an external load resistor increases with increasing light level. In practice, one measures the voltage drop appearing across the load resistor.

A variety of photodiode structures are available. No single photodiode structure can best meet all requirements. Perhaps the two most common structures are the *planar diffused photodiode*, shown in Figure 6-6a, and the *Schottky photodiode*, shown in Figure 6-6b. The planar diffused photodiode is fabricated by growing a layer of oxide over a slice of high-resistivity silicon, etching a hole in the oxide and diffusing boron into the silicon through the hole. This structure leads to devices with high breakdown voltage and low leakage current. The circuitry for operation of the photodiode is also indicated, including the load resistor.

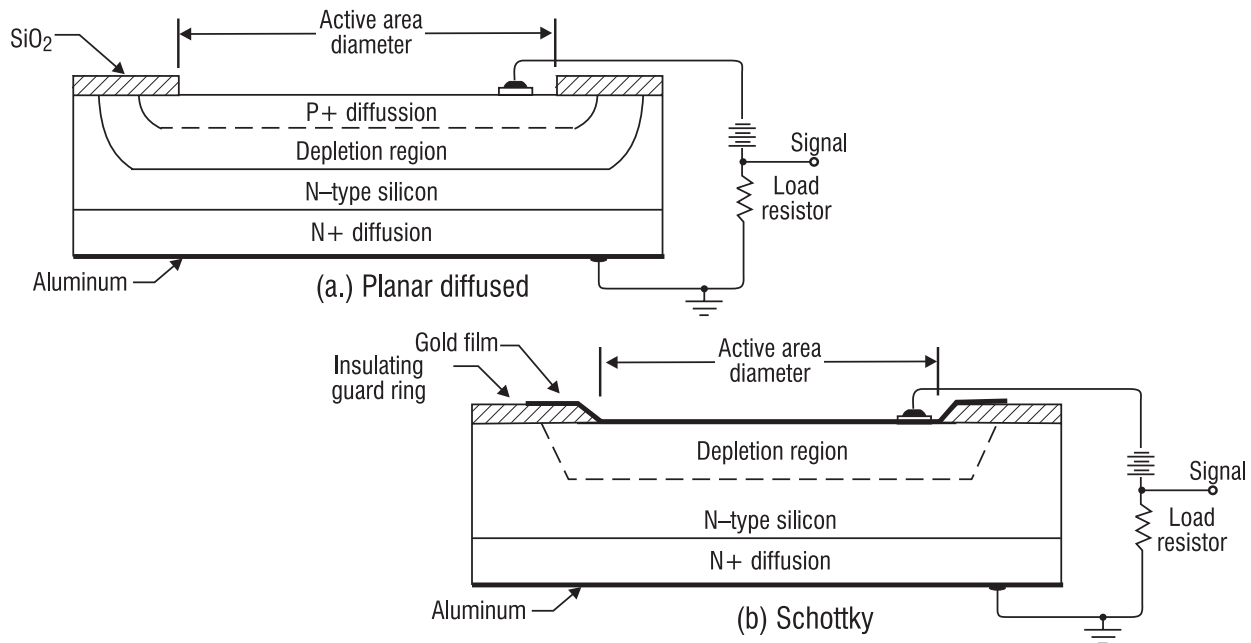


Figure 6-6 Photodiode structures: (a) Planar diffused photodiode; (b) Schottky photodiode

The Schottky barrier photodiode is formed at a junction between a metallic layer and a semiconductor. If the metal and the semiconductor have work functions related in the proper way, this can be a rectifying barrier. The junction is fabricated by oxidation of the silicon surface, then etching of a hole in the oxide, followed by the evaporation of a thin transparent and conducting gold layer. The insulation guard rings serve to reduce the leakage current through the device.

A number of different semiconductor materials are in common use as photodiodes. They include silicon for use in the visible, near ultraviolet, and near infrared; germanium and indium gallium arsenide in the near infrared; and indium antimonide, indium arsenide, mercury cadmium telluride, and germanium doped with elements like copper and gold in the longer-wavelength infrared.

The most frequently encountered type of photodiode is silicon. *Silicon photodiodes* are widely used as the detector elements in optical disks and as the receiver elements in optical fiber telecommunication systems operating at wavelengths around 800 nm. Silicon photodiodes

respond over the approximate spectral range of 400–1100 nm, covering the visible and part of the near infrared regions. The spectral responsivity (in A/watt) of typical commercial silicon photodiodes is shown in Figure 6-7. The responsivity reaches a peak value around 0.7 amp/watt near 900 nm, decreasing at longer and shorter wavelengths. Optional models provide somewhat extended coverage in the infrared or ultraviolet regions. Silicon photodiodes are useful for detection of many of the most common laser wavelengths, including argon, HeNe, AlGaAs, and Nd:YAG.

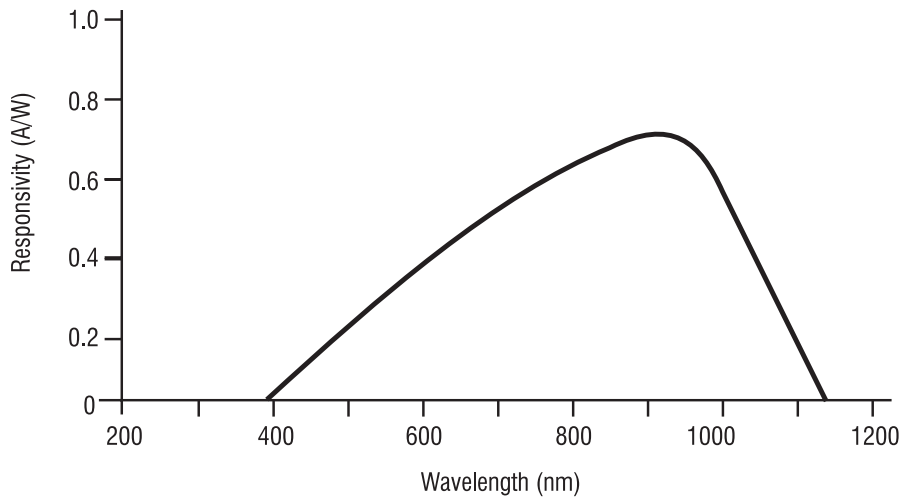


Figure 6-7 Responsivity as a function of wavelength for typical silicon photodiodes

In practice, *silicon photodiodes* have become the detector of choice for many photonics applications within their spectral range. They use well-developed technology and are widely available. They represent the most widely used type of detector for lasers operating in the visible and near infrared portions of the spectrum.

Figure 6-4 shows the spectral D^* (or detectivity) for a number of commercially available detectors operating in the infrared spectrum. The figure includes both photovoltaic detectors (denoted PV) and photoconductive detectors (denoted PC). The choice of detector will depend on the wavelength region that is desired. For example, for a laser operating at 5 μm , an indium antimonide photovoltaic detector would be suitable.

Figure 6-4 also indicates the detectivity for “ideal” detectors, that is, detectors whose performance is limited only by fluctuations in the background of incident radiation, and that do not contribute noise themselves. Available detectors approach the ideal performance limits fairly closely.

PIN photodiodes

Another common type of semiconductor structure used in photodiodes is the so-called *PIN structure*. This structure was developed to increase the frequency response of photodiodes. The device has a layer of nearly intrinsic semiconductor material bounded on one side by a relatively thin layer of highly doped p-type semiconductor and on the other side by a relatively thick layer of n-type semiconductor. Hence it is called a PIN device.

Light that is absorbed in the intrinsic region produces free electron-hole pairs, provided that the photon energy is high enough. These carriers are swept across the region with high velocity and are collected in the heavily doped regions. The frequency response of PIN photodiodes can be very high, of the order of 10^{10} Hz. This is higher than the frequency response of pn junctions without the intrinsic region.

Avalanche photodiodes

Another variation of the photodiode is the *avalanche photodiode*. The avalanche photodiode offers the possibility of internal gain; it is sometimes referred to as a “solid-state photomultiplier.” The most widely used material for avalanche photodiodes is silicon, but they have been fabricated from other materials, such as germanium.

An avalanche photodiode has a diffused pn junction, with surface contouring to permit high reverse bias voltage without surface breakdown. A large internal electric field leads to multiplication of the number of charge carriers through ionizing collisions. The signal is thus increased, to a value perhaps 100–200 times greater than that of a nonavalanche device. The detectivity is also increased, provided that the limiting noise is not from background radiation. Avalanche photodiodes cost more than conventional photodiodes and they require temperature-compensation circuits to maintain the optimum bias, but they represent an attractive choice when high performance is required.

Photomultipliers

Previously, we described photoemissive detectors in which current flows directly from a photocathode to an anode. We turn now to an important photoemissive detector that provides for amplification of the current. This is the *photomultiplier*. This device has a photoemissive cathode and a number of secondary emitting stages called dynodes. The dynodes are arranged so that electrons from each dynode are delivered to the next dynode in the series. Electrons emitted from the cathode are accelerated by an applied voltage to the first dynode, where their impact causes emission of numerous secondary electrons. These electrons are accelerated to the next dynode and generate even more electrons. Finally, electrons from the last dynode are accelerated to the anode and produce a large current pulse in the external circuit. The photomultiplier is packaged as a vacuum tube.

Figure 6-8 shows a cross-sectional diagram of a typical photomultiplier tube structure. This tube has a transparent end window with the underside coated with the photocathode material.

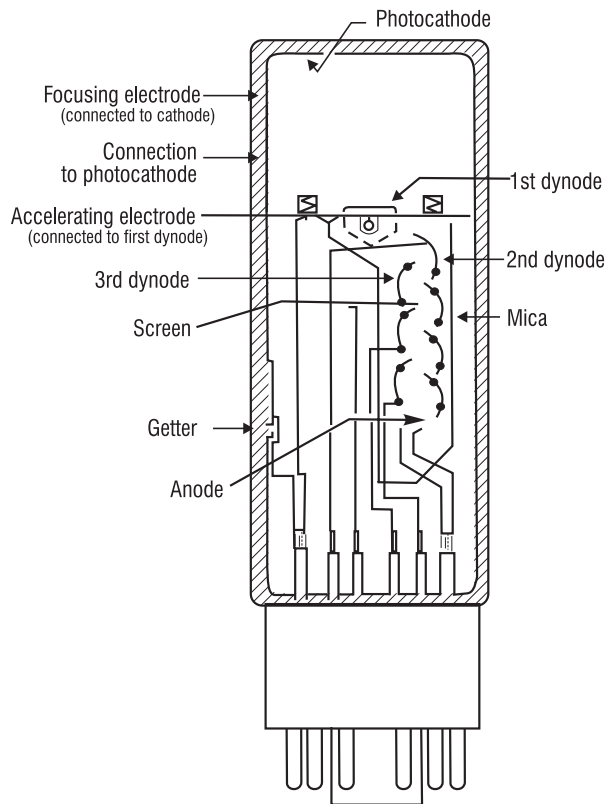


Figure 6-8 Diagram of typical photomultiplier tube structure

Figure 6-9 shows the principles of operation of the tube. Photoelectrons emitted from the cathode strike the first dynode, where they produce 1 to 8 secondary electrons per incident electron. These are accelerated to the second dynode, where the process is repeated. After several such steps the electrons are collected at the anode and flow through the load resistor. Voltages of 100 to 300 volts accelerate electrons between dynodes, so that the total tube voltage may be from 500 to 3000 volts from anode to cathode, depending on the number of dynodes.

The *current gain* of a photomultiplier is the *ratio of anode current to cathode current*. Typical values of gain may be in the range 100,000 to 1,000,000. Thus 100,000 or more electrons reach the anode for each photon striking the cathode. Photomultiplier tubes can in fact detect the arrival of a single photon at the cathode.

Figure 6-10 shows the gain as a function of the voltage from the anode to the cathode for a typical photomultiplier tube. This high gain process means that photomultiplier tubes offer the highest available responsivity in the ultraviolet, visible, and near infrared portions of the spectrum. But their response does not extend to wavelengths longer than about 1000 nm.

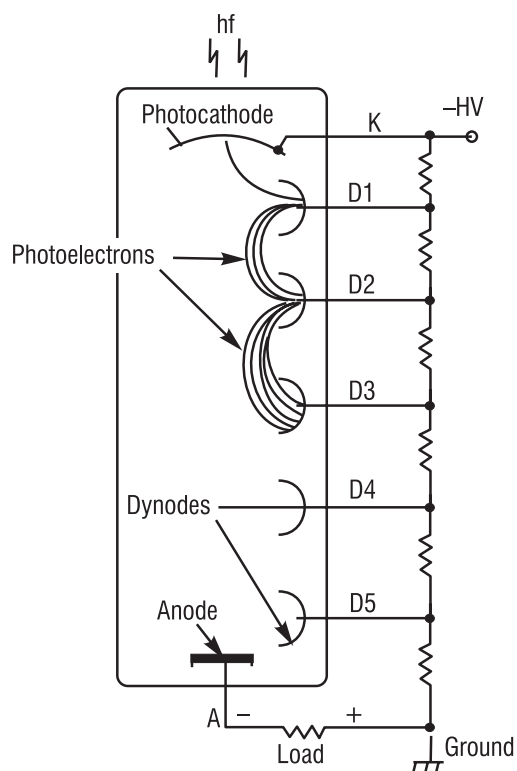


Figure 6-9 Principles of photomultiplier operation. The dynodes are denoted $D1$, $D2$, etc.

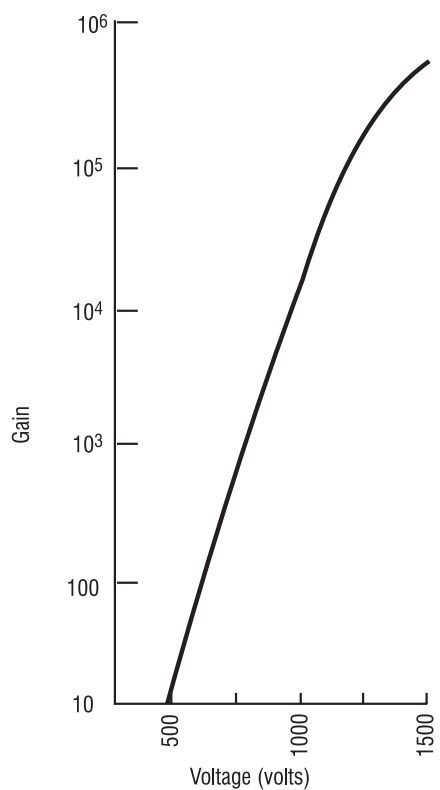


Figure 6-10 Photomultiplier gain as a function of applied voltage

Photomultiplier tubes are used in many photonics applications, such as air-pollution monitoring, star tracking, photometry, and radiometry.

B. Thermal detectors

The second broad class of optical detectors, thermal detectors, responds to the total energy absorbed, regardless of wavelength. Thus thermal detectors do not have a long-wavelength cutoff in their response, as photon detectors do. The value of D^* for a thermal detector is independent of wavelength. Thermal detectors generally do not have as rapid a response as do photon detectors. For many photonics applications, they are often not used in the wavelength region in which photon detectors are most effective ($\leq 1.55 \mu\text{m}$). They are often used at longer wavelengths.

Bolometers and thermistors

In perhaps the most common manifestation of thermal detectors, the optical energy is absorbed by an element whose properties change with temperature. As the light energy is absorbed, the temperature of the element increases and the change in its properties is sensed. The temperature-measuring elements include *bolometers* and *thermistors*. Bolometers and thermistors respond to the change in electrical resistivity that occurs as temperature rises. Bolometers use metallic elements; thermistors use semiconductor elements. The bolometer or thermistor is in a circuit in series with a voltage source, so that current flows through it and, as the resistance changes, the voltage drop across the element changes, providing a sensing mechanism.

Thermocouples

In another manifestation, light is absorbed by an element to which a thermocouple is attached. The thermocouple is a device formed of two dissimilar metals joined at two points. Thermocouples may be fabricated from wires, but for detector applications they are often fabricated as thin films. The device generates a potential difference, which is a measure of the temperature difference between the points. One point is held at a constant reference temperature. The second point is in contact with the absorber. The light energy heats the absorber and the thermocouple junction in contact with it. This causes the voltage generated by the thermocouple to change, giving a measure of the temperature rise of the absorber and of the incident light energy.

To enhance the performance of the thermocouples, often there are a number of thermocouples in series, perhaps as many as 100. The “hot” junctions are all attached close together. This type of device is called a *thermopile*.

Figure 6-11 shows values of $D^*(\lambda, 1000, 1)$ (see section I.C) for some thermal detectors, including thermistors, bolometers, thermopiles and pyroelectric detectors, which will be described later. The values are independent of wavelength. In the visible and near infrared, the values of D^* for thermal detectors tend to be lower than for good photon detectors, but the response does not decrease at long wavelength.

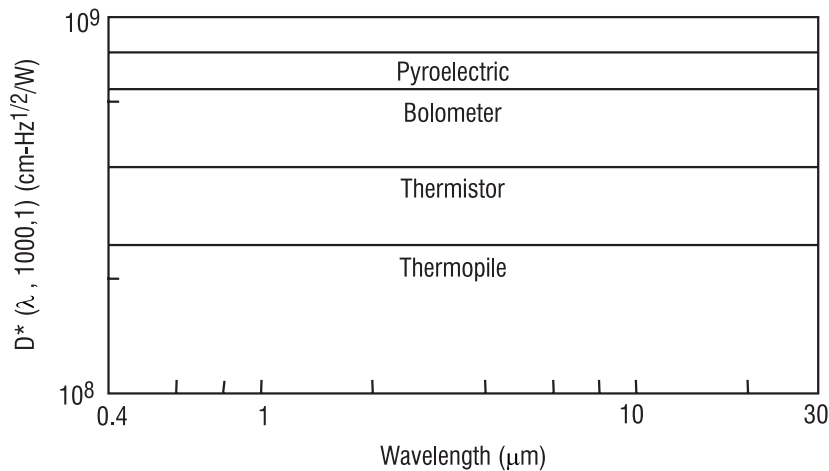


Figure 6-11 Detectivity (D^*) as a function of wavelength for several typical thermal detectors. The temperature of operation is 295 K.

Calorimeters

Measurements of pulse energy are frequently made using a *calorimeter*, which represents a common thermal detector system. Calorimetric measurements yield a simple determination of the total energy in an optical pulse, but calorimeters usually do not respond rapidly enough to follow the pulse shape. Calorimeters designed for photonics measurements often use blackbody absorbers with low thermal mass and with temperature-measuring devices in contact with the absorber to measure the temperature rise. Knowledge of the thermal mass coupled with measurement of the temperature rise yields the energy in the optical pulse.

A variety of calorimeter designs have been developed for measuring the total energy in an optical pulse or for integrating the output from a continuous optical source. Since the total energy in a pulse is usually not large, the calorimetric techniques are rather delicate. The absorbing medium must be small enough that the absorbed energy may be rapidly distributed throughout the body. It must be thermally isolated from its surroundings so that the energy is not lost.

A commonly encountered calorimeter design, the so-called cone calorimeter, uses a small, hollow carbon cone, shaped so that radiation entering the base of the cone will not be reflected back out of the cone. Such a design is a very efficient absorber. Thermistor beads or thermocouples are placed in contact with the cone. The thermistors form one element of a balanced bridge circuit, the output of which is connected to a display or meter. As the cone is heated by a pulse of energy, the resistance of the bridge changes, leading to an imbalance of the bridge and a voltage pulse that activates the display. The pulse decays as the cone cools to ambient temperature. The magnitude of the voltage pulse gives a measure of the energy in the pulse. Two identical cones may be used to form a conjugate pair in the bridge circuit. This approach allows cancellation of drifts in the ambient temperature.

Pyroelectric detectors

Another type of thermal detector is the *pyroelectric detector*. Pyroelectric detectors respond to the change in electric polarization that occurs in certain classes of crystalline materials (like lithium tantalate) as their temperatures change. The *change in polarization*, called the pyroelectric effect, may be measured as an open-circuit voltage or as a short-circuit current. Because they respond to changes in temperature, pyroelectric devices are useful as detectors for only pulsed or chopped radiation.

The response speed of pyroelectric detectors is fast, faster than that of other thermal detectors like thermistors and thermopiles. Pyroelectric detectors are fast enough to detect very short optical pulses.

The spectral detectivity D^* of pyroelectric detectors was shown in Figure 6-11. It tends to be higher than the detectivity of thermistor or thermopile detectors, and it is independent of wavelength.

III. CALIBRATION

The response of an optical detector in current (or voltage) per unit input of power is often taken as the nominal value specified by the manufacturer. But, for precise work, the detector may have to be calibrated by the user. Accurate absolute measurements of optical power or energy are difficult. A good calibration requires very careful work.

A. Response of detector

One widely used calibration method involves measurement of the total energy in the laser beam (with a calorimetric energy meter) at the same time that the detector response is determined. The temporal history of the energy delivery is known from the shape of the detector output. Since the power integrated over time must equal the total energy, the detector calibration is obtained in terms of laser power per unit of detector response.

In one common approach, you can use a calorimeter to calibrate a detector, which is then used to monitor the laser output from one pulse to another. A small fraction of the laser beam is diverted by a beam splitter to the detector, while the remainder of the laser energy goes to the calorimeter. The total energy arriving at the calorimeter is determined. The temporal history of the detector output gives the pulse shape. Then numerical or graphical integration yields the calibration of the response of the detector relative to the calorimeter. The calibration may be in terms of power or energy in the laser pulse. If you know the fraction of the total beam energy diverted to the detector, you can calibrate the detector response in terms of the energy in the pulse. If the pulse shape is stable from pulse to pulse, you can use the results of the numerical or graphical integration to determine the peak power in the pulse.

If the response of the calorimeter is fast, it can be used for measurement of power from a continuous source. The temperature of the absorber will reach an equilibrium value dependent on the input power. Such units are available commercially as laser power meters. Compared to the power meters based on silicon or other photodiodes, the power meters based on absorbing cones or disks are useful over a wider range of wavelength and do not require use of a compensating factor to adjust for the change in response as the laser wavelength changes.

After the calibration is complete, you can remove the calorimeter and use the main portion of the beam for the desired application. The detector, receiving the small portion of the beam directed to it by the beam splitter, acts as a pulse-to-pulse monitor.

B. Techniques to limit beam power

Filters

Quantitative measurements of laser output involve several troublesome features. The intense laser output tends to overload and saturate the output of detectors if they are exposed to the full power. Thus, absorbing filters may be used to cut down the input to the detector. A suitable filter avoids saturation of the detector, keeps it in the linear region of its operating characteristics, shields it from unwanted background radiation, and protects it from damage. Many types of attenuating filters have been used, including *neutral-density filters*, *semiconductor wafers* (like silicon), and *liquid filters*.

We note that filters also may saturate and become nonlinear when exposed to high irradiance. If a certain attenuation is measured for a filter exposed to low irradiance, the attenuation may be less for a more intense laser beam. Thus, a measurement must be performed at a low enough irradiance so that the filter does not become saturated.

Beam splitters

The use of beam splitters also can provide attenuation of an intense laser beam. If the beam is incident on a transparent dielectric material inserted at an angle to the beam, there will be specular reflection of a portion of the beam. One may measure this reflected beam, which will contain only a small fraction of the incident power. The fraction may be determined using Fresnel's equations. The calculation requires knowledge of the geometry and the index of refraction of the dielectric material.

Lambertian reflectors

Another method for attenuating the beam before detection is to allow it to fall normally on a diffusely reflecting massive surface, such as a magnesium oxide block. The arrangement is shown in Figure 6-12. The angular distribution of the reflected light is proportional to the angle θ between the normal to the surface and the direction of observation. Thus, the power reflected is maximum along the normal to the surface and decreases to zero at 90 degrees to the surface. This dependency is called *Lambert's cosine law*, and a surface that follows this law is called a *Lambertian surface*. Many practical surfaces follow this relation, at least approximately. The power P_{detector} that reaches the detector after reflection from such a surface is

$$P_{\text{detector}} = P_{\text{tot}} \frac{A_d}{\pi D^2} \cos \theta \quad (6-4)$$

where P_{tot} is the total laser power, A_d is the area of the detector (or its projection on a plane perpendicular to the line from the target to the detector), and D is the distance from the target to the detector. This approximation is valid when D is much larger than the detector dimensions and the transverse dimension of the laser beam.

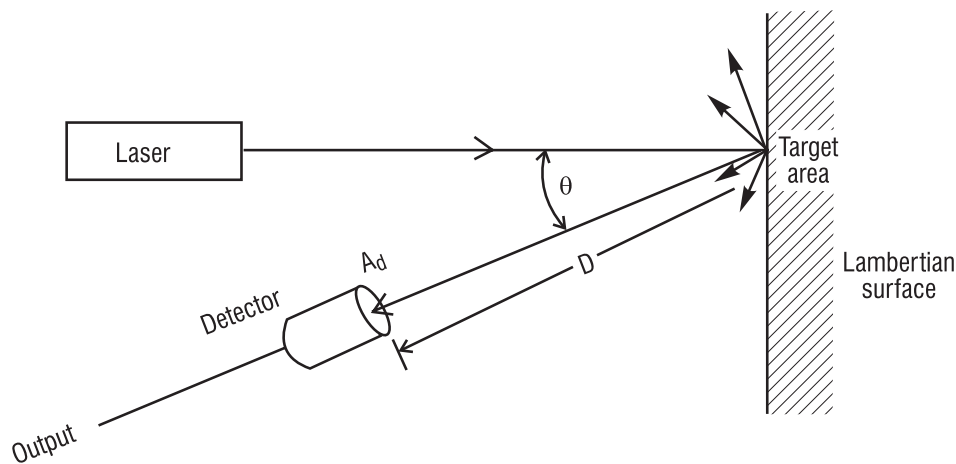


Figure 6-12 Arrangement for measuring laser power using a Lambertian reflector to attenuate the power reaching the detector. D is the distance from the surface target to the detector and A_d is the area of the detector.

With a Lambertian reflector, the power incident on the photosurface can be adjusted in a known way by changing the distance D or the angle θ . The beam may be spread over a large enough area on the Lambertian surface so that the surface is not damaged. The distance D is made large enough to ensure that the detector is not saturated. The measurement of the power received by the detector, plus some easy geometric parameters, gives the fraction of the beam power reaching the detector.

Example 4

A laser beam with total power 10 watts is incident at normal incidence on a Lambertian surface. How much power reaches a detector with an area of 0.5 cm^2 at an angle of 45° if the detector area is 30 cm from where the beam strikes the target reflecting surface?

Solution

According to Equation 6-4, the power reaching the detector is

$$\begin{aligned} P_{\text{detector}} &= 10 \text{ W} \times 0.5 \text{ cm}^2 / \{\pi \times (30 \text{ cm})^2\} \times \cos 45^\circ \\ &= 10 \times 0.5 / (3.1416 \times 900) \times 0.707 \text{ W} \\ P_{\text{det}} &= 0.00125 \text{ W} \end{aligned}$$

C. Electrical calibration

It is also possible to calibrate power meters electrically. It is assumed that the deposition of a given amount of energy in the absorber provides the same response, independent of whether the energy is optical or electrical.

The absorbing element is heated by an electrical resistance heater. The electrical power dissipation is determined from electrical measurements of the current and voltage. The measured response of the instrument to the known electrical input provides the calibration.

Accurate absolute measurement of optical power is difficult. Thus, one must use great care in the calibration of optical detectors.

IV. CIRCUITRY FOR OPTICAL DETECTORS

The basic power supply for an optical detector contains a voltage source and a load resistor in series with the detector. As the irradiance on the detector element changes, the current in the circuit changes and the voltage drop across the load resistor changes. Measurement of the voltage drop provides the basis for the optical power measurement.

A variety of different circuits may be used, depending on the detector type and on the application. A full description of all the types of detector circuits is beyond the scope of this module. We shall describe electrical circuitry used with two representative types of detectors, the photovoltaic detector and the photoconductive detector.

A. Basic circuit for a photovoltaic detector

A photovoltaic detector requires no bias voltage; it is a voltage generator itself. A basic circuit for a photovoltaic detector is shown in Figure 6-13. This shows the conventional symbol for a photodiode at the left. The symbol includes the arrow representing incident light. The incident light generates a voltage from the photodiode, which causes current to flow through the load resistor R_L . The resulting voltage drop across the resistor again is available as a signal to be monitored.

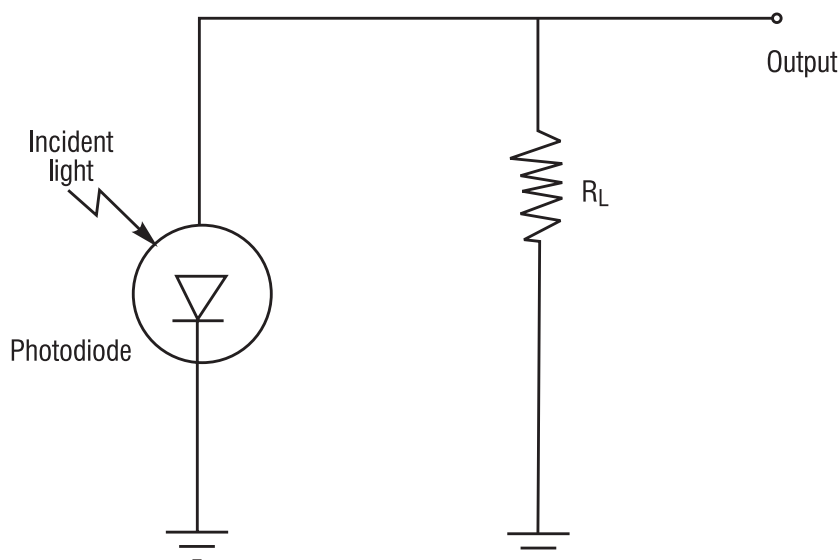


Figure 6-13 Basic circuit for operation of a photovoltaic detector. The symbol for a photodiode is indicated. The load resistor is R_L .

OPTICAL DETECTORS AND HUMAN VISION

In this configuration it is assumed that the value of the load resistor is much larger than the value of the shunt resistance of the detector. The shunt resistance is the resistance of the detector element in parallel with the load resistor in the circuit. The value of the shunt resistance is specified by the manufacturer and for silicon photodiodes may be a few megohms to a few hundred megohms.

Disadvantages of this circuit are that the response is nonlinear (it is logarithmic) and the signal depends on the shunt resistance of the detector, which may vary in different production batches of detectors.

Practical loads that need to be driven are usually much lower than those that can be used with the photovoltaic diode. In order to counter this disadvantage, an amplifier can be used as a buffer between an acceptable high load resistor for the diode and a much lower useful load resistance. Figure 6-14 shows this configuration.

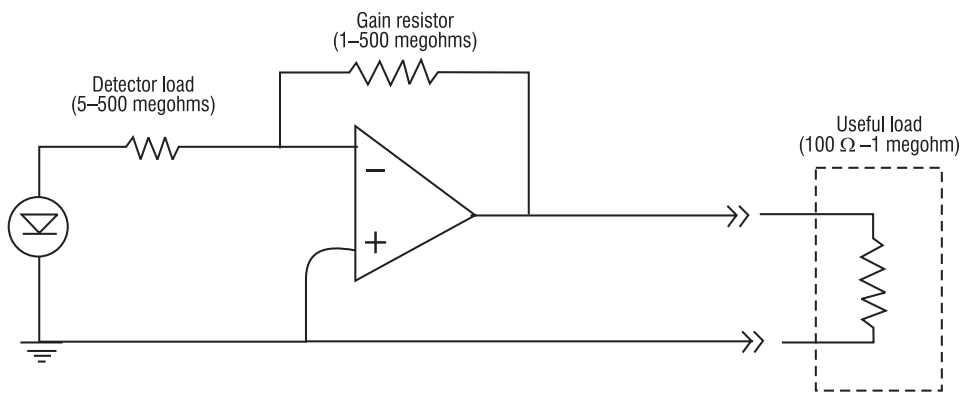


Figure 6-14 Circuit for photovoltaic detector operation with a high detector load resistance driving a useful load of lower resistance

This circuit has a linear response to the incident light intensity. It also is a low noise circuit because it has almost no leakage current, so that shot noise is low.

B. Basic circuit for a photoconductive detector

We previously noted that photodiodes may be operated in a photoconductive mode. Figure 6-15 shows a circuit which provides this type of operation. The diode is reverse biased, so that the operation is in the third quadrant of Figure 6-5. The photocurrent produces a voltage across the load resistor which is in parallel with the shunt resistance of the detector. The shunt resistance is nearly constant. One may use large values of load resistance, so as to obtain large values of signal, and still obtain linear variation of the output with the optical power.

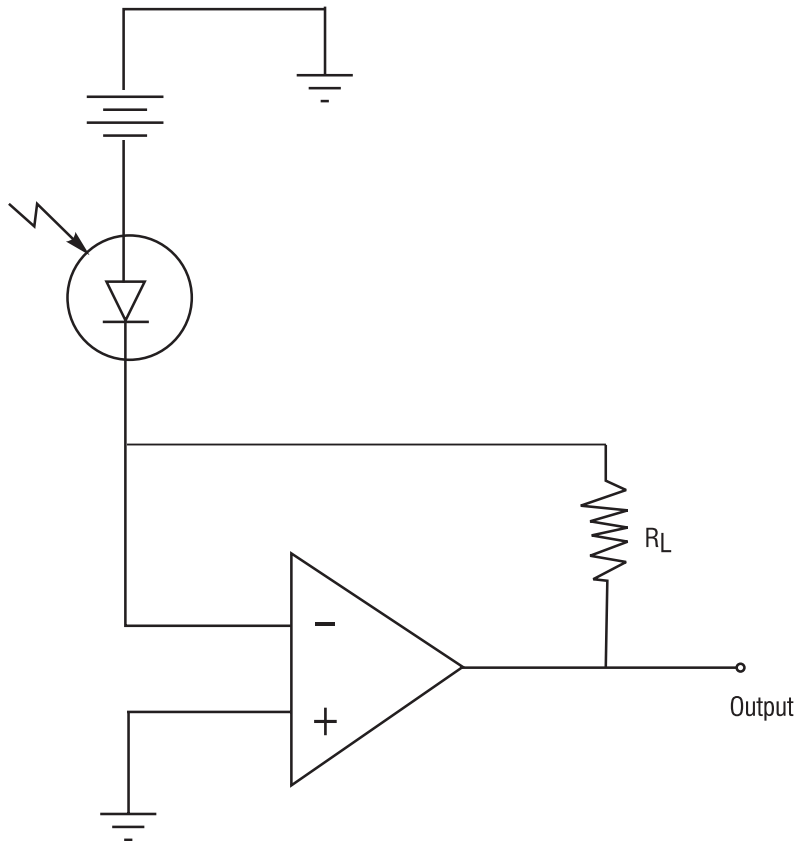


Figure 6-15 Circuit for operation of a photodiode in the photoconductive mode. The load resistor is R_L .

This circuit can provide very high-speed response. It is possible to obtain rise times of one nanosecond or below with this type of circuit. The biggest disadvantage of this circuit is the fact that the leakage current is relatively large, so that the shot noise is increased.

V. HUMAN VISION

A. The eye as an optical detector

An important optical detector is the human eye. In some respects, the eye can be regarded as a specialized type of detector system, with attributes similar to those of the other detectors that we have considered. In common with other optical detectors, the eye is a *square-law detector*, responding to the incident radiant power, which is proportional to the square of the electric field in the light wave.

The eye has a spectral response that covers the range approximately from 400 to 700 nm, the range that is called the visible spectrum. At longer and shorter wavelengths, the eye is not able to detect incident optical energy.

B. Structure of the eye

The eye can be considered as a complete optical system, including packaging, a variable aperture, a curved corneal surface and a lens that provide for imaging, a variable focus capability, a photosensor, and an output to a computer, the brain. Figure 6-16 shows a simplified diagram of the structure of the eye. The eye is approximately spherical in shape and is contained within a tough, fibrous envelope of tissue called the sclera. The sclera covers all the eyeball except for a portion of its front. At the front of the eyeball is the cornea, which has a refractive index around 1.38. The cornea is a transparent membrane that allows light to enter the eyeball and that contributes significantly to the focusing capability of the eye. Behind the cornea is the iris, an adjustable aperture that expands in dim light and contracts in bright light, controlling the amount of light which enters the eyeball. The pupil of the eye is the opening in the center of the aperture defined by the iris. Light entering the eye passes through the pupil.

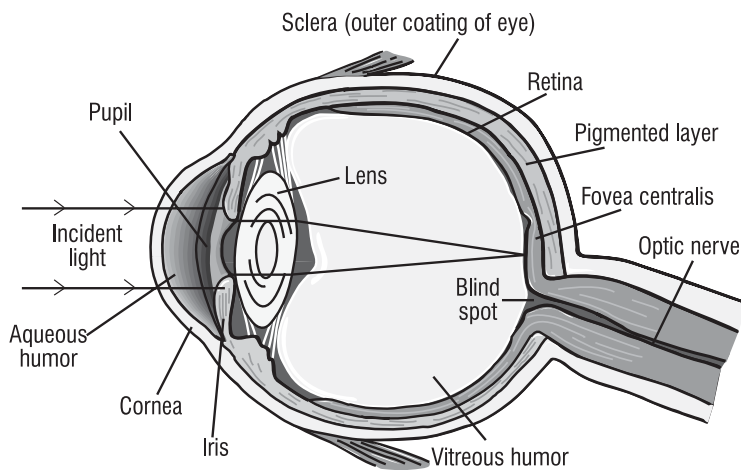


Figure 6-16 *Structure of the human eye*

The region behind the cornea contains a transparent liquid called the aqueous humor with refractive index around 1.34. Then there is the lens of the eye, a capsule of fibrous jelly-like material, with refractive index varying from 1.41 in the center to 1.39 at the periphery. The shape of the lens can be changed by muscles attached to it. This allows for fine focusing of light entering the eye.

After the lens is a transparent thin jelly called the vitreous humor. It has a refractive index around 1.34. Finally, covering most of the back surface of the eyeball is the retina, the photosensitive medium that serves as the actual detector material.

The cells in the retina are of two types, called rods and cones. The rods and cones serve different functions, the cones providing the most sensitive vision near the center of the retina and the rods the peripheral vision farther out in the retina. The rods also are more sensitive in dim light than are the cones, so that the rods tend to dominate night vision.

Near the center of the retina is a slight depression, called the fovea centralis, that contains only cones. This region provides the most acute vision.

The rods and cones are connected through nerve fibers to the optic nerve, which emerges from the back of the eyeball. The rods and cones receive the optical image and transmit it through the

nerve fibers to the brain. At the point where the optic nerve exits the eyeball, there are no rods or cones, so there is a small blind spot at that position.

C. Operation of the eye

The eye is an imaging system. The substantial refraction of incoming light energy by the cornea and the action of the lens combine to form an image of the pattern of incident light on the retina. Because the index of refraction of the lens (about 1.40) is not too different from that of the aqueous and vitreous media (about 1.34), much of the refraction of light entering the eyeball occurs at the cornea, as mentioned earlier.

When a normal eye is relaxed, light from very distant objects is focused on the retina. The light rays from the distant object enter the eye as parallel rays. The eye is said to be focused at infinity.

Fine focusing of light coming from points other than infinity is accomplished by changing the shape of the lens. Muscles attached to the lens and to the eyeball accomplish this. In this way the eye may form a sharp focus of nearby objects. This process is called accommodation. In some cases, where the cornea is too coned, and the length of the eyeball is a bit too long or too short, *corneal sculpting*, performed with lasers, can be carried out to improve visual acuity.

The spectral response of the eye is shown in Figure 6-17. The y-axis is the relative response of the eye, normalized to unity at its peak near 555 nm, as a function of wavelength. This curve is the so-called photopic response, which is characteristic of the cones. It is usually stated to cover the range from 400 to 700 nm, but there is some lower response at longer wavelengths. The peak response is in the green portion of the spectrum, near 555 nm.

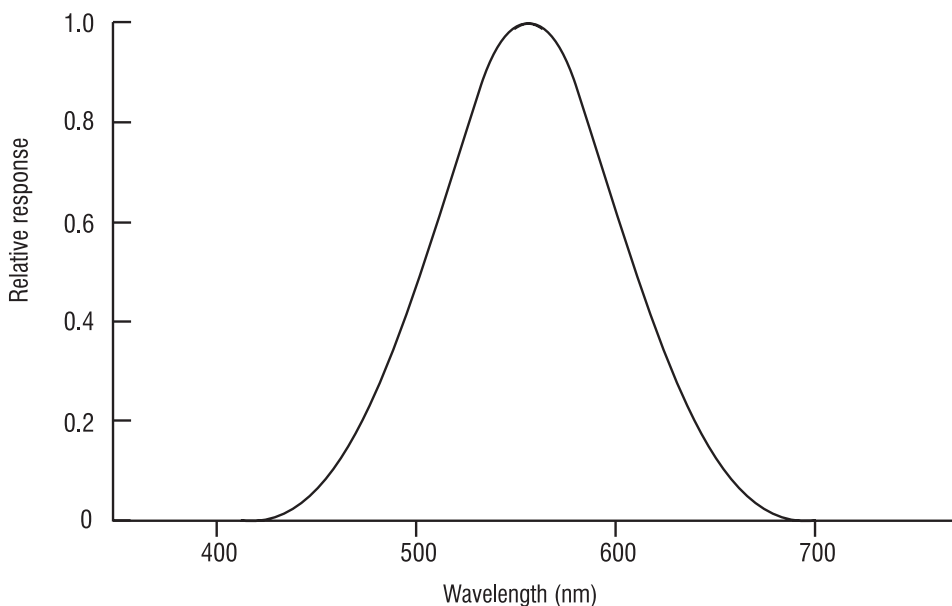


Figure 6-17 Relative spectral response of the eye (photopic response) as a function of wavelength, normalized to unity at its peak value near 555 nm

The rods in the eye have a somewhat different response, called the scotopic response. The peak of the scotopic response is shifted toward the blue.

The interaction of light with the structures of the eye leads to the phenomenon called vision. Vision may be considered to be the sensation in the consciousness of a human observer when the retina of the eye is stimulated by optical energy with appropriate wavelength. The process of vision begins with photochemical changes that occur within the retinal cells when light is absorbed by them. It continues as the complex organic molecules produced in the photochemical processes cause signals to propagate through the nerve fibers to the brain. Finally, in a very important portion of the process, the brain interprets the signals as images corresponding to external objects. This is the process by which the observer becomes aware of optical images through visual sensations that arise because of stimulation of the retina by optical energy.

D. Color

Human vision includes the sensation of color. Color may be considered to be the aspect of things that is caused by differing qualities of the light reflected or emitted by them, other than spatial and temporal inhomogeneities. It is definable in terms of the observer as the appearance of objects or light sources described in terms of the individual's perception of them, involving *brightness*, *saturation* and *hue*.

Taken together, the three attributes of brightness, saturation and hue make up the sensation of color. We will discuss these three attributes one at a time.

Brightness

For *brightness*, consider a series of neutral grays, ranging from white at one end to black at the other. White evokes the greatest sensation of brightness and black the least. Other of the neutral grays are in between. A colored sample may be compared with the series of neutral grays. It will evoke the same sensation of brightness as some member of the group of grays. Brightness is then defined as the attribute of any color that allows it to be considered as equivalent in the sensation produced by some member of the series of neutral grays.

Saturation

The *saturation* is the attribute that describes the extent by which a color departs from a neutral gray and approaches a pure color.

Hue

The attribute of *hue* is the property of color by which it can be perceived as ranging from red through orange, yellow, green, etc. It is related to a property called the dominant wavelength of the light, which will be defined later.

We may clarify these concepts by considering the so-called *chromaticity diagram*. The chromaticity diagram allows us to specify any color in terms of numbers. The chromaticity diagram, usually presented in full color, is shown in Figure 6-18 in black and white. Usually the interior of the diagram is filled in with varying colors.

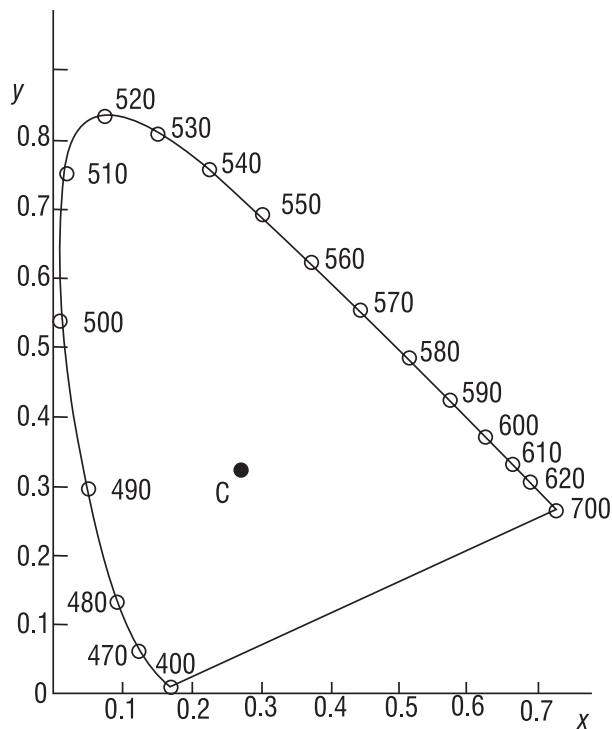


Figure 6-18 Black-and-white version of the chromaticity diagram. The white light point is denoted C . The numbers around the edge of the curve denote the wavelengths in nanometers of pure spectral colors.

The wing-like boundary edge of the curve represents pure colors of the visible electromagnetic spectrum. These colors represent monochromatic light of one wavelength and are denoted by the wavelength in nanometers, ranging from 400 to 700 nm. A straight line from 400 to 700 nm completes the closed curve.

The interior of the curve represents all colors. Shades of blue would be found inside the curve near the number 480, shades of yellow near 575, etc. The point marked C represents “white light,” or average daylight. Any color within the diagram can be expressed quantitatively by the two coordinates x and y .

With the aid of the chromaticity diagram, the hue of a color can be expressed in terms of its principal wavelength. For a given color, with coordinates x and y in the diagram, a line is drawn from the point C through the point given by x and y and extended to the edge of the diagram, where it intersects the edge at some pure spectral color. That wavelength of that color is the principal wavelength for the given color.

The purity, related to the saturation, may be found in the same way. On the line from C through the coordinates x and y to the edge of the diagram, the purity of the color expressed by x and y is the distance from C to the point represented by x and y , expressed as a percentage of the distance from C to the edge of the diagram. Thus, the purity of white light is 0% and the purity of a spectral color at the edge of the diagram is 100%.

E. Defects of vision

Vision can be imperfect in a number of ways. Some imperfections arise because there is an incorrect relation between the positions of various parts of the eye. In a normal, relaxed eye, parallel light rays entering the eye will be focused on the retina, as shown in Figure 6-16. For very distant objects, the light rays coming from the object will be nearly parallel and the image of the object will be focused on the retina of the relaxed eye.

If the eyeball is too long, parallel light rays will come to a focus in front of the retina. For this eyeball, the most distant object that will be in focus on the retina of the relaxed eye will be at a distance less than infinity. In this case, the eye is said to be nearsighted. The condition is called *myopia*.

If the eyeball is too short, the focus of parallel light rays will be behind the retina. The eye is then said to be farsighted. This condition is called *hyperopia*.

Another defect arises when the surface of the cornea is not spherical. It may be more sharply curved along one great circle than along another. This leads to a condition called *astigmatism*. Astigmatism makes it impossible to focus clearly on horizontal and vertical lines at the same time.

The conditions of myopia, hyperopia, and astigmatism may all be alleviated by the use of corrective lenses.

Another defect of vision, which does not arise from improper relation of the different parts of the eye is *color blindness*. Color blindness, also called color vision deficiency, involves abnormalities that cause a person to be unable to distinguish certain colors, or to perceive colors differently than most people. Color blindness arises from inherited defects in the pigment in cone cells in the retina. It may take on a wide range of degrees of severity, from very mild to a situation in which the eye sees only shades of gray. Color blindness is a lifelong condition. It may disqualify people from certain occupations.

Laboratory

In this laboratory, you will set up and operate circuitry for a silicon PIN photodiode and will use the circuitry to measure chopped HeNe laser light and argon laser light and will determine the relative response of the detector system at several wavelengths.

Materials

Photodiode (Centronic OSD100-5T or equivalent)

Operational amplifier (National Semiconductor LF356 or equivalent)

Electric motor with toothed chopper wheel (Laser Precision CTX-534 Variable Speed Optical Chopper or equivalent)

Helium-neon laser (few milliwatt output)

Resistors (selection of values, kilohms to megohms)

Neutral-density filters (selection of different values, with total neutral density at least 4)

Oscilloscope

Power meter (Spectra-Physics model 405 or equivalent)

Argon ion laser (line tunable, with at least 4 visible wavelengths available)

DC voltage source

PROCEDURE

The first part of the *Procedure* will involve fabrication of a circuit to operate a photodiode as a photovoltaic detector. You will use the circuit to measure chopped laser light and to measure the responsivity of the photodiode.

First, you will set up circuitry for using the photodiode in the photovoltaic mode. Figure 6-19 shows the experimental arrangement. In this arrangement, the photodiode operates as a photovoltaic detector.

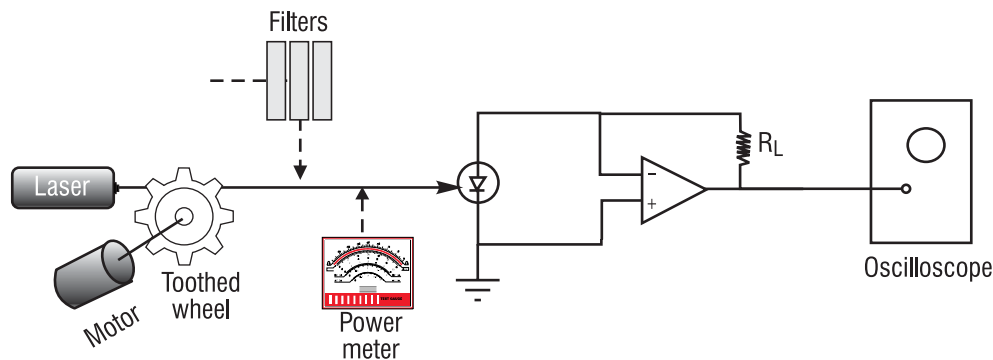


Figure 6-19 Experimental arrangement for measurements with photodiode operated in a photovoltaic mode

The toothed wheel is mounted on the electric motor. When it rotates, it chops the light. That is, it periodically blocks the HeNe laser light from reaching the detector. The speed of the motor should be adjusted so that the light is blocked 1000 times per second. This is a standard measurement condition. If the wheel has 10 teeth, for example, the motor should rotate at 100 revolutions per second.

Assemble the circuit as shown in Figure 6-19. The load resistor should be much smaller than the value of the shunt resistance of the photodiode, which is specified by the manufacturer. The output of the circuit will be hooked to the input of the oscilloscope.

The oscilloscope screen should show a square wave with a frequency of 1000 Hertz. Use the voltage calibration of the oscilloscope to measure the peak voltage of the signal. Then insert the power meter into the laser beam in front of the photodiode and measure the power in the beam. Calculate the responsivity of the detector and compare it to the manufacturer's specification. Remove the power meter.

Next, check the linearity of the detector response by inserting neutral-density filters into the path of the beam as indicated. Gradually increase the number of neutral-density filters and

OPTICAL DETECTORS AND HUMAN VISION

record the total neutral density and the peak voltage at each step. Increase the neutral density to at least 4, increasing the sensitivity of the oscilloscope display as necessary. Plot the peak voltage as a function of neutral density on semilogarithmic paper. The plot should be a straight line.

Next, you will operate the photodiode in the photoconductive mode. The experimental arrangement is shown in Figure 6-20. Note that a DC voltage source is added in this figure.

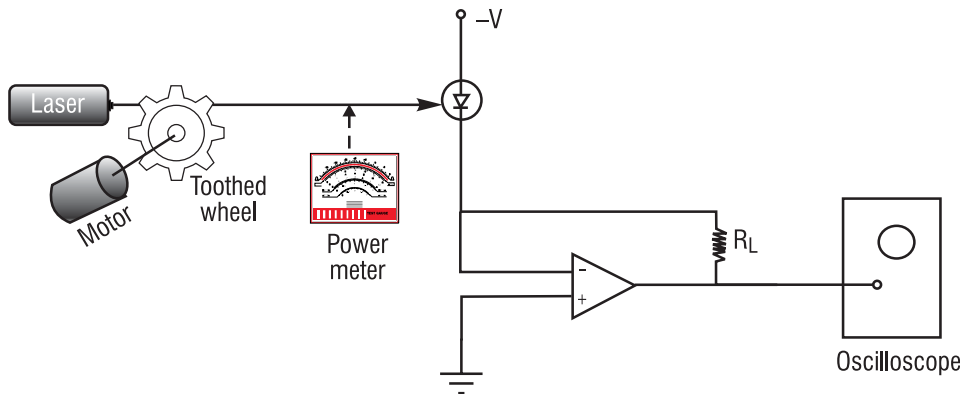


Figure 6-20 Experimental arrangement for measurements with photodiode operated in a photoconductive mode

Hook up the circuit as shown in Figure 6-20. The load resistor should be relatively large, in the megohm regime. The output of the circuit is used as the input of the oscilloscope.

The output on the oscilloscope screen should be a 1000-Hz square wave. Use the voltage calibration of the oscilloscope to measure the peak voltage of the signal. Then insert the power meter into the laser beam in front of the photodiode and measure the power in the beam. Calculate the responsivity of the detector and compare it to the manufacturer's specification.

Next, you will investigate the effect of varying the load resistor. Remove the power meter and change the value of the load resistor. Use several different values of load resistor, and for each one record the value of the peak signal. Plot the peak signal as a function of the value of the load resistor. How does the signal vary with load resistance?

Now you will measure the responsivity as a function of wavelength. One measurement is already available, at 633 nm. Use the line-tunable argon laser to obtain values for at least four different visible wavelengths.

Replace the helium-neon laser in Figure 6-20 with the argon laser. Replace the load resistor with the same value that was used for the responsivity measurement at 633 nm. For each of four different argon laser wavelengths, measure the peak voltage on the oscilloscope and the laser power reading with the power meter in the same way that you measured them at 633 nm. If the argon laser power is too high, insert neutral-density filters in front of the photodiode and the power meter to reduce it to an appropriate value. Record the results and calculate the responsivity for each wavelength. Plot the responsivity as a function of wavelength and compare the result to the manufacturer's specification.

DATA TABLE

Detector responsivity measurement (photovoltaic mode)

Voltage measurement _____

Laser power _____

Calculated detector responsivity _____

Manufacturer's quoted responsivity _____

Linearity measurement

Measurement number	Neutral density	Voltage
1.	_____	_____
2.	_____	_____
3.	_____	_____
4.	_____	_____

(Plot the results on semilog paper.)

Detector responsivity measurement (photoconductive mode)

Voltage measurement _____

Laser power _____

Calculated detector responsivity _____

Manufacturer's quoted responsivity _____

Effect of load resistor

Measurement number	Load resistor	Signal voltage
1.	_____	_____
2.	_____	_____
3.	_____	_____
4.	_____	_____

(Plot the results.)

Responsivity vs wavelength

Measurement number	Voltage	Power	Calculated responsivity	Manufacturer's responsivity
1.	_____	_____	_____	_____
2.	_____	_____	_____	_____
3.	_____	_____	_____	_____
4.	_____	_____	_____	_____

(Plot the results, including the value for 633 nm obtained earlier.)

Problems

1. Define detector responsivity, noise equivalent power, quantum efficiency, detectivity, and rise time.
2. Define sources of detector noise, including shot noise, Johnson noise, $1/f$ noise, and photon noise. Explain methods employed to reduce these noise sources in the detection of optical radiation.
3. Describe and explain important types of photodetectors, including photon detectors, thermal detectors, photoemissive detectors, photoconductive detectors, photovoltaic detectors, and photomultiplier detectors. Describe the spectral response of each type.
4. Draw and explain a typical circuit used with a photovoltaic detector.
5. Draw and explain a typical circuit used with a photoconductive detector.
6. Describe concepts related to human vision, including structure of the eye, the formation of images by the eye, and common defects of vision.
7. With an irradiance of 0.001 W/cm^2 incident on a detector of area 0.5 cm^2 and with a bandwidth of 2 Hz, the ratio of the noise voltage to the signal voltage is 10. What is the noise equivalent power of the detector?
8. A detector has a noise equivalent power of $5 \times 10^{-10} \text{ watts/(Hz)}^{1/2}$ and an area of 0.2 cm^2 . What is its value of D^* (detectivity)?
9. A detector has a responsivity of 0.12 ampere per watt at a wavelength of $1.06 \mu\text{m}$. What is the quantum efficiency of the detector?
10. A laser beam with total power of 22 watts is incident at normal incidence on a Lambertian surface. How much power reaches a detector with an area of 0.1 cm^2 at an angle of 22 degrees located at a distance of 50 cm from where the beam strikes the surface?

References

- J. D. Vincent, *Fundamentals of Infrared Detector Operation and Testing*, Wiley, New York, 1990.
- P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, Wiley, New York, 1962, Chapters 8–10.
- E. L. Dereniak and D. G. Crowe, *Optical Radiation Detectors*, Wiley, New York, 1984.

Optical Waveguides and Fibers

Ajoy Ghatak and K. Thyagarajan
Department of Physics
Indian Institute of Technology
New Delhi, India

The dramatic reduction of transmission loss in optical fibers coupled with equally important developments in the area of light sources and detectors has brought about a phenomenal growth of the fiber optic industry during the past two decades. The birth of optical fiber communication coincided with the fabrication of low-loss optical fibers and room-temperature operation of semiconductor lasers in 1970. Ever since, the scientific and technological progress in this field has been so phenomenal that we are already in the fifth generation of optical fiber communication systems within a brief span of 30 years. Recent developments in optical amplifiers and wavelength division multiplexing (WDM) are taking us to a communication system with almost “zero” loss and “infinite” bandwidth. Indeed, optical fiber communication systems are fulfilling the increased demand on communication links, especially with the proliferation of the Internet. This module, *Optical Waveguides and Fibers*, is an introduction to the basics of fiber optics, discussing especially the characteristics of optical fibers as regards their application to telecommunication (to be covered in Module 1.8, *Fiber Optic Telecommunication*) and fiber optic sensors.

Prerequisites

Prior to the module, you are expected to have covered Modules 1-1, *Nature and Properties of Light*; Module 1-2, *Light Sources and Laser Safety*; Module 1-3, *Basic Geometrical Optics*; and Module 1-4, *Basic Physical Optics*. In addition, you should be able to manipulate and use algebraic formulas involving trigonometric functions and deal with units.

Objectives

The objective of the module is to present the basic characteristics of optical fibers that are important from the point of view of their applications in telecommunication and sensing. Characteristics such as attenuation, pulse dispersion, single-mode and multimode fibers, graded-index, fibers and zero-dispersion wavelength are covered.

When you finish this module, you will be able to

- Describe how light is guided through optical fibers.
- Differentiate between multimode and single-mode fibers.
- Calculate the numerical aperture (NA), intermodal dispersion, and material dispersion.
- Understand zero-dispersion wavelength and dispersion-shifted fibers.
- Know the importance of plastic optical fibers with regard to their application in communication.
- Describe how optical fibers can be used in sensing.

Fiber Optics as a Career

Although not apparent, fiber optics plays a very significant role in our daily lives. When we make long-distance or international telephone calls or browse through the Internet, we are communicating through optical fibers. Without a backbone of an optical fiber network, it would be almost impossible to have all the current facilities of communication and information transfer. The entire world is being "fibered." Having a basic understanding of the guiding properties of optical fibers and recognizing their applications in communication and sensing can provide a person with an added advantage in terms of job prospects in many different areas. Fiber optics is being used in diverse applications such as cold lighting in museums, construction of smart civil structures, and networking for communication. Among many issues associated with day-to-day handling of optical fibers, a fiber optic technician may have to solve problems regarding joining of optical fibers, location of fiber breaks, and fiber end preparation. The field of fiber optics is as exciting as ever and continues to expand even faster than microelectronics. The true age of optics, in which fiber optics is a very important part, is here and is expected to contribute in making the 21st century the century of photonics.

Abbreviations

CSF	Conventional single-mode fiber
DSF	Dispersion-shifted fiber
FOG	Fiber optic gyroscope
LD	Laser diode
LED	Light-emitting diode
MMF	Multimode fiber
NA	Numerical aperture
nm	Nanometer
Pa	Pascal
PIF	Parabolic-index fiber
POF	Plastic optical fiber
ps	Picosecond
SIF	Step-index fiber
TIR	Total internal reflection
WDM	Wavelength-division multiplexing

Opening Demonstrations

Fiber optics is an exciting field, and the demonstration exercises that follow are sure to kindle one's interest in this fascinating area.

- 1. Demonstration of light guidance through optical fibers.** Take a 2-meter length or so of optical fiber—preferably a large-diameter plastic optical fiber—and make perpendicular cuts at both the ends. Couple light from a laser pointer into one of the ends and see for

yourself how the light is guided through the fiber. If the room is dark, you should be able to observe the entire fiber glowing all along its length due to scattering (Rayleigh scattering). Observe the light emerging from the other end of the fiber. Try tying the fibers in a knot and you will notice that light still is guided, in spite of the severe bend of the fiber. If you use a long reel of the optical fiber, you should still see the emerging light at the other end and the glow along the entire reel. Figure 7-1 shows a photograph of light propagating through a long optical fiber.

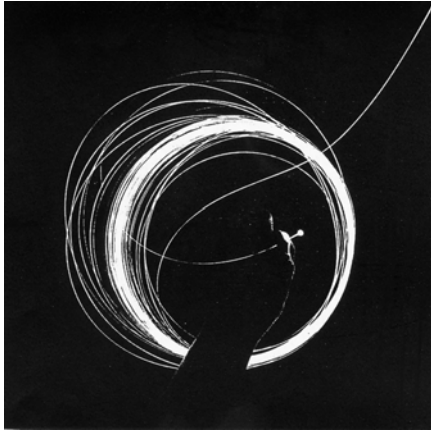


Figure 7-1 A long, thin optical fiber transmitting a light beam
(Photograph courtesy Dr. Chynoweth)

2. **Differentiate between two kinds of fibers.** Take short pieces (~ 2 m) of a multimode optical fiber and a single-mode optical fiber. Focus the light from a HeNe laser onto the input end of the fiber using a lens such as a microscope objective, and observe the output by projecting it onto a white screen. The output from the multimode fiber will show a speckle pattern while that from a single-mode fiber will show a very uniform beam much like the output from a good laser. In the former case, the light propagates via a large number of modes, and the interference between the various modes causes the speckle pattern.
3. **Simple demonstration of sensing property of optical fibers.** Perform the previous demonstration using a multimode fiber reel by fixing the entire setup on a table and observing the output speckle pattern as you bring your hand close to the optical fiber reel. The warmth of the hand is enough to change the phase difference between the interfering modes, which in turn changes the interference pattern. The speckle pattern will keep changing as you move your hand closer to and farther from the reel. If you try to press or twist the fiber, you will immediately notice the change in the speckle pattern. Such changes in the pattern due to external influence demonstrate in a very simple fashion the possibility of using optical fibers for sensing.
4. **Demonstration of Rayleigh scattering in an optical fiber.** Rayleigh scattering is responsible for the blue color of the sky and the red color of the setting sun. It is very interesting to demonstrate the wavelength dependence of Rayleigh scattering using a long optical fiber. Couple white light from a lamp, such as a tungsten halogen lamp emitting white light, into a 1-km length or so of multimode optical fiber and look into the output.

Notice the color of the light. Cut the fiber, leaving about 1 m from the input end of the fiber, and repeat the experiment with this 1 m of fiber. You will see that, in the former case, the emerging light looks reddish while in the latter case it looks white. This difference is due to the *decrease of loss with increase in wavelength* due to Rayleigh scattering. Thus, shorter wavelengths toward the blue region have suffered greater scattering out of the fiber than longer wavelengths near the red region. Even though all wavelengths at the input end are coupled, there is more power in the red part at the output end, thereby giving the light a reddish color.

Basic Concepts

I. HISTORICAL INTRODUCTION

Communication implies transfer of information from one point to another. When it is necessary to transmit information, such as speech, images, or data, over a distance, one generally uses the concept of carrier wave communication. In such a system, the information to be sent modulates an electromagnetic wave such as a radio wave, microwave, or light wave, which acts as a carrier. This modulated wave is then transmitted to the receiver through a channel and the receiver demodulates it to retrieve the imprinted signal. The carrier frequencies associated with TV broadcast ($\sim 50\text{--}900\text{ MHz}$) are much higher than those associated with AM radio broadcast ($\sim 600\text{ kHz--}20\text{ MHz}$). This is due to the fact that, in any communication system employing electromagnetic waves as the carrier, **the amount of information that can be sent increases as the frequency of the carrier is increased.**¹ Obviously, TV broadcast has to carry much more information than AM broadcasts. Since optical beams have frequencies in the range of 10^{14} to 10^{15} Hz, the use of such beams as the carrier would imply a tremendously large increase in the information-transmission capacity of the system as compared to systems employing radio waves or microwaves.

In a conventional telephone hookup, voice signals are converted into equivalent electrical signals by the microphone and are transmitted as electrical currents through metallic (copper or aluminum) wires to the local telephone exchange. Thereafter, these signals continue to travel as electric currents through metallic wire cable (or for long-distance transmission as radio/microwaves to another telephone exchange) usually with several repeaters in between. From the local area telephone exchange, at the receiving end, these signals travel via metallic wire pairs to the receiver telephone, where they are converted back into corresponding sound waves. Through such cabled wire-pair telecommunication systems, one can at most send 48 simultaneous telephone conversations intelligibly. On the other hand, in an optical communication system that uses glass fibers as the transmission medium and light waves as carrier waves, it is distinctly possible today to have 35,000 or more simultaneous telephone

¹ The information-carrying capacity of an electromagnetic carrier is approximately proportional to the difference between the maximum and the minimum frequencies (technically known as *bandwidth* of the channel) that can be transmitted through the communication channel. The higher one goes in the electromagnetic spectrum in frequency scale, the higher the bandwidth and hence the information-carrying capacity of such a communication system. That is why historically the trend in carrier wave communication has been always toward bandwidths of higher and higher frequencies.

conversations (equivalent to a transmission speed of about 2.5 Gbit/s) through one glass fiber no thicker than a human hair. This large information-carrying capacity of a light beam is what generated interest among communication engineers and caused them to explore the possibility of developing a communication system using light waves as carrier waves.

The idea of using light waves for communication can be traced as far back as 1880 when Alexander Graham Bell invented the photophone (see Figure 7-2) shortly after he invented the telephone in 1876. In this remarkable experiment, speech was transmitted by modulating a light beam, which traveled through air to the receiver. The flexible reflecting diaphragm (which could be activated by sound) was illuminated by sunlight. The reflected light was received by a parabolic reflector placed at a distance of about 200 m. The parabolic reflector concentrated the light on a photoconducting selenium cell, which formed a part of a circuit with a battery and a receiving earphone. Sound waves present in the vicinity of the diaphragm vibrated the diaphragm, which led to a consequent variation of the light reflected by the diaphragm. The variation of the light falling on the selenium cell changed the electrical conductivity of the cell, which in turn changed the current in the electrical circuit. This changing current reproduced the sound on the earphone.

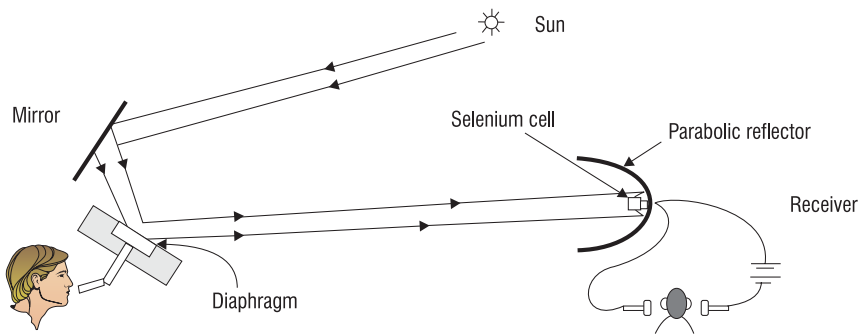


Figure 7-2 Schematic of the photophone invented by Bell. In this system, sunlight was modulated by a vibrating diaphragm and transmitted through a distance of about 200 meters in air to a receiver containing a selenium cell connected to the earphone.

After succeeding in transmitting a voice signal over 200 meters using a light signal, Bell wrote to his father: “I have heard a ray of light laugh and sing. We may talk by light to any visible distance without any conducting wire.” To quote from Maclean: “In 1880 he (Graham Bell) produced his ‘photophone’ which to the end of his life, he insisted was ‘... the greatest invention I have ever made, greater than the telephone...’ Unlike the telephone, though, it had no commercial value.”

The modern impetus for telecommunication with carrier waves at optical frequencies owes its origin to the discovery of the laser in 1960. Earlier, no suitable light source was available that could reliably be used as the information carrier.² At around the same time, telecommunication traffic was growing very rapidly. It was conceivable then that conventional telecommunication systems based on, say, coaxial cables, radio and microwave links, and wire-pair cable, could soon reach a saturation point. The advent of lasers immediately triggered a great deal of

² We may mention here that, although incoherent sources like light-emitting diodes (LED) are also often used in present-day optical communication systems, it was discovery of the laser that triggered serious interest in the development of optical communication systems.

investigation aimed at examining the possibility of building optical analogues of conventional communication systems. The very first such modern optical communication experiments involved laser beam transmission through the atmosphere. However, it was soon realized that shorter-wavelength laser beams could not be sent in open atmosphere through reasonably long distances to carry signals, unlike, for example, the longer-wavelength microwave or radio systems. This is due to the fact that a laser light beam (of wavelength about $1\text{ }\mu\text{m}$) is severely attenuated and distorted owing to scattering and absorption by the atmosphere. Thus, for reliable light-wave communication under terrestrial environments it would be necessary to provide a “guiding” medium that could protect the signal-carrying light beam from the vagaries of the terrestrial atmosphere. This guiding medium is the optical fiber, a hair-thin structure that guides the light beam from one place to another as was shown in Figure 7-1. The guidance of the light beam through the optical fiber takes place because of total internal reflection, which we discuss in the following section.

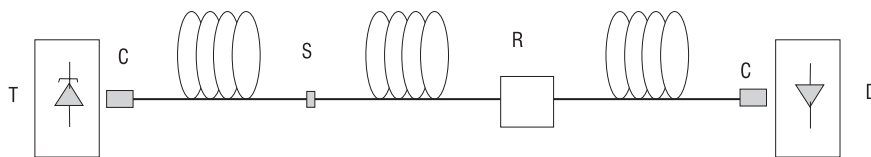


Figure 7-3 A typical fiber optic communication system: T, transmitter; C, connector; S, splice; R, repeater; D, detector

In addition to the capability of carrying a huge amount of information, optical fibers fabricated with recently developed technology are characterized by extremely low losses³ ($< 0.2\text{ dB/km}$), as a consequence of which the distance between two consecutive repeaters (used for amplifying and reshaping the attenuated signals) could be as large as 250 km. We should perhaps mention here that it was the epoch-making paper of Kao and Hockham in 1966 that suggested that optical fibers based on silica glass could provide the necessary transmission medium if metallic and other impurities could be removed. Indeed, this 1966 paper triggered the beginning of serious research in developing low-loss optical fibers. In 1970, Kapron, Keck, and Maurer (at Corning Glass in USA) were successful in producing silica fibers with a loss of about 17 dB/km at a wavelength of 633 nm . (Kapron, Keck, and Maurer) Since then, the technology has advanced with tremendous rapidity. By 1985 glass fibers were routinely produced with extremely low losses ($< 0.2\text{ dB/km}$). Figure 7-3 shows a typical optical fiber communication system. It consists of a transmitter, which could be either a laser diode or an LED, the light from which is coupled into an optical fiber. Along the path of the optical fiber are splices, which are permanent joints between sections of fibers, and repeaters that boost the signal and correct any distortion that may have occurred along the path of the fiber. At the end of the link, the light is detected by a photodetector and electronically processed to retrieve the signal.

II. TOTAL INTERNAL REFLECTION (TIR)

At the heart of an optical communication system is the optical fiber that acts as the transmission channel carrying the light beam loaded with information. As mentioned earlier, the guidance of

³ Attenuation is usually measured in decibels (dB). We define attenuation in Section VI.

the light beam (through the optical fiber) takes place because of the phenomenon of total internal reflection (TIR), which we will now discuss. You learned about critical angles, TIR, etc. in Module 1-3, *Basic Geometrical Optics*. You need now to refresh your memory and apply these ideas more directly to the physics of optical fibers. We first define the refractive index (n) of a medium:

$$n = \frac{c}{v} \quad (7-1)$$

where c ($\approx 3 \times 10^8$ m/s) is the speed of light in free space and v represents the velocity of light in that medium. For example, for light waves, $n \approx 1.5$ for glass and $n \approx 1.33$ for water.

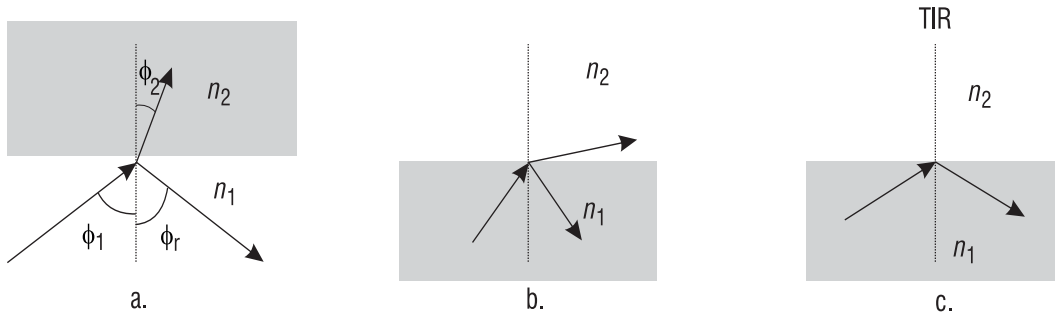


Figure 7-4 (a) A ray of light incident on a denser medium ($n_2 > n_1$). (b) A ray incident on a rarer medium ($n_2 < n_1$). (c) For $n_2 < n_1$, if the angle of incidence is greater than critical angle, it will undergo total internal reflection.

As you know, when a ray of light is incident at the interface of two media (like air and glass), the ray undergoes partial reflection and partial refraction as shown in Figure 7-4a. The vertical dotted line represents the normal to the surface. The angles ϕ_1 , ϕ_2 , and ϕ_r represent the angles that the incident ray, refracted ray, and reflected ray make with the normal. According to Snell's law and the law of reflection,

$$n_1 \sin \phi_1 = n_2 \sin \phi_2 \quad \text{and} \quad \phi_1 = \phi_2 \quad (7-2)$$

Further, the incident ray, reflected ray, and refracted ray lie in the same plane. In Figure 7-4a, since $n_2 > n_1$ we must have (from Snell's law) $\phi_2 < \phi_1$, i.e., the ray will bend toward the normal. On the other hand, if a ray is incident at the interface of a rarer medium ($n_2 < n_1$), the ray will bend away from the normal (see Figure 7-4b). The angle of incidence, for which the angle of refraction is 90° , is known as the critical angle and is denoted by ϕ_c . Thus, when

$$\phi_1 = \phi_c = \sin^{-1} \left(\frac{n_2}{n_1} \right) \quad (7-3)$$

$\phi_2 = 90^\circ$. When the angle of incidence exceeds the critical angle (i.e., when $\phi_1 > \phi_c$), there is no refracted ray and we have total internal reflection (see Figure 7-4c).

Example 7-1

For the glass-air interface, $n_1 = 1.5$, $n_2 = 1.0$, and the critical angle is given by

$$\phi_c = \sin^{-1} \left(\frac{1}{1.5} \right) \approx 41.8^\circ$$

On the other hand, for the glass-water interface, $n_1 = 1.5$, $n_2 = 1.33$, and

$$\phi_c = \sin^{-1} \left(\frac{1.33}{1.5} \right) \approx 62.5^\circ.$$

The phenomenon of total internal reflection can be very easily demonstrated through a simple experiment as shown in Figure 7-5. A thick, semicircular glass disc is immersed in a glass vessel filled with water. A laser beam from a HeNe laser or a laser pointer is directed toward the center of the semicircular disc so that it is incident normally on the glass surface and goes undeviated toward O as shown in the figure. The angle of incidence (at the water-glass interface) is increased by rotating the glass disc about the point O , always keeping the incident ray normal to the circular disc. Eventually, when the angle of incidence exceeds the critical angle ($\approx 62.5^\circ$), the laser beam undergoes TIR—as shown in Figure 7-5. The ray trace can be clearly seen when viewed from the top. If one puts a drop of ink in the water, the light path becomes very beautiful to look at! The experiment is very simple and we urge the reader to carry it out using a laser pointer.

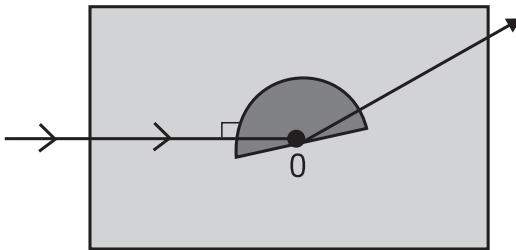


Figure 7-5 A simple laboratory experiment to demonstrate the phenomenon of TIR

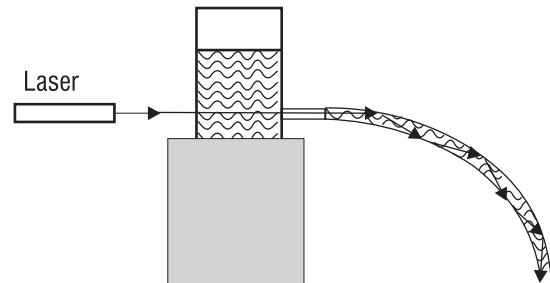


Figure 7-6 The concept of Tyndall's original experiment demonstrating the phenomenon of light guidance by TIR

We should mention that the phenomenon of guidance by multiple total internal reflections was first demonstrated by John Tyndall in 1854. In this demonstration, Tyndall showed that light travels along the curved path of water emanating from an illuminated vessel (see Figure 7-6). Actually, light undergoes TIR at the water-air interface, along the water path, as shown in Figure 7-6.

III. THE OPTICAL FIBER

Figure 7-7a shows an optical fiber, which consists of a (cylindrical) central dielectric core clad by a material of slightly lower refractive index. The corresponding refractive index distribution (in the transverse direction) is given by:

$$\begin{aligned} n &= n_1 & \text{for } r < a \\ n &= n_2 & \text{for } r > a \end{aligned} \quad (7-4)$$

where n_1 and n_2 ($< n_1$) represent respectively the refractive indices of core and cladding and a represents the radius of the core. We define a parameter Δ through the following equations.

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_2^2} \quad (7-5)$$

When $\Delta \ll 1$ (as is indeed true for silica fibers where n_1 is very nearly equal to n_2) we may write

$$\Delta = \frac{(n_1 + n_2)(n_1 - n_2)}{2n_1^2} \approx \frac{(n_1 - n_2)}{n_1} \approx \frac{(n_1 - n_2)}{n_2} \quad (7-6)$$

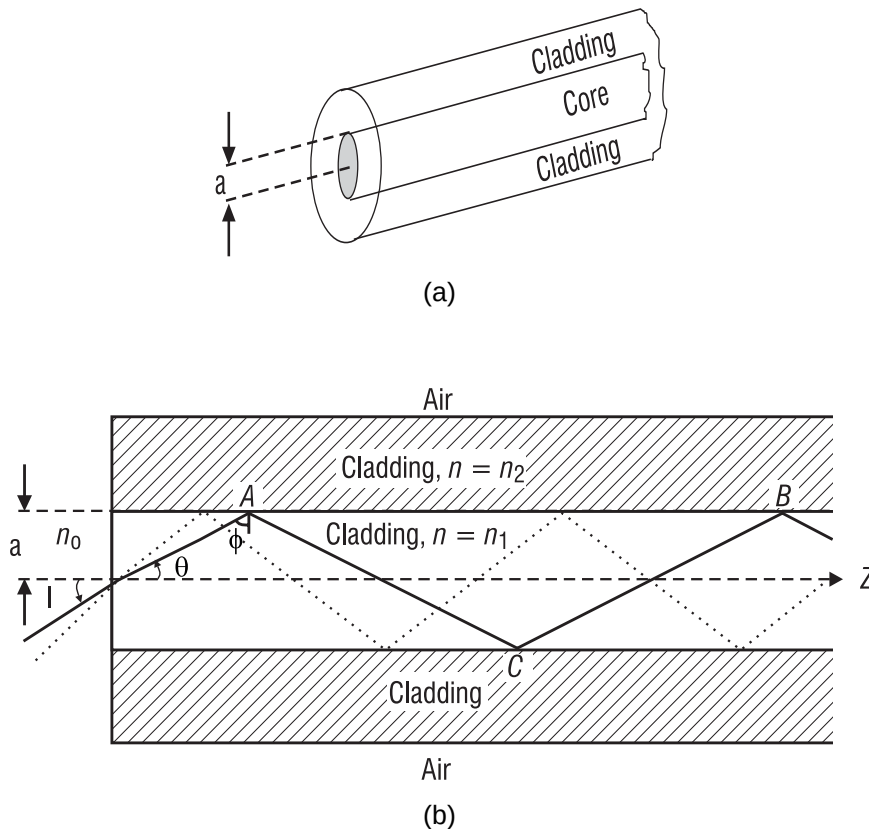


Figure 7-7 (a) A glass fiber consists of a cylindrical central core clad by a material of slightly lower refractive index. (b) Light rays impinging on the core-cladding interface at an angle greater than the critical angle are trapped inside the core of the fiber.

For a typical (multimode) fiber, $a \approx 25 \mu\text{m}$, $n_2 \approx 1.45$ (pure silica), and $\Delta \approx 0.01$, giving a core index of $n_1 \approx 1.465$. The cladding is usually pure silica while the core is usually silica doped with germanium. Doping by germanium results in a typical increase of refractive index from n_2 to n_1 .

Now, for a ray entering the fiber core at its end, if the angle of incidence ϕ at the internal core-cladding interface is greater than the critical angle $\phi_c [= \sin^{-1}(n_2/n_1)]$, the ray will undergo TIR at that interface. Further, because of the cylindrical symmetry in the fiber structure, this ray will suffer TIR at the lower interface also and therefore be guided through the core by repeated total internal reflections. Even for a bent fiber, light guidance can occur through multiple total internal reflections (see Figures 7-1 and 7-6).

The necessity of a clad fiber (Figure 7-7)—rather than a bare fiber with no cladding—is clear. For transmission of light from one place to another, the fiber must be supported. Supporting structures, however, may considerably distort the fiber, thereby affecting the guidance of the light wave. This is avoided by choosing a sufficiently thick cladding. Further, in a fiber bundle, in the absence of the cladding, light can leak through from one fiber to another. The idea of adding a second layer of glass (namely, the cladding) came (in 1955) from Hopkins and Kapany in the United Kingdom. However, during that time the use of optical fibers was mainly in image transmission rather than in communication. Indeed, the early pioneering works in fiber optics (in the 1950s) were by Hopkins and Kapany in the United Kingdom and by Van Heel in Holland. Their work led to the use of the fibers in optical devices and medical instruments.

We may mention here that the retina of the human eye (see Module 1-2, *Light Sources and Safety*, and Module 1-6, *Optical Detectors and Human Vision*) consists of a large number of rods and cones that have the same kind of structure as the optical fiber. They consist of dielectric cylindrical rods surrounded by another dielectric of slightly lower refractive index. The core diameters are in the range of a few microns. The light absorbed in these “light guides” generates electrical signals, which are then transmitted to the brain through various nerves.

It is interesting to know why optical fibers are made of glass. Quoting from Professor W.A. Gambling, a pioneer in the field of fiber optics:

We note that glass is a remarkable material which has been in use in ‘pure’ form for at least 9000 years. The compositions remained relatively unchanged for millennia and its uses have been widespread. The three most important properties of glass which makes it of unprecedented value are:

1. First, there is a wide range of accessible temperatures where its viscosity is variable. The viscosity can be controlled—unlike most materials, like water and metals, which remain liquid until they are cooled down to their freezing temperatures and then suddenly become solid. Glass, on the other hand, does not solidify at a discrete freezing temperature but gradually becomes stiffer and stiffer and eventually becomes hard. In the transition region it can be easily drawn into a thin fiber.
2. The second most important property is that highly pure silica is characterized with extremely low light transmission loss; i.e., it is highly transparent. Today, in most commercially available silica fibers, 96% of the power gets transmitted after propagating through 1 km of optical fiber. This indeed represents a truly remarkable achievement.

3. The third most remarkable property is the intrinsic strength of glass. It can support a stress of about $2,000,000 \text{ lb/in}^2$ so that a glass fiber of the type used in the telephone network with a diameter ($125 \text{ }\mu\text{m}$) of twice the thickness of a human hair, can support a load of 40 pounds.

IV. THE COHERENT BUNDLE

A large number of fibers put together form what is known as a bundle. If the fibers are not aligned, i.e. they are all jumbled up, the bundle is said to be an *incoherent bundle*. However, if the fibers are aligned properly, i.e., if the relative positions of the fibers in the input and output ends are the same, the bundle is said to be a *coherent bundle*. If a particular fiber in this coherent bundle is illuminated at one of its ends, there will be a bright spot at the other end of the **same** fiber. Thus a coherent bundle will transmit the image from one end to another (see Figure 7-8). On the other hand, in an incoherent bundle the output image will be scrambled. Because of this property, an incoherent bundle can be used as a coder; the transmitted image can be decoded by using a similar bundle at the output end. In a bundle, since there can be hundreds of thousands of fibers, decoding without the original bundle configuration would be extremely difficult. Incoherent bundles are also used in illumination such as in traffic lights or road signs [see, e.g., “Schott Is Lighting the Way Home”] or even for lighting applications in buildings wherein the light source is removed from relatively inaccessible areas and fibers are used to guide light from a lamp [see, e.g., “Fiber Optic Technology Put to Work—Big Time”]. They can also be used as cold light sources (i.e., light sources giving only light and no heat) by cutting off the heat radiation with a filter at the input to the fiber bundle. The light emerging from the bundle is also free from UV radiation and is suitable for illumination of paintings in museums.

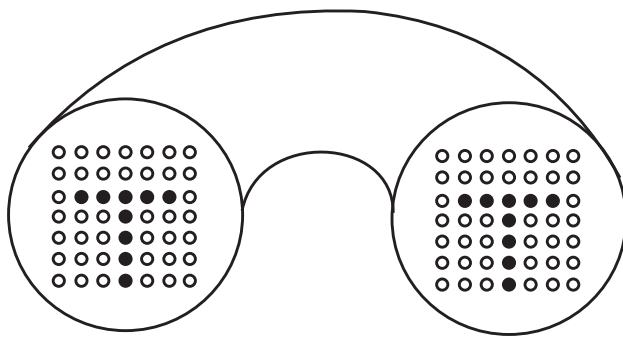


Figure 7-8 A bundle of aligned fibers. A bright (or dark) spot at the input end of the coherent fiber bundle produces a bright (or dark) spot at the output end. Thus, an image will be transmitted (in the form of bright and dark spots) through a bundle of aligned fibers, forming, for example, the letter T shown above.

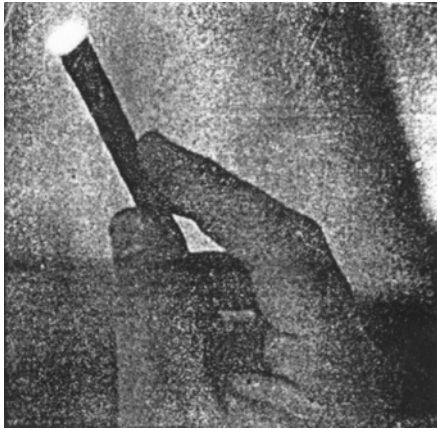


Figure 7-9 (a) An optical fiber medical probe called an endoscope enables doctors to examine the inner parts of the human body. (b) A stomach ulcer as seen through an endoscope [Photographs courtesy United States Information Service, New Delhi]

Perhaps the most important application of a coherent bundle is in a fiber optic endoscope where it can be inserted inside a human body and the interior of the body can be viewed from outside. For illuminating the portion that is to be seen, the bundle is enclosed in a sheath of fibers that carry light from outside to the interior of the body (see Figure 7-9). A state-of-the-art fiberscope can have about 10,000 fibers, which would form a coherent bundle of about 1 mm in diameter, capable of resolving objects 70 μm across. Fiber optic bundles can also be used for viewing internal machine parts that are otherwise inaccessible.

V. THE NUMERICAL APERTURE (NA)

We return to Figure 7-7b and consider a ray that is incident on the entrance face of the fiber core, making an angle i with the fiber axis. Let the refracted ray make an angle θ with the same axis. Assuming the outside medium to have a refractive index n_0 (which for most practical cases is unity), we get

$$\frac{\sin i}{\sin \theta} = \frac{n_1}{n_0} \quad (7-7)$$

Obviously, if this refracted ray is to suffer total internal reflection at the core-cladding interface, the angle of incidence ϕ must satisfy the equation,

$$\sin \phi (= \cos \theta) > \frac{n_2}{n_1} \quad (7-8)$$

Since $\sin \theta = \sqrt{1 - \cos^2 \theta}$, we will have

$$\sin \theta < \left[1 - \left(\frac{n_2}{n_1} \right)^2 \right]^{\frac{1}{2}} \quad (7-9)$$

Let i_m represent the maximum half-angle of the acceptance cone for rays at the input end. Applying Snell's law at the input end and using Equations 7-5, 7-7, and 7-9, we must have $I < i_m$, where

$$\sin i_m = (n_1^2 - n_2^2)^{1/2} = n_1 \sqrt{2\Delta} \quad (7-10)$$

and we have assumed $n_0 = 1$; i.e., the outside medium is assumed to be air. Thus, if a cone of light is incident on one end of the fiber, it will be guided through it provided the half-angle of the cone is less than i_m . This half-angle is a measure of the light-gathering power of the fiber. We define the numerical aperture (NA)—see Module 1-3, *Basic Geometrical Optics*—of the fiber by the following equation:

$$NA = \sin i_m = \sqrt{n_1^2 - n_2^2} = n_1 \sqrt{2\Delta} \quad (7-11)$$

Example 7-2

For a typical step-index (multimode) fiber with $n_1 \approx 1.45$ and $\Delta \approx 0.01$, we get

$$\sin i_m = n_1 \sqrt{2\Delta} = 1.45 \sqrt{2 \times (0.01)} = 0.205$$

so that $i_m \approx 12^\circ$. Thus, all light entering the fiber must be within a cone of half-angle 12° .

In a short length of an optical fiber, if all rays between $i = 0$ and i_m are launched, the light coming out of the fiber will also appear as a cone of half-angle i_m emanating from the fiber end. If we now allow this beam to fall normally on a white paper and measure its diameter, we can easily calculate the NA of the fiber (Laboratory B, “NA of a multimode optical fiber”).

VI. ATTENUATION IN OPTICAL FIBERS

Attenuation and pulse dispersion represent the two most important characteristics of an optical fiber that determine the information-carrying capacity of a fiber optic communication system. Obviously, the lower the attenuation (and similarly, the lower the dispersion) the greater can be the required repeater spacing and therefore the lower will be the cost of the communication system. Pulse dispersion will be discussed in the next section, while in this section we will discuss briefly the various attenuation mechanisms in an optical fiber.

The attenuation of an optical beam is usually measured in decibels (dB). If an input power P_1 results in an output power P_2 , the power loss α in decibels is given by

$$\alpha \text{ (dB)} = 10 \log_{10} (P_1/P_2) \quad (7-12)$$

Thus, if the output power is only half the input power, the loss is $10 \log 2 \approx 3$ dB. Similarly, if the power reduction is by a factor of 100 or 10, the power loss is 20 dB or 10 dB respectively. If

96% of the light is transmitted through the fiber, the loss is about 0.18 dB. On the other hand, in a typical fiber amplifier, a power amplification of about 1000 represents a power gain of 30 dB.

Figure 7-10 shows the spectral dependence of fiber *attenuation* (i.e., *loss coefficient per unit length*) as a function of wavelength of a typical silica optical fiber. The losses are caused by various mechanisms such as Rayleigh scattering, absorption due to metallic impurities and water in the fiber, and intrinsic absorption by the silica molecule itself. The Rayleigh scattering loss varies as $1/\lambda_0^4$, i.e., shorter wavelengths scatter more than longer wavelengths. Here λ_0 represents the free space wavelength. This is why the loss coefficient decreases up to about 1550 nm. The two absorption peaks around 1240 nm and 1380 nm are primarily due to traces of OH^- ions and traces of metallic ions. For example, even 1 part per million (ppm) of iron can cause a loss of about 0.68 dB/km at 1100 nm. Similarly, a concentration of 1 ppm of OH^- ion can cause a loss of 4 dB/km at 1380 nm. This shows the level of purity that is required to achieve low-loss optical fibers. If these impurities are removed, the two absorption peaks will disappear. For $\lambda_0 > 1600$ nm the increase in the loss coefficient is due to the absorption of infrared light by silica molecules. This is an intrinsic property of silica, and no amount of purification can remove this infrared absorption tail.

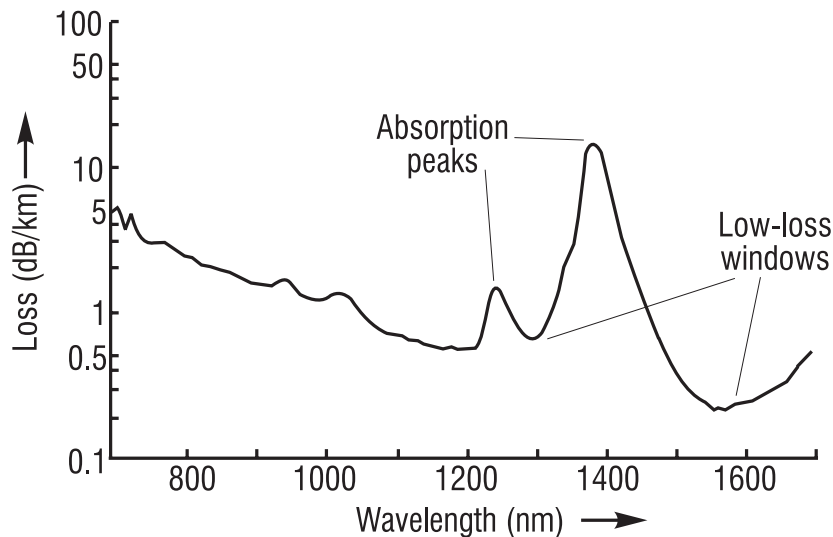


Figure 7-10 Typical wavelength dependence of attenuation for a silica fiber. Notice that the lowest attenuation occurs at 1550 nm [adapted from Miya, Hasaka, and Miyashita].

As you see, there are two windows at which loss attains its minimum value. The first window is around 1300 nm (with a typical loss coefficient of less than 1 dB/km) where, fortunately (as we will see later), the material dispersion is negligible. However, the loss attains its absolute minimum value of about 0.2 dB/km around 1550 nm. The latter window has become extremely important in view of the availability of erbium-doped fiber amplifiers.

Example 7-3

Calculation of losses using the dB scale become easy. For example, if we have a 40-km fiber link (with a loss of 0.4 dB/km) having 3 connectors in its path and if each connector has a loss of 1.8 dB, the total loss will be the sum of all the losses in dB; or $0.4 \text{ dB/km} \times 40 \text{ km} + 3 \times 1.8 \text{ dB} = 21.4 \text{ dB}$.

Example 7-4

Let us assume that the input power of a 5-mW laser decreases to $30 \mu\text{W}$ after traversing through 40 km of an optical fiber. Using Equation 7-12, attenuation of the fiber in dB/km is therefore $[10 \log (166.7)]/40 \approx 0.56 \text{ dB/km}$.

VII. PULSE DISPERSION IN STEP-INDEX FIBERS (SIF)

In digital communication systems, information to be sent is first coded in the form of pulses and these pulses of light are then transmitted from the transmitter to the receiver, where the information is decoded. The larger the number of pulses that can be sent per unit time and still be resolvable at the receiver end, the larger will be the transmission capacity of the system. A pulse of light sent into a fiber broadens in time as it propagates through the fiber. This phenomenon is known as *pulse dispersion*, and it occurs primarily because of the following mechanisms:

1. Different rays take different times to propagate through a given length of the fiber. We will discuss this for a step-index multimode fiber and for a parabolic-index fiber in this and the following sections. In the language of wave optics, this is known as *intermodal dispersion* because it arises due to different modes traveling with different speeds.⁴
2. Any given light source emits over a range of wavelengths, and, because of the intrinsic property of the material of the fiber, different wavelengths take different amounts of time to propagate along the same path. This is known as *material dispersion* and will be discussed in Section IX.
3. Apart from intermodal and material dispersions, there is yet another mechanism—referred to as *waveguide dispersion* and important only in single-mode fibers. We will briefly discuss this in Section XI.

In the fiber shown in Figure 7-7, the rays making larger angles with the axis (those shown as dotted rays) have to traverse a longer optical path length and therefore take a longer time to reach the output end. Consequently, the pulse broadens as it propagates through the fiber (see

⁴ We will have a very brief discussion about modes in Section XI. We may mention here that the number of modes in a step index fiber is about $V^2/2$ where the parameter V will be defined in Section XI. When $V < 4$, the fiber supports only a few modes and it is necessary to use wave theory. On the other hand, if $V > 8$, the fiber supports many modes and the fiber is referred to as a multimoded fiber. When this happens, ray optics gives accurate results and one need not use wave theory.

OPTICAL WAVEGUIDES AND FIBERS

Figure 7-11). Even though two pulses may be well resolved at the input end, because of the broadening of the pulses they may not be so at the output end. Where the output pulses are not resolvable, no information can be retrieved. Thus, the smaller the pulse dispersion, the greater will be the information-carrying capacity of the system.

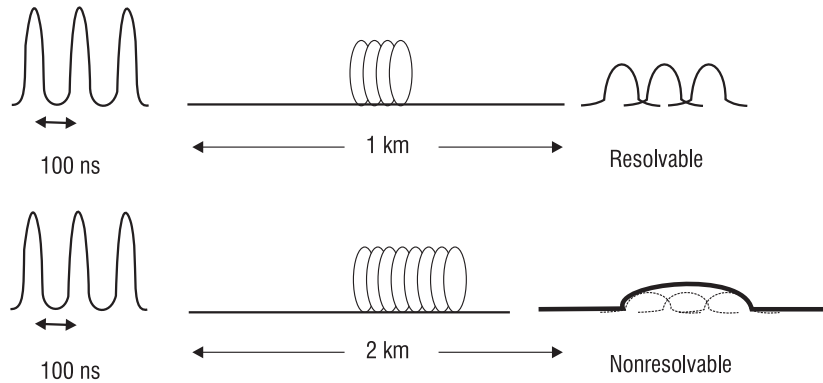


Figure 7-11 Pulses separated by 100 ns at the input end would be resolvable at the output end of 1 km of the fiber. The same pulses would not be resolvable at the output end of 2 km of the same fiber.

We will now derive an expression for the *intermodal dispersion* for a step-index fiber. Referring back to Figure 7-7b, for a ray making an angle θ with the axis, the distance AB is traversed in time.

$$t_{AB} = \frac{AC + CB}{c/n_1} = \frac{AB/\cos\theta}{c/n_1} \quad (7-13)$$

or

$$t_{AB} = \frac{n_1(AB)}{c \cos\theta} \quad (7-14)$$

where c/n_1 represents the speed of light in a medium of refractive index n_1 , c being the speed of light in free space. Since the ray path will repeat itself, the time taken by a ray to traverse a length L of the fiber would be

$$t_L = \frac{n_1 L}{c \cos\theta} \quad (7-15)$$

The above expression shows that the time taken by a ray is a function of the angle θ made by the ray with the z -axis (fiber axis), which leads to pulse dispersion. If we assume that all rays lying between $\theta = 0$ and $\theta = \theta_c = \cos^{-1}(n_2/n_1)$ (see Equation 7-8) are present, the time taken by The following extreme rays for a fiber of length L would be given by

$$t_{\min} = \frac{n_1 L}{c} \quad \text{corresponding to rays at } \theta = 0 \quad (7-16)$$

$$\boxed{t_{\max} = \frac{n_1^2 L}{cn_2}} \quad \text{corresponding to rays at } \theta = \theta_c = \cos^{-1}(n_2/n_1) \quad (7-17)$$

Hence, if all the input rays were excited simultaneously, the rays would occupy a time interval at the output end of duration

$$\tau_i = t_{\max} - t_{\min} = \frac{n_1 L}{c} \left[\frac{n_1}{n_2} - 1 \right]$$

or, finally, the intermodal dispersion in a multimode SIF is

$$\boxed{\tau_i \cong \frac{n_1 L}{c} \Delta \approx \frac{L}{2n_1 c} (NA)^2} \quad (7-18)$$

where Δ has been defined earlier [see Equations 7-5 and 7-6] and we have used Equation 7-11. The quantity τ_i represents the pulse dispersion due to different rays taking different times in propagating through the fiber, which, in wave optics, is nothing but the intermodal dispersion and hence the subscript i . Note that the pulse dispersion is proportional to the square of NA . Thus, to have a smaller dispersion, one must have a smaller NA , which of course reduces the acceptance angle i_m and hence the light-gathering power. If, at the input end of the fiber, we have a pulse of width τ_1 , after propagating through a length L of the fiber the pulse will have a width τ_2 given approximately by

$$\boxed{\tau_2^2 = \tau_1^2 + \tau_i^2} \quad (7-19)$$

Example 7-5

For a typical (multimoded) step-index fiber, if we assume $n_1 = 1.5$, $\Delta = 0.01$, $L = 1$ km, we would get

$$\boxed{\tau_i = \frac{1.5 \times 1000}{3 \times 10^8} \times 0.01 = 50 \text{ ns/km}} \quad (7-20)$$

That is, a pulse traversing through the fiber of length 1 km will be broadened by 50 ns. Thus, two pulses separated by, say, 500 ns at the input end will be quite resolvable at the end of 1 km of the fiber. However, if consecutive pulses were separated by, say, 10 ns at the input end, they would be absolutely unresolvable at the output end. Hence, in a 1-Mbit/s fiber optic system, where we have one pulse every 10^{-6} s, a 50-ns/km dispersion would require repeaters to be placed every 3 to 4 km. On the other hand, in a 1-Gbit/s fiber optic communication system, which requires the transmission of one pulse every 10^{-9} s, a dispersion of 50 ns/km would result in intolerable broadening even within 50 meters or so. This would be highly inefficient and uneconomical from a system point of view.

From the discussion in the above example it follows that, for a very-high-information-carrying system, it is necessary to reduce the pulse dispersion. Two alternative solutions exist—one involves the use of near-parabolic-index fibers and the other involves single-mode fibers. We look at these next.

VIII. PARABOLIC-INDEX FIBERS (PIF)

In a step-index fiber such as that pictured in Figure 7-7, the refractive index of the core has a constant value. By contrast, in a parabolic-index fiber, the refractive index in the core decreases continuously (in a quadratic fashion) from a maximum value at the center of the core to a constant value at the core-cladding interface. The refractive index variation is given by

$$\begin{array}{ll} n^2(r) = n_1^2 \left[1 - 2\Delta \left(\frac{r}{a} \right)^2 \right] & 0 < r < a \quad \text{core} \\ n^2(r) = n_2^2 = n_1^2 (1 - 2\Delta) & r > a \quad \text{cladding} \end{array} \quad (7-21)$$

with Δ as defined in Equation 7-5. For a typical (multimode) parabolic-index silica fiber, $\Delta \approx 0.01$, $n_2 \approx 1.45$, and $a \approx 25 \mu\text{m}$. On the other hand, for a typical plastic fiber (see Section XII), $n_1 \approx 1.49$, $n_2 \approx 1.40$, and $a \approx 500 \mu\text{m}$.

Since the refractive index decreases as one moves away from the center of the core, a ray entering the fiber is continuously bent toward the axis of the fiber, as depicted in Figure 7-12.

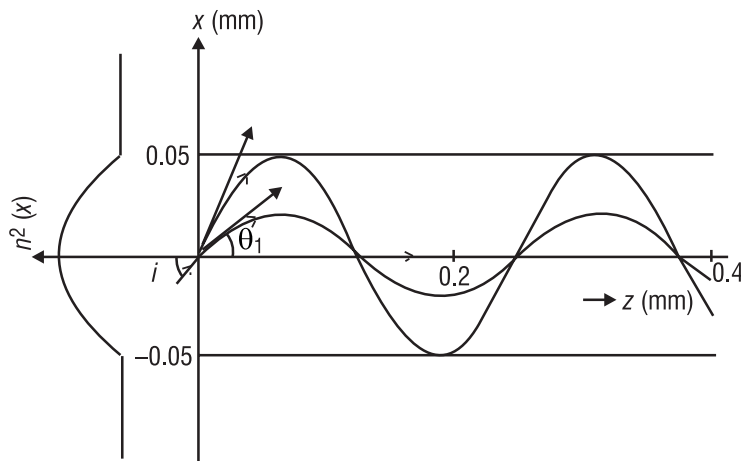


Figure 7-12 Different ray paths in a parabolic-index fiber

This follows from Snell's law because the ray continuously encounters a medium of lower refractive index and hence bends continuously away from the normal. Even though rays making larger angles with the fiber axis traverse a longer path, they do so now in a region of lower refractive index (and hence greater speed). The longer path length is almost compensated for by a greater average speed such that all rays take approximately the same amount of time in traversing the fiber. This leads to a much smaller pulse dispersion. The detailed calculations are

a bit involved [see, e.g., Chapters 4 and 5, Ghatak and Thyagarajan]. The final result for the intermodal dispersion in a parabolic-index fiber (PIF) is given by

$$\tau_{\text{im}} = \frac{n_2 L}{2c} \left(\frac{n_1 - n_2}{n_2} \right)^2 \approx \frac{n_2 L}{2c} \Delta^2 \approx \frac{L}{8cn_1^3} (NA)^4 \quad \text{Pulse dispersion in multimode PIF} \quad (7-22)$$

Note that, as compared to a step-index fiber, the pulse dispersion is proportional to the fourth power of NA . For a typical (multimode parabolic-index) fiber with $n_2 \approx 1.45$ and $\Delta \approx 0.01$, we would get

$$\tau_{\text{im}} \approx 0.25 \text{ ns/km} \quad (7-23)$$

Comparing with Equation 7-20 we find that for a parabolic-index fiber the pulse dispersion is reduced by a factor of about 200 in comparison to a step-index fiber. This is why first- and second-generation optical communication systems used near-parabolic-index fibers. To further decrease the pulse dispersion, it is necessary to use single-mode fibers, which will be discussed in Section XI. However, in all fiber optic systems we will have material dispersion (which is a characteristic of the material itself and not of the waveguide). We will discuss material dispersion in the following section.

IX. MATERIAL DISPERSION

We first define the *group index*. To do this we return to Equation 7-1 where we noted that the velocity of light in a medium is given by

$$v = c/n \quad (7-24)$$

Here n is the refractive index of the medium, which, in general, depends on the wavelength. The dependence of the refractive index on wavelength leads to what is known as *dispersion*, discussed in Module 1.3, *Basic Geometrical Optics*. In Figure 7-13 we have shown a narrow pencil of a white light beam incident on a prism. Since the refractive index of glass depends on the wavelength, the angle of refraction will be different for different colors. For example, for crown glass the refractive indices at 656.3 nm (orange), 589.0 nm (yellow), and 486.1 nm (green) are respectively given by 1.5244, 1.5270, and 1.5330. Thus, if the angle of incidence $i = 45^\circ$ the angle of refraction, r , will be $r = 27.64^\circ$, 27.58° , and 27.47° for the orange, yellow, and blue colors respectively. The incident white light will therefore *disperse* into its constituent colors—the dispersion will become more evident at the second surface of the prism as seen in Figure 7-13.

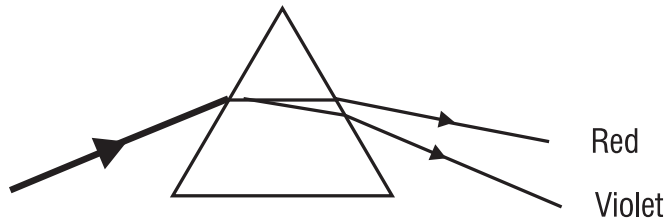


Figure 7-13 Dispersion of white light as it passes through a prism

Now, the quantity v defined by Equation 7-24 is usually referred to as the *phase velocity*. However, a pulse travels with what is known as the *group velocity*, which is given by

$$v_g = c/n_g \quad (7-25)$$

where n_g is known as the group index and, in most cases its value is slightly larger than n . In Table 7.1 we have tabulated n and n_g for pure silica for different values of wavelength lying between 700 nm and 1600 nm. The corresponding spectral variation of n and n_g for pure silica is given in Figure 7-14.

Table 7.1 Values of n , n_g , and D_m for pure silica

λ_0 (nm)	$n(\lambda_0)$	$n_g(\lambda_0)$	D_m (ps/nm-km)
700	1.45561	1.47154	-172.902
750	1.45456	1.46924	-135.313
800	1.45364	1.46744	-106.609
850	1.45282	1.46601	-84.2077
900	1.45208	1.46489	-66.382
950	1.45139	1.46401	-51.9441
1000	1.45075	1.46332	-40.0577
1050	1.45013	1.46279	-30.1214
1100	1.44954	1.46241	-21.6951
1150	1.44896	1.46214	-14.4511
1200	1.44839	1.46197	-8.14213
1250	1.44783	1.46189	-2.57872
1300	1.44726	1.46189	2.38579
1350	1.44670	1.46196	6.86631
1400	1.44613	1.46209	10.9539
1450	1.44556	1.46229	14.7211
1500	1.44498	1.46253	18.2268
1550	1.44439	1.46283	21.5187
1600	1.44379	1.46318	24.6358

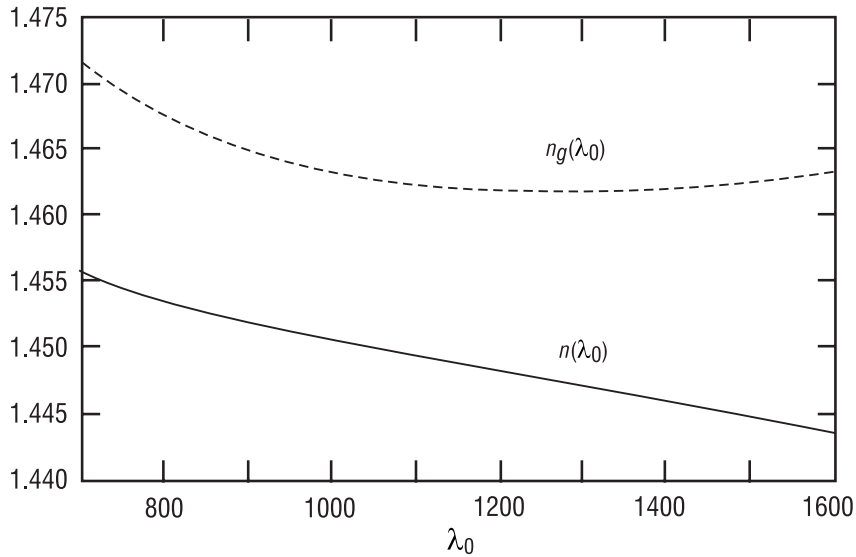


Figure 7-14 Variation of n and n_g with wavelength for pure silica. Notice that n_g has a minimum value around 1270 nm.

In Sections VIII and IX we considered the broadening of an optical pulse due to different rays taking different amounts of time to propagate through a certain length of the fiber. However, every source of light has a certain wavelength spread, which is often referred to as the spectral width of the source. Thus, a white light source (like the sun) has a spectral width of about 300 nm. On the other hand, an LED has a spectral width of about 25 nm and a typical laser diode (LD) operating at 1300 nm has a spectral width of about 2 nm or less. Each wavelength component will travel with a slightly different group velocity through the fiber. This results in broadening of a pulse. This broadening of the pulse is proportional to the length of the fiber and to the spectral width of the source. We define the *material dispersion coefficient* D_m , which is measured in ps/km-nm. D_m represents the material dispersion in picoseconds per kilometer length of the fiber per nanometer spectral width of the source. At a particular wavelength, the value of D_m is a characteristic of the material and is (almost) the same for *all* silica fibers. The values of D_m for different wavelengths are tabulated in Table 7.1. A negative D_m implies that the longer wavelengths travel faster; similarly, a positive value of D_m implies that shorter wavelengths travel faster. However, in calculating the pulse broadening, only the magnitude should be considered.

Example 7-6

The LEDs used in early optical communication systems had a spectral width $\Delta\lambda_0$ of about 20 nm around $\lambda_0 = 825$ nm. Using D_m in Table 7.1 (at 850 nm), such a pulse will broaden by

$$\tau_m = D_m \times L \times \Delta\lambda = 84.2 \text{ (ps/km-nm)} \times 1 \text{ (km)} \times 20 \text{ (nm)} \sim 1700 \text{ ps} = 1.7 \text{ ns}$$

in traversing a 1-km length of the fiber. It is very interesting to note that, if we carry out a similar calculation around $\lambda_0 \approx 1300$ nm, we will obtain a much smaller value of τ_m ; thus

$$\tau_m = D_m \times L \times \Delta\lambda = 2.4 \text{ (ps/km-nm)} \times 1 \text{ (km)} \times 20 \text{ (nm)} \sim 0.05 \text{ ns}$$

in traversing 1-km length of the fiber. The very small value of τ_m is due to the fact that n_g is approximately constant around $\lambda_0 = 1300$ nm, as shown in Figure 7-14. Indeed, the wavelength $\lambda_0 \approx 1270$ nm is usually referred to as the zero material dispersion wavelength, and it is because of such low material dispersion that the optical communication systems shifted their operation to around $\lambda_0 \approx 1300$ nm.

Example 7-7

In the optical communication systems that are in operation today, one uses laser diodes (LD) with $\lambda_0 \approx 1550$ nm having a spectral width of about 2 nm. Thus, for a 1-km length of the fiber, the material dispersion τ_m becomes

$$\tau_m = D_m \times L \times \Delta\lambda = 21.5 \text{ (ps/km-nm)} \times 1 \text{ (km)} \times 2 \text{ (nm)} \sim 43 \text{ ps}$$

the positive sign indicating that higher wavelengths travel more slowly than lower wavelengths. [Notice from Table 7.1 that, for $\lambda_0 \geq 1300$ nm, n_g increases with λ_0 .]

X. DISPERSION AND MAXIMUM BIT RATE

We may mention here briefly that, in a digital communication system employing light pulses, pulse broadening would result in an overlap of pulses, resulting in loss of resolution and leading to errors in detection. Thus pulse broadening is one of the mechanisms (other than attenuation) that limits the distance between two repeaters in a fiber optic link. It is obvious that, the larger the pulse broadening, the smaller will be the number of pulses per second that can be sent down a link. Different criteria based on slightly different considerations are used to estimate the maximum permissible bit rate ($B_{\max}$) for a given pulse dispersion. However, it is always of the order of $1/\tau$, where τ is the pulse dispersion. In one type of extensively used coding (known as NRZ) we have

$$B_{\max} \approx 0.7/\tau \quad (7-26)$$

This formula takes into account (approximately) only the limitation imposed by the pulse dispersion in the fiber. In an actual link the source and detector characteristics would also be taken into account while estimating the maximum bit rate (see Module 1-8, *Fiber Optic Telecommunication*). It should also be pointed out that, in a fiber, the pulse dispersion is caused, in general, by intermodal dispersion, material dispersion, and waveguide dispersion. However, waveguide dispersion is important only in single-mode fibers and may be neglected in carrying out analysis for multimode fibers. Thus (considering multimode fibers), if τ_i and τ_m are the dispersion due to intermodal and material dispersions respectively, the total dispersion is given by

$$\tau = \sqrt{\tau_i^2 + \tau_m^2} \quad (7-27)$$

Example 7-8

We consider a step-index multimode fiber with $n_1 = 1.46$, $\Delta = 0.01$, operating at 850 nm. For such a fiber, the intermodal dispersion (for a 1-km length) is

$$\tau_i = \frac{n_1 L \Delta}{c} \approx \frac{1.46 \times 1000 \times 0.01}{3 \times 10^8} \approx 49 \text{ ns}$$

which is usually written as

$$\tau_i \approx 49 \text{ ns/km}$$

If the source is an LED with $\Delta\lambda = 20 \text{ nm}$, using Table 7.1, the material dispersion τ_m is 1.7 ns/km [see Example 7-6]. Thus, in step-index multimode fibers, the dominant pulse-broadening mechanism is intermodal dispersion and the total dispersion is given by

$$\tau = \sqrt{\tau_i^2 + \tau_m^2} = 49 \text{ ns/km}$$

Using Equation 7-26, this gives a maximum bit rate of about

$$B_{\max} \approx 0.7/\tau \approx 14 \text{ Mbit-km/s}$$

Thus a 10-km link can at most support 1.4 Mbit/s, since τ increases by a factor of 10, causing $B_{\max}$ to decrease by the same factor.

Example 7-9

Let us now consider a parabolic-index multimode fiber with $n_1 = 1.46$, $\Delta = 0.01$, operating at 850 nm with an LED of spectral width 20 nm. For such a fiber, the intermodal dispersion, using Equation 7-22, is

$$\tau_{\text{im}} = \frac{n_1}{2c} \Delta^2 L \approx 0.24 \text{ ns/km}$$

The material dispersion is again 1.7 ns/km. Thus, in this case the dominant mechanism is material dispersion rather than intermodal dispersion. The total dispersion is

$$\tau = \sqrt{0.24^2 + 1.7^2} = 1.72 \text{ ns/km}$$

This gives a maximum bit rate of about

$$B_{\max} \approx 0.7/\tau \approx 400 \text{ Mbit-km/s}$$

giving a maximum permissible bit rate of 20 Mbit/s for a 20-km link.

Example 7-10

If we now shift the wavelength of operation to 1300 nm and use the parabolic-index fiber of the previous example, we see that the intermodal dispersion remains the same at 0.24 ns/km while the material dispersion (for an LED of $\Delta\lambda_0 = 20 \text{ nm}$) becomes 0.05 ns/km (see Example 7-6). The

material dispersion is now negligible in comparison to intermodal dispersion. Thus the total dispersion and maximum bit rate are respectively given by

$$\tau \approx \sqrt{0.24^2 + 0.05^2} \approx 0.25 \text{ ns/km}; \Rightarrow B_{\max} = 2.8 \text{ Gbit-km/s}$$

Example 7-11

If, in Example 7-10, we replace the LED with a laser diode of spectral width 2 nm, the material dispersion becomes 0.17 ns/km, which is now smaller than the intermodal dispersion. The total dispersion is

$$\tau = \sqrt{0.24^2 + 0.17^2} = 0.29 \text{ ns/km}$$

giving a maximum bit rate of

$$B_{\max} = 0.7/\tau \approx 2.4 \text{ Gbit-km/s}$$

We should reiterate that, in the examples discussed above, the maximum bit rate has been estimated by considering the fiber only. In an actual link, the temporal response of the source and detector must also be taken into account.

XI. SINGLE-MODE FIBERS

While discussing step-index fibers, we considered light propagation inside the fiber as a set of many rays bouncing back and forth at the core-cladding interface (see Figure 7-7). There the angle θ could take a continuum of values lying between 0 and $\cos^{-1}(n_2/n_1)$, i.e.,

$$0 < \theta < \cos^{-1}(n_2/n_1)$$

For $n_2 = 1.5$ and $\Delta \approx \frac{n_1 - n_2}{n_1} = 0.01$, we would get $n_2/n_1 \simeq$ and $\cos^{-1}(\frac{n_2}{n_1}) = 8.1^\circ$, so

$$0 < \theta < 8.1^\circ$$

Now, when the core radius (or the quantity Δ) becomes very small, ray optics does not remain valid and one has to use the more accurate wave theory based on Maxwell's equations. This wave theory is fairly involved [see, e.g., Chapters 7 and 8 of Ghatak and Thyagarajan]. Here we just give some of the important results for a step-index fiber.

In wave theory, one introduces the parameter

$$V = \frac{2\pi}{\lambda_0} a \sqrt{n_1^2 - n_2^2} = \frac{2\pi}{\lambda_0} a n_1 \sqrt{2\Delta} \approx \frac{2\pi}{\lambda_0} a n_2 \sqrt{2\Delta} \quad (7-28)$$

where Δ has been defined earlier (see Equation 7-5) and $n^1 \simeq n^2$. The quantity V is often referred to as the “*V-number*” or the “*waveguide parameter*” of the fiber. It can be shown that, if

$$V < 2.4045$$

only one guided mode (as if there is only one discrete value of θ) is possible and the fiber is known as a *single-mode fiber*. Further, for a step-index single-mode fiber, the corresponding (discrete) value of θ is approximately given by the following empirical formula

$$\cos \theta \approx 1 - \Delta \left[1 - \left(1.1428 - \frac{0.996}{V} \right)^2 \right] \quad (7-29)$$

We may mention here that because of practical considerations the value of Δ ranges from about 0.002 to about 0.008.

Example 7-12

Consider a step-index fiber (operating at 1300 nm) with $n_2 = 1.447$, $\Delta = 0.003$, and $a = 4.2 \mu\text{m}$. Thus,

$$V = \frac{2\pi}{1.3(\mu\text{m})} \times 4.2(\mu\text{m}) \times 1.447 \times \sqrt{0.006} \approx 2.275$$

Thus the fiber will be single moded and the corresponding value of θ —using Equation 7-29—will be about $\theta = 3.1^\circ$. It may be mentioned that for the given fiber we may write

$$V = \frac{2\pi}{\lambda_0(\mu\text{m})} \times 4.2(\mu\text{m}) \times 1.447 \times \sqrt{0.006} \approx \frac{2.958}{\lambda_0(\mu\text{m})}$$

Thus, for

$$\lambda_0 > 2.958/2.4045 = 1.23 \mu\text{m}$$

which guarantees that $V < 2.4045$, the fiber will be single moded. The wavelength for which $V = 2.4045$ is known as the *cutoff wavelength* and is denoted by λ_c . In this example, $\lambda_c = 1.23 \mu\text{m}$ and the fiber will be single moded for $\lambda_0 > 1.23 \mu\text{m}$.

Example 7-13

For reasons that will be discussed later, the fibers used in current optical communication systems (operating at $1.55 \mu\text{m}$) have a small value of core radius and a large value of Δ . A typical fiber (operating at $\lambda_0 \approx 1.55 \mu\text{m}$) has $n_2 = 1.444$, $\Delta = 0.0075$, and $a = 2.3 \mu\text{m}$. Thus, at $\lambda_0 = 1.55 \mu\text{m}$, the V-number is,

$$V = \frac{2\pi}{1.55(\mu\text{m})} \times 2.3(\mu\text{m}) \times 1.444 \times \sqrt{0.015} \approx 1.649$$

The fiber will be single moded (at $1.55 \mu\text{m}$) with $\theta = 5.9^\circ$. Further, for the given fiber we may write

$$V = \frac{2\pi}{\lambda_0(\mu\text{m})} \times 2.3(\mu\text{m}) \times 1.444 \times \sqrt{0.015} \approx \frac{2.556}{\lambda_0(\mu\text{m})}$$

and therefore the cutoff wavelength will be $\lambda_c = 2.556/2.4045 = 1.06 \mu\text{m}$.

A. Spot size of the fundamental mode

A single-mode fiber supports only one mode that propagates through the fiber. This mode is also referred to as the *fundamental mode* of the fiber. The transverse field distribution associated with the fundamental mode of a single-mode fiber is an extremely important quantity. It determines various important parameters like splice loss at joints, launching efficiencies, bending loss, etc. For most single-mode fibers, the fundamental mode-field distributions can be approximated by a Gaussian function, which may be written in the form

$$\psi(r) = A e^{-r^2/w^2} \quad (7-30)$$

where w is referred to as the spot size of the mode-field pattern. When $r = w$, the value of ψ is equal to $1/e$ of the value A at $r = 0$. For a step-index (single-mode) fiber, one has the following empirical expression for w [Marcuse]:

$$\frac{w}{a} \approx \left(0.65 + \frac{1.619}{V^{3/2}} + \frac{2.879}{V^6} \right); \quad 0.8 < V < 2.5 \quad (7-31)$$

where a is the core radius and V is the V -number given by Equation 7-28. We may mention here that the light coming from a HeNe laser (or from a laser pointer) has a transverse intensity distribution very similar to that coming from a single-mode fiber except that the spot size for the HeNe laser is much larger. The quantity $2w$ is also referred to as the mode-field diameter (MFD) of the fiber and is a very important property of single-mode fibers. In fact, MFD is a more important property than core diameter in the case of single-mode fibers, since it determines the splice loss across a joint, bending loss, dispersion, etc. of single-mode fibers.

Example 7-14

Consider a step-index fiber (operating at 1300 nm) with $n_2 = 1.447$, $\Delta = 0.003$, and $a = 4.2 \mu\text{m}$. For this fiber (see Example 7-12), $V \approx 2.28$. Using Equation 7-31, with $V = 2.28$ and $a = 4.2 \text{ mm}$, one obtains $w \approx 4.8 \mu\text{m}$. The same fiber will have a V -value of 1.908 at $\lambda_0 = 1550 \text{ nm}$, giving a value of the spot size $\approx 5.5 \mu\text{m}$. Thus the spot size increases with wavelength.

Example 7-15

For a step-index fiber (operating at 1550 nm) with $n_2 = 1.444$, $\Delta = 0.0075$, and $a = 2.3 \mu\text{m}$ (see Example 7-13), $V \approx 1.65$, giving $w \approx 3.6 \mu\text{m}$. The same fiber will have a V -value of 1.97 at $\lambda_0 = 1300 \text{ nm}$, giving a value of the spot size $\approx 3.0 \mu\text{m}$.

B. Splice loss due to transverse misalignment

The most common misalignment at a joint between two similar fibers is the transverse misalignment similar to that shown in Figure 7-18. Corresponding to a transverse misalignment of u , the power loss in decibels is given by

$$\alpha \text{ (dB)} \approx 4.34 \left(\frac{u}{w} \right)^2 \quad (7-32)$$

Thus a larger value of w will lead to a greater tolerance to transverse misalignment. For $w \approx 5 \mu\text{m}$, and a transverse offset of $1 \mu\text{m}$, the loss at the joint will be approximately 0.18 dB. On the other hand, for $w \approx 3 \mu\text{m}$, a transverse offset of $1 \mu\text{m}$ will result in a loss of about 0.5 dB.

Example 7-16

Consider a single-mode fiber operating at 1300 nm with $w = 5 \mu\text{m}$. For the splice loss to be below 0.1 dB, we obtain [from Equation 7-32]

$$u < \sqrt{\alpha/4.34} w = \sqrt{0.1/4.34} (5 \mu\text{m}) \approx 0.76 \mu\text{m}$$

Thus we must have $u < 0.76 \mu\text{m}$. For a low-loss joint, the transverse alignment is very critical and connectors for single-mode fibers require precision matching and positioning for achieving low loss.

C. Waveguide dispersion

In Section IX we discussed material dispersion that results from the dependence of the refractive index of the fiber on wavelength. Even if we assume the refractive indices n_1 and n_2 to be independent of λ_0 , the group velocity of each mode *does* depend on the wavelength. This leads to what is known as *waveguide dispersion*. The detailed theory is rather involved [see, e.g., Chapter 10, Ghatak and Thyagarajan]; we may mention here two important points:

1. The waveguide dispersion is usually negative for a given single-mode fiber. The magnitude increases with an increase in wavelength.
2. If the core radius a (of a single-mode fiber) is made smaller and the value of Δ is made larger, the magnitude of the waveguide dispersion increases. Thus we can tailor the waveguide dispersion by changing the refractive index profile.

The following two examples demonstrate how one can tailor the zero-dispersion wavelength by changing the fiber parameters.

Example 7-17

We consider the fiber discussed in Example 7-12 for which $n_2 = 1.447$, $\Delta = 0.003$, and $a = 4.2 \mu\text{m}$. The variations of the waveguide dispersion (τ_w), material dispersion (τ_m), and total dispersion

($\tau_{\text{tot}} = \tau_w + \tau_m$) with λ_0 are shown in Figure 7-15. From the figure it can be seen that the total dispersion passes through zero around $\lambda_0 \approx 1300$ nm. This is known as *zero total-dispersion wavelength* and represents an extremely important parameter.

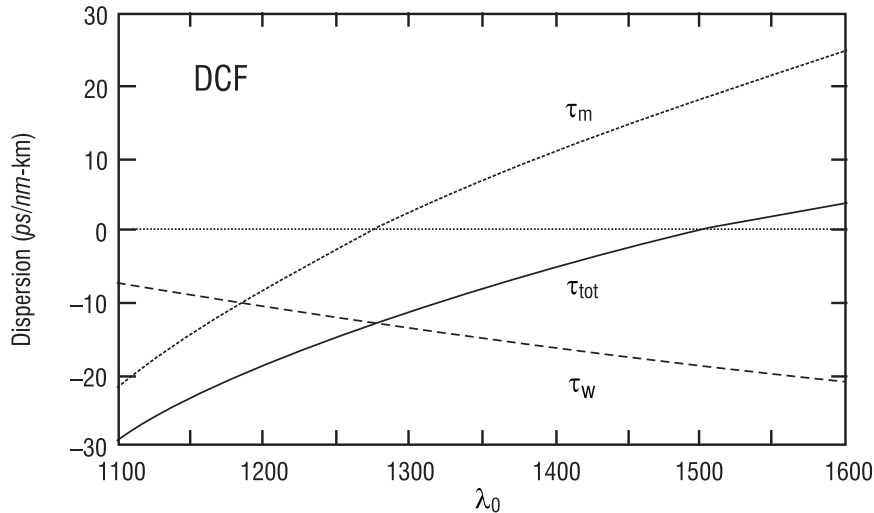


Figure 7-15 The variations of τ_m , τ_w , and τ_{tot} with λ_0 for a typical conventional single-mode fiber (CSF) with parameters given in Example 7-17. The total dispersion passes through zero at around $\lambda_0 \approx 1300$ nm, known as zero total dispersion wavelength.

Example 7-18

We next consider the fiber discussed in Example 7-13 for which $n_2 = 1.444$, $\Delta = 0.0075$, and $a = 2.3$ μm . For this fiber, at $\lambda_0 \approx 1550$ nm,

$\tau_w = -20$ ps/km-nm, as seen in Figure 7-16.

On the other hand, the material dispersion at this wavelength—per km and per unit wavelength interval in nm—is given by Table 7.1 as

$$D_m = \tau_m = +21 \text{ ps/km-nm}$$

We therefore see that the two expressions are of opposite sign and almost cancel each other. Physically, because of waveguide dispersion, longer wavelengths travel more slowly than shorter wavelengths. And, because of material dispersion, longer wavelengths travel faster than shorter wavelengths. So the two effects compensate for each other, resulting in a zero total dispersion around 1550 nm. Thus we have been able to shift the zero-dispersion wavelength by changing the fiber parameters. These are known as the dispersion-shifted fibers, the importance of which will be discussed in the next section. The variations of τ_m , τ_w , and τ_{tot} with λ_0 are plotted in Figure 7-16, showing clearly that τ_{tot} is near zero at $\lambda_0 = 1550$ nm.

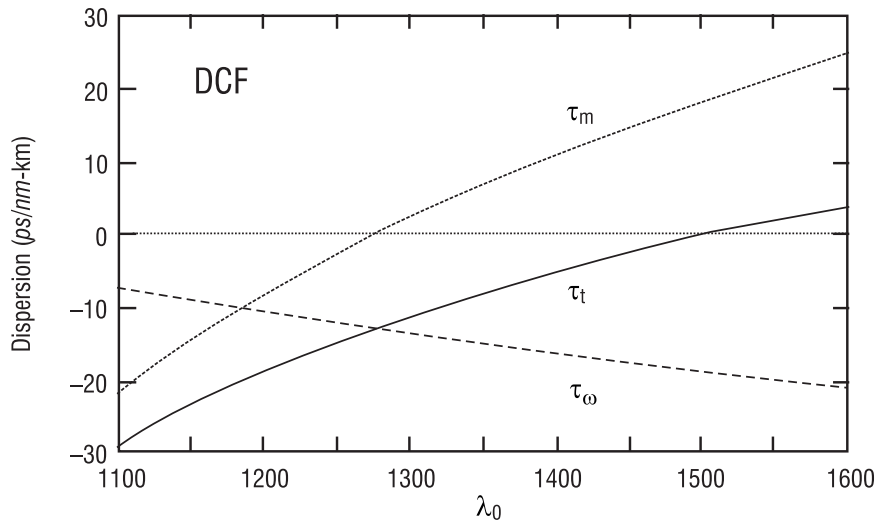


Figure 7-16 The variations of τ_m , τ_w , and τ_{tot} with λ_0 for a typical dispersion-shifted single-mode fiber (DSF) with parameters given in Example 7-18. The total dispersion passes through zero at around $\lambda_0 \approx 1550$ nm.

D. Dispersion-shifted fibers

In Section VI we learned that the attenuation of a silica fiber attains its minimum value of about 0.2 dB/km at around $\lambda_0 \approx 1550$ nm. The second- and third-generation optical communication systems operated around $\lambda_0 \approx 1300$ nm, where the dispersion was extremely small but the loss was about 1 dB/km, and therefore the repeater spacing was limited by the loss in the fiber. Since the lowest loss lies at around $\lambda_0 \approx 1550$ nm, if the zero-dispersion wavelength could be shifted to the $\lambda_0 \approx 1550$ -nm region, one could have both minimum loss and very low dispersion. This would lead to very-high-bandwidth systems with very long (~ 100 km) repeater spacings. Apart from this, extremely efficient optical fiber amplifiers capable of amplifying optical signals in the 1550-nm band have also been developed. Thus, shifting the operating wavelength from 1310 nm to 1550 nm would be very advantageous. As discussed in Example 7-18, by reducing the core size and increasing the value of Δ , we can shift the zero-dispersion wavelength to 1550 nm, which represents the low-loss window. Indeed, the current fourth-generation optical communication systems operate at 1550 nm, using dispersion-shifted single-mode fibers with repeater spacing of about 100 km, carrying about 10 Gbit/s of information (equivalent to about 150,000 telephone channels) through one hair-thin single-mode fiber.

We may mention here that, if one is interested in carrying out accurate calculations for total dispersion, one may use the software described in Ghatak, Goyal, and Varshney.

XII. PLASTIC OPTICAL FIBERS (POF)

Plastic optical fibers made from plastic materials such as polymethyl methacrylate PMMA ($n = 1.49$), polystyrene ($n = 1.59$), polycarbonates ($n = 1.5\text{--}1.57$), fluorinated polymers, and so on. These fibers share the advantages of glass optical fibers in terms of insensitivity to electromagnetic interference, small size and weight, low cost, and potential capability of

carrying information at high rates. The most important attribute of POFs is their large core diameters of around 1 mm as compared to glass fibers with cores of 50 μm or 62.5 μm . Such a large diameter results in easier alignment at joints. They are also more durable and flexible than glass fibers. In addition, they usually have a large NA , resulting in larger light-gathering power (see Section VI).

Plastic optical fibers' performance lies somewhere between conventional copper wires and glass optical fibers. Copper wires are expensive and suffer from electromagnetic interference. By comparison, plastic optical fibers are cheaper and are free from interference. In addition, signals through copper wires can be tapped while it is very difficult to tap signals from optical fibers. Compared to glass fibers, POFs are much easier to connect because of their large diameters. Coupling of light from a source is also very efficient due to large NA and large core diameter. Thus, although glass optical fibers dominate long-distance data communication, POFs are expected to provide *low-cost solutions to short-distance applications* such as local area networks (LAN) and high-speed Internet access. At gigabit rates of transmission, glass fibers are at least 30% more expensive than POFs, while the cost of copper increases dramatically.

Attenuation is one of the important parameters of an optical fiber. Figure 7-17 shows typical loss spectra of 1-mm-diameter step-index (SI) and graded-index (GI) PMMA-based POF. There are three low-loss windows, at 570 nm, 650 nm, and 780 nm. The loss of SI POF at the 650-nm window is about 110 dB/km. This is, of course, very large compared to silica fibers, which have typical losses of about a few dB/km in this wavelength region. The large losses are due to Rayleigh scattering, intrinsic absorption of the material itself, and impurities and absorption due to vibrational modes of the molecules. Because of the high losses, these fibers are used in only short-distance ($\sim$ a few hundred meters) communication links.

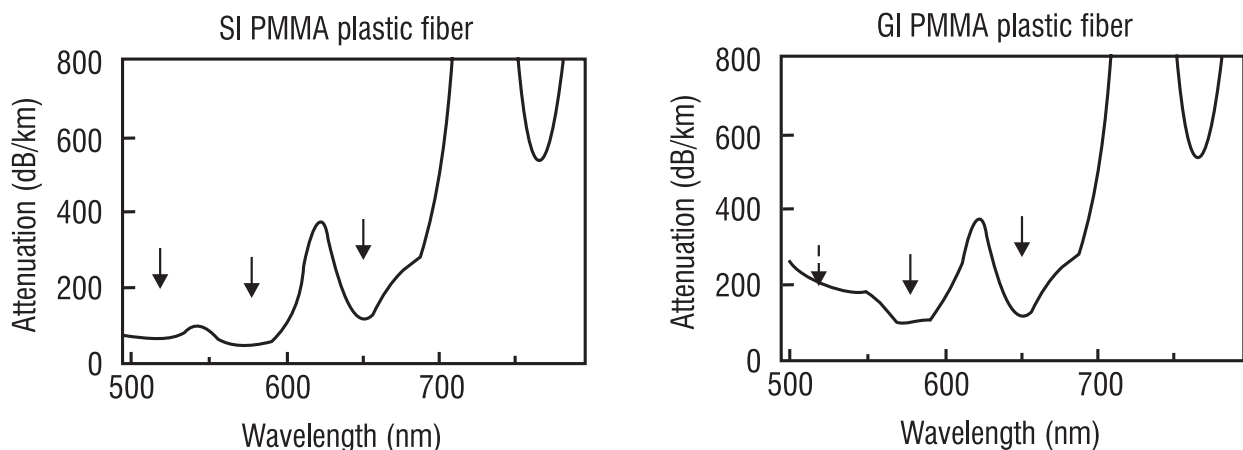


Figure 7-17 Typical attenuation spectra of 1-mm-diameter (a) step-index (SI) and (b) graded-index (GI) PMMA plastic fiber [Adapted from Koeppen, Shi, Chen, and Garito]

Example 7-19

An SI PMMA fiber has a core index of 1.49 and a cladding index of 1.40. Thus the NA [from Equation 7-11] is 0.51 giving an acceptance angle [from Equation 7-11] of 31° . For such a POF, since the value of Δ (in this case ~ 0.06) is not very small, we must use Equation 7-5 and not the approximate expression given by Equation 7-6.

Example 7-20

The coupling efficiency η of light from an LED of an emitting area of radius a_s to a step-index optical fiber of core radius a is given [Keiser] as

$$\eta = \begin{cases} (NA)^2 & \text{for } a > a_s \\ \left(\frac{a}{a_s}\right)^2 (NA)^2 & \text{for } a < a_s \end{cases} \quad (7-33)$$

Thus, if the LED has an emitting radius of $35 \mu\text{m}$ and the fiber has an NA of 0.51 with a core diameter of 1 mm, the coupling efficiency [using the first part of Equation 7-33] will be about 25%. If the total power emitted by the LED is 10 mW, the power coupled will be 2.5 mW. For a typical step-index multimode glass optical fiber with $a \approx 25 \mu\text{m}$ and $NA \approx 0.2$, the corresponding efficiency will be—using the second equation above—about 2%, giving a coupled power of 0.2 mW.

Example 7-21

Any transverse misalignment between the two cores across a joint between two fibers leads to a loss. The efficiency of coupling across a transverse misalignment between two step-index multimode fibers [Keiser] is given by Equation 7-34,

$$\eta = \frac{2}{\pi} \left\{ \cos^{-1}(\tilde{d}) - \tilde{d} \left[1 - (\tilde{d})^2 \right]^{1/2} \right\}; \quad \tilde{d} = \frac{d}{2a} \quad (7-34)$$

where d is the transverse separation between the cores of the two fibers and a is the core radius of the fibers. For misalignments of $d = 0.1 a$ and $0.5 a$, the efficiencies as given by the above equation are 94% and 69% respectively. For a 1-mm-diameter POF, these misalignments correspond to $50 \mu\text{m}$ and $250 \mu\text{m}$ respectively. For a glass fiber of radius $25 \mu\text{m}$, these correspond to $2.5 \mu\text{m}$ and $12.5 \mu\text{m}$ respectively. This shows the much larger tolerance of POFs compared with glass silica fibers.

As discussed in Section VIII, pulse dispersion in multimode fibers can be reduced by using graded-index fibers with a near-parabolic refractive index variation in the core of the fiber.

Using various techniques, near-parabolic-index POFs have been fabricated (Koike, Ishigure, and Nihei). Such fibers can have very small pulse dispersion and hence can be used for high-bandwidth ($\sim$ hundreds of Mbit/s to Gbit/s) transmission over short lengths ($\sim$ 100 m). It is in fact the emergence of such a large-bandwidth graded-index plastic fiber that has spurred its use in many data-transfer applications.

Example 7-22

The coupling efficiency across a transverse misalignment of d between two identical parabolic-index fibers of radii a is given by (see, e. g., Keiser)

$$\eta = \frac{2}{\pi} \left\{ \cos^{-1}(\tilde{d}) - \frac{\tilde{d}}{3} \left[1 - (\tilde{d})^2 \right]^{1/2} (5 - 2\tilde{d}^2) \right\}; \quad \tilde{d} = \frac{d}{2a} \quad (7-35)$$

We can approximate the above equation by $\eta \approx 1 - \frac{16\tilde{d}}{3\pi}$, which is accurate to within 1% for $\tilde{d} < 0.2$. Thus, for $d = 0.1a$ and $0.5a$, the coupling efficiencies would be about 91% and 59% respectively.

Example 7-23

Consider a step-index POF with core and cladding refractive indices given by 1.58 and 1.39 respectively. Using Equation 7-18 for the pulse dispersion in a step-index optical fiber, we obtain a pulse dispersion of 720 ns/km. For a link 100 m long, the dispersion would be 72 ns.

Example 7-24

Consider a parabolic-index POF with $n_1 = 1.490$ and $n_2 = 1.475$ (implying $NA \sim 0.21$ on the axis). We will calculate the pulse dispersion caused by intermodal dispersion only. Using Equation 7-22 we obtain the pulse dispersion as 0.25 ns/km. For a 100-m-long link, the intermodal dispersion would be 25 ps. In addition to the intermodal dispersion, the actual dispersion in such parabolic-index fibers would be determined by material dispersion and also by any mode-coupling effects that may take place in the fiber. In Koike, Ishigure, and Nihei, measurements have shown about 3-GHz bandwidth of graded-index POFs.

Example 7-25

The material dispersion of PMMA is about 0.305 ns/nm-km at 650 nm (Ishigure, Nihei, and Koike). If an LED source with a spectral width of 20 nm is used with a PMMA optical fiber, the material dispersion of the fiber will be $0.305 \times 20 = 6.1$ ns/km. If the spectral width of the source is reduced to 1 nm, the material dispersion reduces to 0.305 ns/km. In the presence of both intermodal and material dispersion, the total dispersion is given by Equation 7-27. Thus, if the intermodal dispersion of the fiber is 0.25 ns/km, the total dispersion for the two spectral widths is given by 6.1 ns/km and 0.394 ns/km respectively. For a 100-m-long fiber, the total dispersions are 0.61 ns and 0.039 ns. Thus, using Equation 7-38, the above dispersion values correspond to maximum bit rates of about 1.1 Gbit/s and 18 Gbit/s respectively using the NRZ scheme. This example shows the importance of reduced spectral width and use of graded-index fibers for achieving high-data-rate communication.

Apart from short-distance communication applications, POFs are expected to find applications in many areas such as lighting for decorative signs and road signs, or in museums (where the property of fibers in non-UV transmission is very useful), in image transmission as in endoscopes, and in sensing applications.

There is also interest in fluorescent or scintillating fibers, which are ordinary plastic fibers doped with special elements. When light falls on a fluorescent fiber, it excites the dopants in the fiber. When the atoms drop to a lower energy state, they emit light. Part of this light is collected within the fiber NA and is transmitted to both ends of the fiber. In scintillating fibers, light is generated when radiations such as α and β particles, or γ rays, are incident on the fiber. Many applications are being envisaged for fluorescent fibers, particularly due to their ability to capture light that falls anywhere along the length of the fiber. Applications include intrusion alarms and size determination. Scintillating optical fibers find applications in particle energy measurement, track detection, etc. For further details, readers may see Spillman and Gravel.

Example 7-26

Polycarbonate fibers with a core index of 1.58 and cladding index of 1.39 are available from some vendors. The NA of such fibers is 0.75, giving an acceptance angle of 49° . Such large-NA fibers provide a large coupling efficiency from light source to fiber.

XIII. FIBER OPTIC SENSORS

Although the most important application of optical fibers is in the field of transmission of information, optical fibers capable of sensing various physical parameters and generating information are also finding widespread use. The use of optical fibers for such applications offers the same advantages as in the field of communication: lower cost, smaller size, more accuracy, greater flexibility, and greater reliability. As compared to conventional electrical sensors, fiber optic sensors are immune to external electromagnetic interference and can be used

in hazardous and explosive environments. A very important attribute of fiber optic sensors is the possibility of having distributed or quasi-distributed sensing geometries, which would otherwise be too expensive or complicated using conventional sensors. With fiber optic sensors it is possible to measure pressure, temperature, electric current, rotation, strain, and chemical and biological parameters with greater precision and speed. These advantages are leading to increased integration of such sensors in civil structures such as bridges and tunnels, process industries, medical instruments, aircraft, missiles, and even cars.

Fiber optic sensors can be broadly classified into two categories: *extrinsic* and *intrinsic*. In the case of extrinsic sensors, the optical fiber simply acts as a device to transmit and collect light from a sensing element, which is external to the fiber. The sensing element responds to the external perturbation, and the change in the characteristics of the sensing element is transmitted by the return fiber for analysis. The optical fiber here plays no role other than that of transmitting the light beam. On the other hand, in the case of intrinsic sensors, the physical parameter to be sensed directly alters the properties of the optical fiber, which in turn leads to changes in a characteristic such as intensity, polarization, or phase of the light beam propagating in the fiber.

A large variety of fiber optic sensors has been demonstrated in the laboratory, and some are already being installed in real systems. In the following, we will discuss some important examples of fiber optic sensors.

A. Extrinsic fiber optic sensors

Figure 7-18 shows a very simple sensor based on the fact that transmission through a fiber joint depends on the alignment of the fiber cores [see Equations 7-33 and 7-34]. Light coupled into a multimode optical fiber couples across a joint into another fiber. The light is detected by a photodetector. Any deviation of the fiber pair from perfect alignment is immediately sensed by the detector. A misalignment of magnitude equal to the core diameter of the fiber results in zero transmission. The first 20% of transverse displacement gives an approximately linear output. Thus, for a 50- μm -core-diameter fiber, approximately 10- μm misalignment will be linear. The sensitivity will of course become better with decrease in core diameter, but, at the same time, the range of displacements will also reduce.

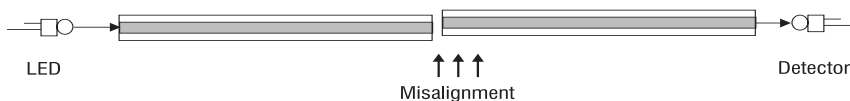


Figure 7-18 A change in the transverse alignment between two fibers changes the coupling and hence the power falling on the detector.

The misalignment between the fibers could be caused by various physical parameters, such as acoustic waves and pressure. Thus, if one of the probe fibers has a short free length while the other has a longer length, acoustic waves impinging on the sensor will set the fibers into vibration, which will result in a modulation of the transmitted light intensity leading to an acoustic sensor. Using such an arrangement, deep-sea noise levels in the frequency range of 100 Hz to 1 kHz and transverse displacements of a few tenths of a nanometer have been

measured (Spillman and Gravel). Using the same principle, any physical parameter leading to a relative displacement of the fiber cores can be sensed using this geometry.

Figure 7-19 shows a modification of the sensor in the form of a probe. Here light from an LED coupled into a multimode fiber passes through a fiber optic splitter to the probe. The probe is in the form of a reflecting diaphragm in front of the fiber, as shown. Light emanating from the fiber is reflected by the diaphragm, passes again through the splitter, and is detected by a photodetector. Any change in the external pressure causes the diaphragm to bend, leading to a change in the power coupled into the fiber. Such sensors can be built to measure pressure variations in medical as well as other applications requiring monitoring operating pressures of up to 4 mega Pascal (~ 600 psi) (Farmer and Digges). Such a device can be used for the measurement of pressure in the arteries, bladder, urethra, etc. Several experiments have shown a very good correlation between this sensor and other conventional sensors.

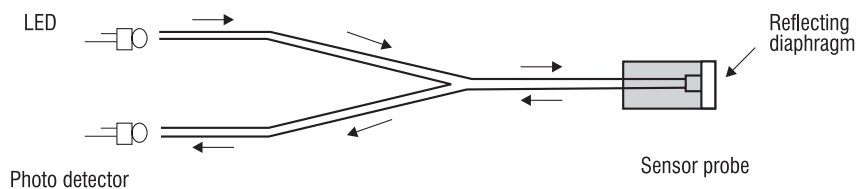


Figure 7-19 *Light returning to the detector changes as the shape of the reflecting diaphragm changes due to changes in external pressure.*

If the diaphragm at the output is removed and the light beam is allowed to fall on the sample, light that is reflected or scattered is again picked up by the fiber and detected and processed by the detector. With analysis of the returning optical beam, information about the physical and chemical properties of the blood can be obtained. Thus, if the scattering takes place from flowing blood, the scattered light beam is shifted in frequency due to the Doppler effect. (Doppler effect refers to the apparent frequency shift of a wave detected by an observer—compared with its true frequency—when there is relative motion between source and observer. You must have noticed the falling frequency of the whistle of a train as it approaches and recedes from you.). The faster the blood cells are moving, the larger will be the shift. Through measurement of the shift in frequency, the blood flow rate can be estimated. By a spectroscopic analysis of the returning optical signal, one can estimate the oxygen content in the blood. One of the most important advantages of using optical fibers in this process is that they do not provoke adverse response from the immune system. They are more durable, more flexible, and potentially safer than alternatives (Katzir, Grifford, and Bartnik).

Another very interesting sensor is the liquid-level sensor shown in Figure 7-20. Light propagating down an optical fiber is total internally reflected from a small glass prism and couples back to the return fiber. As long as the external medium is air, the angle of incidence inside the prism is greater than the critical angle and hence light suffers total internal reflection. As soon as the prism comes in contact with a liquid, the critical angle at the prism-liquid interface reduces and the light is transmitted into the liquid, resulting in a loss of signal. By a proper choice of prism material, such a sensor can be used for sensing levels of various liquids such as water, gasoline, acids, and oils.

OPTICAL WAVEGUIDES AND FIBERS

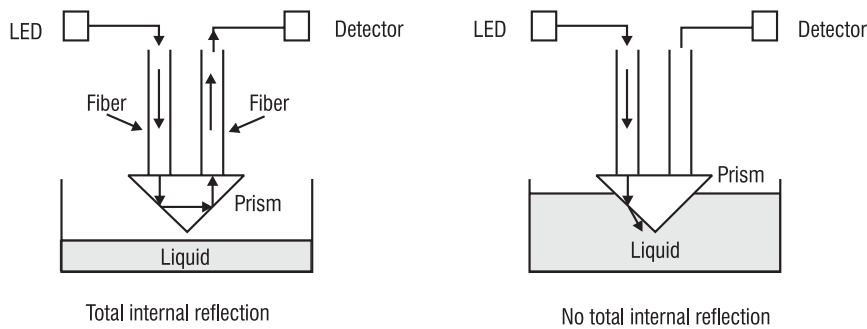


Figure 7-20 A liquid-level sensor based on changes in the critical angle due to liquid level moving up to contact the sides of the prism

Example 7-27

For a prism with refractive index n_p of 1.5, the critical angles with air ($n_a = 1.0$) and water ($n_w = 1.33$) are 41.8° and 62.7° respectively [see Example 7-1]. Thus, if the prism is isosceles right-angled, with two angles as 45° , light that suffers total internal reflection with air as the surrounding medium will suffer only partial internal reflection with water as the surrounding medium, resulting in a loss of signal.

B. Intrinsic sensors

In intrinsic sensors the physical parameter changes some characteristic of the propagating light beam that is sensed. Among the many intrinsic sensors, here we discuss two important examples, namely the Mach-Zehnder interferometric fiber sensor and the fiber optic gyroscope.

Mach-Zehnder interferometric sensor

One of the most sensitive arrangements for a fiber optic sensor is the Mach-Zehnder (MZ) interferometric sensor arrangement shown in Figure 7-21. Light from a laser is passed through a 3-dB fiber optic coupler, which splits the incoming light beam into two equal-amplitude beams in the two single-mode fiber arms. The light beams recombine at the output coupler after passing through the two arms. The output from the output coupler is detected and processed. One of the fiber arms of the interferometer is the sensing arm, which is sensitive to the external parameter to be sensed. The other fiber arm is the reference arm. It is usually coated with a material to make it insensitive to the parameter of measurement. The two fiber arms behave as two paths of an interferometer, and hence the output depends on the phase difference between the beams as they enter the output coupler. If the two fibers are of exactly equal lengths, the entire input light beam appears in the lower fiber and no light comes out of the upper fiber. Any external parameter such as temperature or pressure affects the sensing fiber by changing either the refractive index or the length of the arm, thus changing the phase difference between the two beams as they enter the output coupler. This results in a change in the intensity of the two output arms. Processing of the output leads to a measurement of the external parameter.

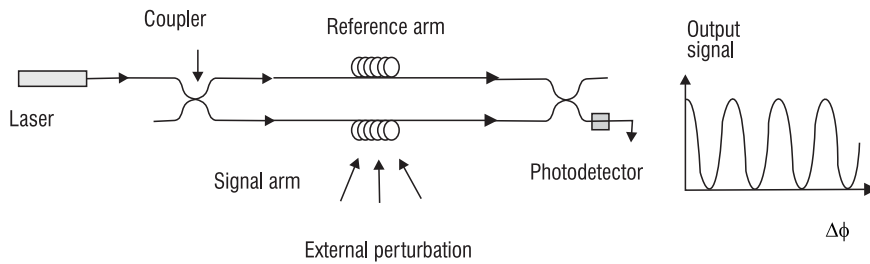


Figure 7-21 The fiber optic Mach-Zehnder interferometric sensor. Phase changes (due to external perturbation on the sensing arm) between the light beams arriving at the output coupler cause changes in intensity at the output.

The MZ sensor is extremely sensitive to external perturbations. For example, the change of phase due to an external pressure that causes both a change in refractive index and a change in the length of the specially coated sensing arm is about 3×10^{-4} rad/Pa-m. Here Pa = 1 N/m^2 represents a Pascal, the unit of pressure. This implies that the change of phase suffered by the beam when the external pressure changes by 1 Pa over 1 m of the fiber is 3×10^{-4} . When someone whispers, the sound pressure corresponds to about 2×10^{-4} Pa at a distance of 1 m. If the length of the sensing arm is 100 m, the corresponding phase change in the light propagating through the sensing arm is 6×10^{-6} rad. Such small changes in phase are detectable by sensitive signal processing.

MZ sensors can be used to sense different physical parameters such as temperature, strain, and magnetic field. These physical parameters cause changes in the phase of the propagating light beam. Such sensors are finding various applications in hydrophones for underwater sound detection. One of the great advantages of such an application is the possibility of configuring the sensors as omnidirectional or highly directional sensors.

Fiber optic rotation sensor—the fiber optic gyroscope (FOG)

One of the more important fiber optic sensors is the *fiber optic gyroscope*, capable of measuring rotation rate. The FOG is a device with no moving parts, with improved lifetime, and of relatively low cost. Thus FOGs are rapidly replacing conventional mechanical gyros for many applications.

The principle of operation of the fiber optic gyroscope is based on the *Sagnac effect*. Figure 7-22 shows a simple FOG configuration. It consists of a loop of polarization-maintaining, single-mode optical fiber connected to a pair of 3-dB directional couplers (capable of splitting the incoming light beam into two equal parts or combining the beams from both input fibers), a polarized source, and a detector. Light from the source is split into two equal parts at the coupler C_1 , one part traveling clockwise and the other counterclockwise in the fiber coil. After traversing the coil the two light beams are recombined at the same coupler and the resulting light energy is detected by a photodetector connected to the coupler C_2 . The source in a FOG is usually a source with a broad spectrum and hence a short coherence length. The source is chosen to avoid any coherent interference between backscattered light from the two counterpropagating beams in the fiber loop. This could be a superluminescent diode or a superfluorescent fiber source.

OPTICAL WAVEGUIDES AND FIBERS

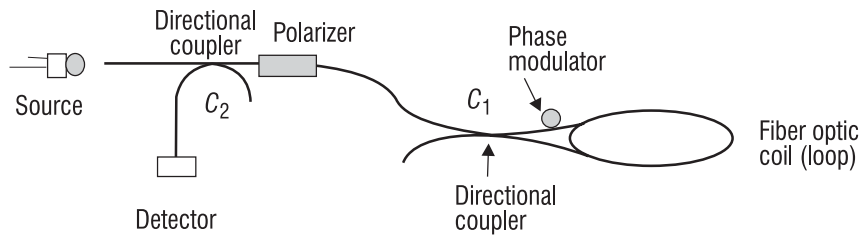


Figure 7-22 A fiber optic gyroscope for rotation sensing based on the Sagnac effect

We first note that, if the loop is not rotating, the clockwise and the counterclockwise beams will take the same time to traverse the loop and hence arrive at the same time with the same phase at the coupler C_1 . On the other hand, when the loop begins to rotate, the times taken by the two beams are different. This can be understood from the fact that, if the loop rotates clockwise, by the time the beams traverse the loop the starting point will have moved and the clockwise beam will take a slightly longer time than the counterclockwise beam to come back to the starting point. This difference of time or phase will result in a change of intensity at the output light beam propagating toward C_2 .

One of the great advantages of a Sagnac interferometer is that the sensor gives no signal for reciprocal stimuli, i.e., stimuli that *act in an identical fashion* on both the beams. Thus a change of temperature affects both the beams (clockwise and counterclockwise) equally and so produces no change in the output.

If the entire loop arrangement rotates with an angular velocity Ω , the phase difference between the two beams is given by

$$\Delta\phi = \frac{8\pi N A \Omega}{c \lambda_0} \quad (7-36)$$

where N is the number of fiber turns in the loop

A is the area enclosed by one turn (which need not be circular)

λ_0 is the free space wavelength of light

Example 7-28

Let us consider a fiber optic gyroscope with a coil of diameter 10 cm, having 1500 turns (corresponding to a total fiber length of $\pi D N \sim 470$ m) and operating at 850 nm. The corresponding phase difference, determined from Equation 7-36 is $\Delta\phi = 1.16 \Omega$ rad. If Ω corresponds to the rotation rate of the Earth (15° per hour) the corresponding phase shift is $\Delta\phi = 8.4 \times 10^{-5}$ rad, a small shift indeed. This phase difference corresponds to a flight time difference between the two beams of $\Delta\tau = \Delta\phi / \omega = \Delta\phi \lambda_0 / (2\pi c) \approx 3.8 \times 10^{-20}$ s.

There are many different ways of operating the gyroscope. One of them is called the closed-loop operation. In this method, a pseudo rotation signal is generated in the gyro to cancel the actual signal caused due to the rotation, thus nulling the output. This is achieved by having a phase

modulator near one end of the loop as shown in Figure 7-22. The counterclockwise-traveling beam encounters the phase modulator later than the clockwise beam. This time difference introduces an artificial phase difference between the two beams. The applied signal on the modulator required to null the signal gives the rotation rate.

Fiber optic gyros capable of measuring from 0.001 deg/h to 100 deg/h are being made. Applications include navigation of aircraft, spacecraft, missiles, manned and unmanned platforms, antenna piloting and tracking, and a compass or north finder. Various applications require FOGs with different sensitivities: Autos require about 10 to 100 deg/h, attitude reference for airplanes requires 1 deg/h, and precision inertial navigation requires gyros with 0.01 to 0.001 deg/h. A Boeing 777 uses an inertial navigation system that has both ring laser gyroscopes and FOGs.

An interesting application involves automobile navigation. The autogyro provides information about the direction and distance traveled and the vehicle's location, which is shown on the monitor in the car. Thus the driver can navigate through a city. Luxury cars from Toyota and Nissan sold in Japan have FOGs as part of their on-board navigation systems. The biggest manufacturer of such gyros is Hitachi of Japan, which manufactures about 3000 gyros per month. [Burns. Bennet. Hotate]

Laboratory

Using the concepts developed in this module, you will be able to perform the following simple experimental projects as part of the laboratory exercise for this module.

- Differentiate between a multimode and a single-mode optical fiber
- Measure NA of a multimode optical fiber
- Measure the attenuation coefficient of a multimode optical fiber
- Determine the mode-field diameter (MFD) of a single-mode fiber
- Determine the splice loss across a multimode fiber joint

Equipment List

HeNe laser

Laser pointer

Multimode and single-mode optical fibers

Microscope objective (20X)

Optical power meter

Translational stages with micromovements (xyz-stack)

Fiber holders (V-groove)

Each of the projects identified above is described in detail below.

A. Multimode and single-mode optical fibers

Take 2-m lengths of multimode and single-mode optical fibers with cutoff wavelength less than 630 nm, so that the fibers are single moded at the HeNe laser wavelength of 633 nm. Couple light into the fibers and observe the output light as it falls on a screen. You should notice that the output from the multimode fiber is characterized by a speckle pattern while that of a single-mode fiber is uniform.

B. NA of a multimode optical fiber

Draw several concentric circles of increasing radii—say, from 0.5 cm to 1.5 cm—on a small paper screen. The screen is positioned in the far field (see Module 1-4, *Basic Physical Optics*) so that the axis of the fiber, at the output end, passes perpendicularly through the center of these circles on the screen (see Figure 7-23). The fiber end, which is mounted on an xyz-stack, is moved slightly toward or away from the screen so that one of the circles there just circumscribes the far-field radiation spot. The distance z between the fiber end and the screen, and the diameter D of the coinciding circle are measured accurately. The NA is calculated using the following equation

$$NA = \sin i_m = \sin [\tan^{-1} (D/2z)] \quad (7-37)$$

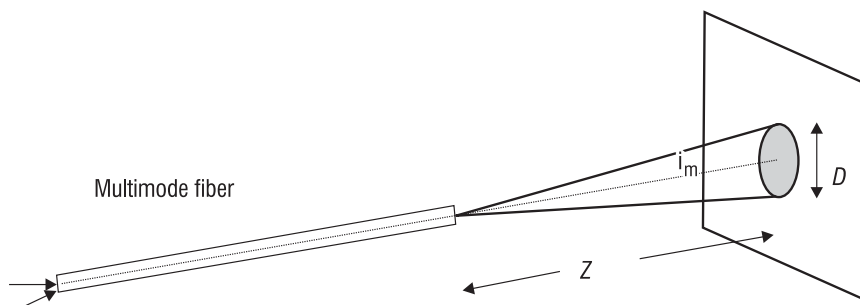


Figure 7-23 Measurement of the diameter D of the spot on a screen placed at a far-field distance z from the output end of a multimode fiber can be used to measure the NA of the fiber.

C. Attenuation measurement

A simple experiment can be performed for measuring the attenuation of the fiber at one specific wavelength. Take a length L (about 1 km) of the fiber and couple the beam from a laser into the fiber. Measure the power P_o exiting at the output end of the fiber. Without disturbing the coupling system, cut off a reference length of 1 m of the fiber from the input end. Measure the power P_i exiting from the 1-m length of the fiber. This will be the input power to the longer portion of the fiber. The attenuation coefficient of the fiber at the wavelength of the laser is then given by

$$\alpha \approx \frac{1}{L} \log \left(\frac{P_i}{P_o} \right) \text{ dB/km}$$

D. MFD measurement of an SMF

The MFD of a single-mode fiber can be easily determined by a measurement of the far-field radiation pattern emerging from the fiber. We have seen that the transverse field distribution of a single-mode fiber is approximately Gaussian in shape. When light exits the single-mode fiber, it diffracts as it propagates through air. The angle of diffraction is given approximately by

$$\theta \approx \frac{\lambda}{\pi w} \quad (7-38)$$

Thus a measurement of θ at a specific chosen wavelength in the single-mode regime can give us the value of w .

To perform the experiment, take a 2-m length of a single-mode fiber, preferably at a visible wavelength such as 633 nm (HeNe laser). Couple light into the fiber using a microscope objective of magnification 20X. Place the photodetector (covered by a pinhole of about 0.1-mm diameter) of an optical power meter at a distance D of about 10 cm from the output end of the fiber. Adjust the detector to coincide with the center of the radiation pattern by maximizing the power. Note the power. Move the photodetector so that the power drops to about $1/e^2$ (~ 0.135) of the maximum. Measure the corresponding distance d moved from the center. Then

$$\theta \approx \frac{d}{D} \approx \frac{\lambda}{\pi w} \quad (7-39)$$

Thus $\text{MFD} = 2w$ can be easily calculated from the measured values of d and D .

E. Splice loss across a multimode fiber joint

In Examples 7-21 and 7-22 we discussed the loss at a joint between a pair of step-index fibers and a pair of parabolic-index multimode fibers. A simple experiment to measure the variation of splice loss across a transverse misalignment can be performed and fitted using the formulas given. Take a multimode optical fiber of about 5-meter length and cut it into two almost equal pieces. Prepare the fiber ends by using a standard fiber cutter. Couple light from a white light source (such as a halogen lamp) into one of the fibers by focusing the light using a microscope objective. Place the other end of the fiber on a V-groove and couple this to the other fiber, which is placed on an xyz-stack. Bring the two ends of the fibers as close as possible—as shown in Figure 7-18—and adjust the transverse position of the fibers so that you couple the maximum amount of light. Now move the fiber in the transverse direction in steps of 10 μm and measure the power at each point. Measure the power exiting from the fiber end after removing the second fiber. Calculate the coupling efficiency (η) and plot as a function of the transverse misalignment. If you have knowledge of the fiber parameters, you can use Equations 7-33 and 7-34 to plot the theoretical variations and compare the experimentally measured variation with the theoretical plot. This experiment will show the sensitivity of the loss to transverse alignment of the two fibers. This is an important consideration while designing connectors for fibers.

A similar experiment can be carried out with single-mode fibers. However, you will need very precise movements to carry out the experiment, since single-mode fibers typically have mode-

field diameters of about 5 to 10 μm . You will have to use Equation 7-32 to carry out the corresponding theoretical analysis.

Problems

1. For a step-index multimode fiber, $n_1 = 1.5$ and $\Delta n = 0.015$, calculate n_2 , NA , and the maximum acceptance angle. (Answer: 1.477, 0.26, 15°)
2. Based on the previous problem, consider the same step-index fiber immersed in water of refractive index 1.33. Calculate the maximum acceptance angle.
3. The power of a 2-mW laser beam decreases to 15 μW after the beam traverses through 25 km of a single-mode optical fiber. Calculate the attenuation of the fiber. (Answer: 0.85 dB/km)
4. A 5-mW laser beam passes through a 26-km fiber of loss 0.2 dB/km. Calculate the power at the output end. (Answer: 1.5 mW)
5. Consider an SIF with $n_1 = 1.5$, $a = 40 \mu\text{m}$, and $\Delta n = 0.015$, operating at 850 nm with a spectral width of 50 nm. a) Is this a single-mode fiber or a multimode fiber? b) Calculate the material dispersion, intermodal dispersion, and total pulse dispersion. [Answer: (b) 4.2 ns/km, 75 ns/km, 75.1 ns/km]
6. Based on the previous problem, consider a parabolic-index fiber with $n_1 = 1.5$, $a = 40 \mu\text{m}$, and $\Delta n = 0.015$, operating at 850 nm with a spectral width of 50 nm. a) Is this a single-mode fiber or a multimode fiber? b) Calculate the material dispersion, intermodal dispersion, and total pulse dispersion. [Answer: (b) 4.2 ns/km, 0.6 ns/km, 4.2 ns/km]
7. Consider a single-mode fiber with $a = 3 \mu\text{m}$ operating with a V-number of 2.3. Calculate the spot size of the fundamental mode. [Answer: 3.4 μm]
8. Consider the SIF given in Problem 5. Calculate the splice loss at a joint between two such identical fibers with transverse misalignments of 5, 10, and 15 μm . [Answer: 0.36 dB, 0.75 dB, 1.2 dB]

References

- Bennet, S. "Fibre Optic gyro system keeps bus riders informed," *Photonics Spectra*, August 1996, pp. 117–120.
- Burns, W.K. "Fiber Optic Gyroscopes—Light is better," *Optics and Photonics News*, May 1998, pp. 28–32.
- Chynoweth, A.G. "Lightwave Communications: The fiber lightguide," *Physics Today*, 29 (5), 28, 1976.
- Farmer, K.R., and T.G. Digges. "A miniature fiber sensor," *Photonics Spectra*, August 1996, pp. 128–129.

- “Fiber Optic Technology Put to Work—Big Time,” *Photonics Spectra*, August 1994, p. 114.
- Gambling, W.A. “Glass, light, and the information revolution, Ninth W.E.S. Turner Memorial Lecture,” *Glass Technology*, Vol. 27 (6), 179, 1986.
- Ghatak, A., I.C. Goyal, and R. Varshney. *Fiber optica: A software for characterizing fiber and integrated optic waveguides*. New Delhi: Viva Books, 1999.
- Ghatak, A., and K. Thyagarajan. *Introduction to fiber optics*. Cambridge: Cambridge University Press, 1998.
- Grifford, R.S., and D.J. Bartnik. “Using optical sensors to measure arterial blood gases,” *Optics and Photonics News*, March 1998, pp. 27–32.
- Hotate, K. “Fiber Optic Gyros,” *Photonics Spectra*, April 1997, p. 108.
- Ishigure, T., E. Nihei, and Y. Koike. “Optimum refractive index profile of the graded index polymer optical fiber, toward gigabit data links,” *Applied Optics*, Vol. 35, 1996, pp. 2048–2053.
- Kao, C.K., and G.A. Hockham. “Dielectric-fibre surface waveguides for optical frequencies,” *Proc. IEEE*, Vol. 113 (7), 1151, 1966.
- Kapron, F.P., D.B. Keck, and R.D. Maurer. “Radiation losses in glass optical waveguides,” *Applied Physics Letters*, Vol. 17, 423, 1970.
- Katzir, A. “Optical Fibers in Medicine,” *Scientific American*, May 1989, pp. 86–91.
- Keiser, G. *Optical Fiber Communications*. New York: McGraw Hill, 1991.
- Koeppen, C., R.F. Shi, W.D. Chen, and A.F. Garito. “Properties of plastic optical fibers,” *Journal of the Optical Society of America*, B Vol. 15, 1998, 727–739.
- Koike, Y., T. Ishigure, and E. Nihei. “High bandwidth graded index polymer optical fiber,” *IEEE Journal of Lightwave Technology*, Vol. 13, 1995, pp. 1475–1489.
- Maclean, D.J.H. *Optical Line Systems*. Chichester: John Wiley, 1996.
- Marcou, J., M. Robiette, and J. Bulabois. *Plastic Optical Fibers*, Chichester: John Wiley and Sons, 1997.
- Marcuse, D. “Loss analysis of single mode fiber splices,” *Bell Systems Tech. Journal*, Vol. 56, 703, 1977.
- Miya, T., Y. Terunuma, T. Hosaka, and T. Miyashita. “An ultimate low loss single mode fiber at 1.55 μm ,” *Electron. Letts.*, Vol. 15, 106, 1979.
- “Schott is lighting the way home,” *Fiberoptic Product News*, February 1997, p. 13.
- Spillman, W.B., and R.L. Gravel. “Moving fiber optic hydrophone,” *Optics Letters*, Vol. 5, 1980, 30–33.

Module
1.8

Fiber Optic Telecommunication

Nick Massa
Springfield Technical Community College
Springfield, Massachusetts

Fiber optics is a major building block in the telecommunication infrastructure. Its high bandwidth capabilities and low attenuation characteristics make it ideal for gigabit transmission and beyond. In this module, you will be introduced to the building blocks that make up a fiber optic communication system. You will learn about the different types of fiber and their applications, light sources and detectors, couplers, splitters, wavelength-division multiplexers, and state-of-the-art devices used in the latest high-bandwidth communication systems. Attention will also be given to system performance criteria such as power and rise-time budgets.

Prerequisites

Before you work through this module, you should have completed Module 1-7, *Basic Principles of Fiber Optics*. In addition, you should be able to manipulate and use algebraic formulas, deal with units, and use basic trigonometric functions such as sine, cosine, and tangent. A basic understanding of wavelength, frequency, and the velocity of light is also assumed.

Objectives

When you finish this module, you will be able to:

- Identify the basic components of a fiber optic communication system
- Discuss light propagation in an optical fiber
- Identify the various types of optical fibers
- Determine the dispersion characteristics for the various types of optical fibersDescribe the various connector types
- Calculate decibel and dBm power
- Calculate the power budget for a fiber optic system
- Calculate the bandwidth of a fiber optic system
- Describe the operation and applications of the various types of fiber optic couplers
- Describe the operation and applications of light-emitting diodes (LEDs)
- Describe the operation and applications of laser diodes (LDs)
- Describe the operation and applications of distributed-feedback (DFB) lasers
- Discuss the differences between LEDs and laser diodes with respect to performance characteristics
- Discuss the differences between the various types of optical detectors with respect to performance characteristics
- Describe how pulse code modulation (PCM) is used in analog-to-digital conversion
- Describe the operation North American Digital Hierarchy
- Describe the difference between internal and external modulation
- Discuss the principles of time-division multiplexing (TDM)
- Discuss the principles of wavelength-division multiplexing (WDM)
- Discuss the principles of dense wavelength-division multiplexing (DWDM)
- Discuss the significance of the International Telecom Union grid (ITU grid)
- Discuss the use of erbium-doped fiber amplifiers (EDFA) for signal regeneration
- Describe the operation and applications of fiber Bragg gratings
- Describe the operation and application of fiber optic circulators
- Describe the operation of a typical fiber optic communication system and the components that make it up

Scenario—Using Fiber Optics in Telecommunication

Michael recently completed an associate degree in laser electro-optics technology at Springfield Technical Community College in Springfield, Massachusetts. Upon graduation he accepted a position as an electro-optics technician at JDS Uniphase Corporation in Bloomfield, Connecticut. The company makes high-speed fiber optic modulators and components that are used in transmitters for the telecommunication and cable television industry.

The company's main focus is on the precision manufacturing of these devices, which requires not only an in-depth knowledge of how the devices work but also an appreciation for the complex manufacturing processes that are required to fabricate the devices to exacting specifications. While Mike was in school, he took courses in optics, fiber optics, and electronics. The background he received, especially in the area of fiber optic testing and measuring, has proven to be invaluable in his day-to-day activities. On the job, Mike routinely works with fusion splicers, optical power meters, and laser sources and detectors, as well as with optical spectrum analyzers and other sophisticated electronic test equipment.

Mike was fortunate in that during his senior year in college he was awarded a full scholarship and internship at JDS Uniphase. The company allowed Mike to complete his degree while working part time. According to Mike, "the experience of working in a high-tech environment while going to school really helps you see the practical applications of what you are learning—which is especially important in a field that is so rapidly changing as fiber optics."

Opening Activities

The field of fiber optics, especially with respect to telecommunication, is a rapidly changing world in which, seemingly, each day a new product or technology is introduced. A good way to start learning about this field is to research the companies that are making major strides in this industry. The Internet is a tremendous source for valuable information on this subject. Try searching the Internet for companies such as:

- Lucent Technologies
- JDS Uniphase
- Ciena
- Alcatel
- Tyco Submarine Systems
- Corning
- AT&T
- Nortel Networks
- Cisco
- Others

Another way to obtain information is to search the Internet for specific topics in fiber optic telecommunication, such as

- Dense wavelength-division multiplexing
- Fiber optic communication
- Dispersion-shifted fiber
- Erbium-doped fiber amplifier
- Fiber optic transmitters
- Fiber optic modulators
- Optical networks
- SONET
- Fiber optic cable

Introduction

Since its invention in the early 1970s, the use of and demand for optical fiber have grown tremendously. The uses of optical fiber today are quite numerous. With the explosion of information traffic due to the Internet, electronic commerce, computer networks, multimedia, voice, data, and video, the need for a transmission medium with the bandwidth capabilities for handling such vast amounts of information is paramount. Fiber optics, with its comparatively infinite bandwidth, has proven to be the solution.

Companies such as AT&T, MCI, and U.S. Sprint use optical fiber cable to carry plain old telephone service (POTS) across their nationwide networks. Local telephone service providers use fiber to carry this same service between central office switches at more local levels, and sometimes as far as the neighborhood or individual home. Optical fiber is also used extensively for transmission of data signals. Large corporations, banks, universities, Wall Street firms, and others own private networks. These firms need secure, reliable systems to transfer computer and monetary information between buildings, to the desktop terminal or computer, and around the world. The security inherent in optical fiber systems is a major benefit. Cable television or community antenna television (CATV) companies also find fiber useful for video services. The high information-carrying capacity, or bandwidth, of fiber makes it the perfect choice for transmitting signals to subscribers.

The *fibering* of America began in the early 1980s. At that time, systems operated at 90 Mb/s. At this data rate, a single optical fiber could handle approximately 1300 simultaneous voice channels. Today, systems commonly operate at 10 Gb/s and beyond. This translates to over 130,000 simultaneous voice channels. Over the past five years, new technologies such as dense wavelength-division multiplexing (DWDM) and erbium-doped fiber amplifiers (EDFA) have been used successfully to further increase data rates to beyond a *terabit per second* (>1000 Gb/s) over distances in excess of 100 km. This is equivalent to transmitting 13 million simultaneous phone calls through a single hair-size glass fiber. At this speed, one can transmit 100,000 books coast to coast in 1 second!

The growth of the fiber optics industry over the past five years has been explosive. Analysts expect that this industry will continue to grow at a tremendous rate well into the next decade and beyond. Anyone with a vested interest in telecommunication would be all the wiser to learn more about the tremendous advantages of fiber optic communication. With this in mind, we hope this module will provide the student with a rudimentary understanding of fiber optic communication systems, technology, and applications in today's information world.

I. BENEFITS OF FIBER OPTICS

Optical fiber systems have many advantages over metallic-based communication systems. These advantages include:

- **Long-distance signal transmission**
The low attenuation and superior signal integrity found in optical systems allow much longer intervals of signal transmission than metallic-based systems. While single-line,

voice-grade copper systems longer than a couple of kilometers (1.2 miles) require in-line signal for satisfactory performance, it is not unusual for optical systems to go over 100 kilometers (km), or about 62 miles, with no active or passive processing.

- **Large bandwidth, light weight, and small diameter**

Today's applications require an ever-increasing amount of bandwidth. Consequently, it is important to consider the space constraints of many end users. It is commonplace to install new cabling within existing duct systems or conduit. The relatively small diameter and light weight of optical cable make such installations easy and practical, saving valuable conduit space in these environments.

- **Nonconductivity**

Another advantage of optical fibers is their dielectric nature. Since optical fiber has no metallic components, it can be installed in areas with electromagnetic interference (EMI), including radio frequency interference (RFI). Areas with high EMI include utility lines, power-carrying lines, and railroad tracks. All-dielectric cables are also ideal for areas of high lightning-strike incidence.

- **Security**

Unlike metallic-based systems, the dielectric nature of optical fiber makes it impossible to remotely detect the signal being transmitted within the cable. The only way to do so is by accessing the optical fiber. Accessing the fiber requires intervention that is easily detectable by security surveillance. These circumstances make fiber extremely attractive to governmental bodies, banks, and others with major security concerns.

- **Designed for future applications needs**

Fiber optics is affordable today, as electronics prices fall and optical cable pricing remains low. In many cases, fiber solutions are less costly than copper. As bandwidth demands increase rapidly with technological advances, fiber will continue to play a vital role in the long-term success of telecommunication.

II. BASIC FIBER OPTIC COMMUNICATION SYSTEM

Fiber optics is a medium for carrying information from one point to another in the form of light. Unlike the copper form of transmission, fiber optics is not electrical in nature. A basic fiber optic system consists of a transmitting device that converts an electrical signal into a light signal, an optical fiber cable that carries the light, and a receiver that accepts the light signal and converts it back into an electrical signal. The complexity of a fiber optic system can range from

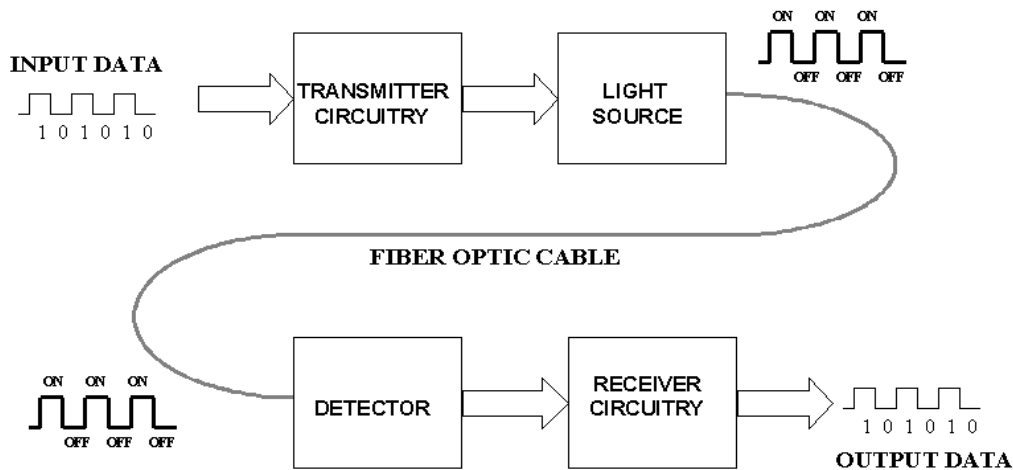


Figure 8-1 Basic fiber optic communication system

very simple (i.e., local area network) to extremely sophisticated and expensive (i.e., long-distance telephone or cable television trunking). For example, the system shown in Figure 8-1 could be built very inexpensively using a visible LED, plastic fiber, a silicon photodetector, and some simple electronic circuitry. The overall cost could be less than \$20. On the other hand, a typical system used for long-distance, high-bandwidth telecommunication that employs wavelength-division multiplexing, erbium-doped fiber amplifiers, external modulation using DFB lasers with temperature compensation, fiber Bragg gratings, and high-speed infrared photodetectors could cost tens or even hundreds of thousands of dollars. The basic question is “how much information is to be sent and how far does it have to go?” With this in mind we will examine the various components that make up a fiber optic communication system and the considerations that must be taken into account in the design of such systems.

III. TRANSMISSION WINDOWS

Optical fiber transmission uses wavelengths that are in the near-infrared portion of the spectrum, just above the visible, and thus undetectable to the unaided eye. Typical optical transmission wavelengths are 850 nm, 1310 nm, and 1550 nm. Both lasers and LEDs are used to transmit light through optical fiber. Lasers are usually used for 1310- or 1550-nm single-mode applications. LEDs are used for 850- or 1300-nm multimode applications.

There are ranges of wavelengths at which the fiber operates best. Each range is known as an operating window. Each window is centered on the typical operational wavelength, as shown in Table 8.1.

Table 8.1: Fiber Optic Transmission Windows

Window	Operating Wavelength
800 – 900 nm	850 nm
1250 – 1350 nm	1310 nm
1500 – 1600 nm	1550 nm

These wavelengths were chosen because they best match the transmission properties of available light sources with the transmission qualities of optical fiber.

IV. FIBER OPTIC LOSS CALCULATIONS

Loss in a system can be expressed as the following:

$$\text{Loss} = \frac{P_{\text{out}}}{P_{\text{in}}} \quad (8-1)$$

where P_{in} is the input power to the fiber and P_{out} is the power available at the output of the fiber. For convenience, *fiber optic loss* is typically expressed in terms of decibels (dB) and can be calculated using Equation 8-2a.

$$\text{Loss}_{\text{dB}} = 10 \log \frac{P_{\text{out}}}{P_{\text{in}}} \quad (8-2a)$$

Oftentimes, loss in optical fiber is also expressed in terms of decibels per kilometer (dB/km)

Example 1

A fiber of 100-m length has $P_{\text{in}} = 10 \mu\text{W}$ and $P_{\text{out}} = 9 \mu\text{W}$. Find the loss in dB/km.

From Equation 8-2

$$\text{Loss}_{\text{dB}} = 10 \log \left(\frac{9 \mu\text{W}}{10 \mu\text{W}} \right) = -0.458 \text{ dB}$$

and since

$$100 \text{ m} = 0.1 \text{ km}$$

the loss is

$$\text{Loss}(\text{dB/km}) = \frac{-0.458 \text{ dB}}{0.1 \text{ km}} = -4.58 \text{ dB/km}$$

$\therefore$ The negative sign implies loss.

Example 2

A communication system uses 10 km of fiber that has a 2.5-dB/km loss characteristic. Find the output power if the input power is 400 mW.

Solution:

From Equation 8-2, and making use of the relationship that $y = 10^x$ if $x = \log y$,

$$\text{Loss}_{\text{dB}} = 10 \log \left(\frac{P_{\text{out}}}{P_{\text{in}}} \right)$$

$$\frac{\text{Loss}_{\text{dB}}}{10} = \log \left(\frac{P_{\text{out}}}{P_{\text{in}}} \right)$$

which becomes, then,

$$10^{\frac{\text{Loss}_{\text{dB}}}{10}} = \left(\frac{P_{\text{out}}}{P_{\text{in}}} \right).$$

So, finally, we have

$$P_{\text{out}} = P_{\text{in}} \times 10^{\frac{\text{Loss}_{\text{dB}}}{10}} \quad (8-2b)$$

For 10 km of fiber with 2.5-dB/km loss characteristic, the loss_{dB} becomes

$$\text{Loss}_{\text{dB}} = 10 \text{ km} \times (-2.5 \text{ dB/km}) = -25 \text{ dB}$$

Plugging this back into Equation 8-2b,

$$P_{\text{out}} = (400 \text{ mW}) \times 10^{\frac{-25}{10}} = 1.265 \text{ mW}$$

Optical power in fiber optic systems is typically expressed in terms of dBm, which is a decibel term that assumes that the input power is 1 mwatt. Optical power here can refer to the power of a laser source or just to the power somewhere in the system. If P in Equation 8-3 is in milliwatts, Equation 8-3 gives the power in dBm, referenced to an input of one milliwatt:

$$P(\text{dBm}) = 10 \log \left(\frac{P}{1 \text{ mW}} \right) \quad (8-3)$$

With optical power expressed in dBm, output power anywhere in the system can be determined simply by expressing the power input in dBm and subtracting the individual component losses, also expressed in dB. It is important to note that an optical source with a power input of 1 mW can be expressed as 0 dBm, as indicated by Equation 8-3. For every 3-dB loss, the power is cut in half. Consequently, for every 3-dB increase, the optical power is doubled. For example, a 3-dBm optical source has a P of 2 mW, whereas a -6-dBm source has a P of 0.25 mW, as can be verified with Equation 8-3.

Example 3

A 3-km fiber optic system has an input power of 2 mW and a loss characteristic of 2 dB/km. Determine the output power of the fiber optic system.

Solution:

Using Equation 8-3, we convert the source power of 2 mW to its equivalent in dBm:

$$\text{Input power}_{\text{dBm}} = 10 \log \left(\frac{2 \text{ mW}}{1 \text{ mW}} \right) = +3 \text{ dBm}$$

The loss_{dB} for the 3-km cable is,

$$\text{Loss}_{\text{dB}} = 3 \text{ km} \times 2 \text{ dB/km} = 6 \text{ dB}$$

Thus, power in dB is (Output power)_{dB} = +3 dBm – 6 dB = –3 dBm

Using Equation 8-3 to convert the output power of –3 dBm back to milliwatts, we have

$$P(\text{dBm}) = 10 \log \frac{P(\text{mW})}{1 \text{ mW}}$$

so that
$$P(\text{mW}) = 1 \text{ mW} \times 10^{\frac{P(\text{dBm})}{10}}$$

Plugging in for $P(\text{dBm}) = -3 \text{ dBm}$, we get for the output power in milliwatts

$$P(\text{mW}) = 1 \text{ mW} \times 10^{\frac{-3}{10}} = 0.5 \text{ mW}$$

Note that one can also use Equation 8-2a to get the same result, where now $P_{\text{in}} = 2 \text{ mW}$ and Loss_{dB} = –6 dB:

$$P_{\text{out}} = P_{\text{in}} \times 10^{\frac{\text{Loss}_{\text{dB}}}{10}}$$

or
$$P_{\text{out}} = 2 \text{ mW} \times 10^{\frac{-6}{10}} = 0.5 \text{ mW}, \text{ the same as above.}$$

V. TYPES OF FIBER

Three basic types of fiber optic cable are used in communication systems:

1. Step-index multimode
2. Step-index single mode
3. Graded-index

This is illustrated in Figure 8-2.

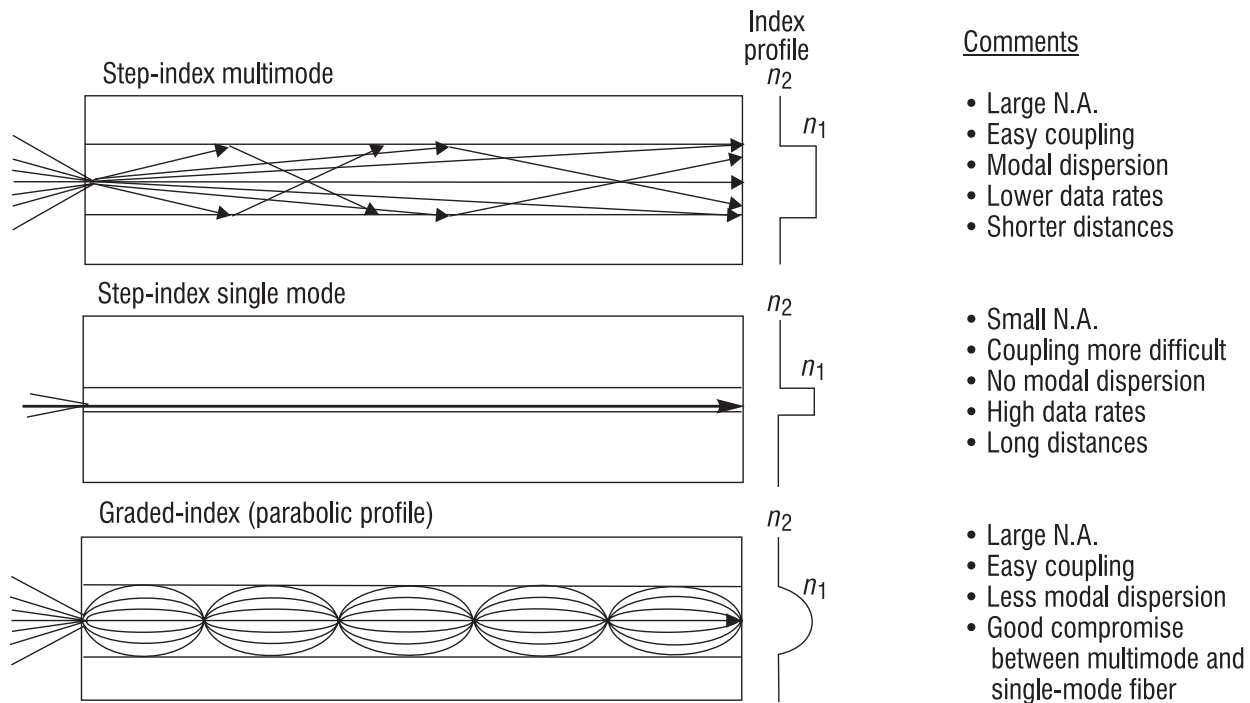


Figure 8-2 Types of fiber

Step-index multimode fiber has an index of refraction profile that “steps” from low to high to low as measured from cladding to core to cladding. Relatively large core diameter and numerical aperture characterize this fiber. The core/cladding diameter of a typical multimode fiber used for telecommunication is 62.5/125 μm (about the size of a human hair). The term “multimode” refers to the fact that multiple *modes* or *paths* through the fiber are possible. Step-index multimode fiber is used in applications that require high bandwidth (< 1 GHz) over relatively short distances (< 3 km) such as a local area network or a campus network backbone.

The major benefits of multimode fiber are: (1) it is relatively easy to work with; (2) because of its larger core size, light is easily coupled to and from it; (3) it can be used with both lasers and LEDs as sources; and (4) coupling losses are less than those of the single-mode fiber. The drawback is that because many modes are allowed to propagate (a function of core diameter, wavelength, and numerical aperture) it suffers from *modal dispersion*. The result of modal dispersion is bandwidth limitation, which translates into lower data rates.

Single-mode step-index fiber allows for only one path, or mode, for light to travel within the fiber. In a multimode step-index fiber, the number of modes M_n propagating can be approximated by

$$M_n = \frac{V^2}{2} \quad (8-4)$$

Here V is known as the normalized frequency, or the **V-number**, which relates the fiber size, the refractive index, and the wavelength. The V -number is given by Equation (8-5)

$$V = \left[\frac{2\pi a}{\lambda} \right] \times \text{N.A.} \quad (8-5)$$

or by Equation 8-6.

$$V = \frac{2\pi a}{\lambda} \times n_1 \times (2 \times \Delta)^{\frac{1}{2}} \quad (8-6)$$

In either equation, a is the fiber core radius, λ is the operating wavelength, N.A. is the numerical aperture, n_1 is the core index, and Δ is the relative refractive index difference between core and cladding.

The analysis of how the V -number is derived is beyond the scope of this module, but it can be shown that by reducing the diameter of the fiber to a point at which the V -number is less than 2.405, higher-order modes are effectively extinguished and single-mode operation is possible.

Example 4

What is the maximum core diameter for a fiber if it is to operate in single mode at a wavelength of 1550 nm if the N.A. is 0.12?

From Equation 8-5,

$$V = \left[\frac{2\pi a}{\lambda} \right] \times \text{N.A.}$$

Solving for a yields

$$a = (V)(\lambda)/(2\pi \text{N.A.})$$

For single-mode operation, V must be 2.405 or less. The maximum core diameter occurs when $V = 2.405$. So, plugging into the equation, we get

$$a_{\max} = (2.405)(1550 \text{ nm})/[(2\pi)(0.12)] = 4.95 \text{ } \mu\text{m}$$

or

$$d_{\max} = 2 \times a = 9.9 \text{ } \mu\text{m}$$

The core diameter for a typical single-mode fiber is between 5 μm and 10 μm with a 125- μm cladding. Single-mode fibers are used in applications in which low signal loss and high data rates are required, such as in long spans where repeater/amplifier spacing must be maximized. Because single-mode fiber allows only one mode or ray to propagate (the lowest-order mode), it does not suffer from modal dispersion like multimode fiber and therefore can be used for higher bandwidth applications. However, even though single-mode fiber is not affected by modal dispersion, at higher data rates *chromatic* dispersion can limit the performance. This problem can be overcome by several methods. One can transmit at a wavelength in which glass has a fairly constant index of refraction ($\sim 1300 \text{ nm}$), use an optical source such as a distributed-feedback laser (DFB laser) that has a very narrow output spectrum, use special dispersion-

compensating fiber, or use a combination of all these methods. In a nutshell, single-mode fiber is used in high-bandwidth, long-distance applications such as long-distance telephone trunk lines, cable TV head-ends, and high-speed local and wide area network (LAN and WAN) backbones. The major drawback of single-mode fiber is that it is relatively difficult to work with (i.e., splicing and termination) because of its small core size. Also, single-mode fiber is typically used only with laser sources because of the high coupling losses associated with LEDs.

Graded-index fiber is a compromise between the large core diameter and N.A. of multimode fiber and the higher bandwidth of single-mode fiber. With creation of a core whose index of refraction decreases parabolically from the core center toward the cladding, light traveling through the center of the fiber experiences a higher index than light traveling in the higher modes. This means that the higher-order modes travel faster than the lower-order modes, which allows them to “catch up” to the lower-order modes, thus decreasing the amount of modal dispersion, which increases the bandwidth of the fiber.

VI. DISPERSION

Dispersion, expressed in terms of the symbol Δt , is defined as pulse spreading in an optical fiber. As a pulse of light propagates through a fiber, elements such as numerical aperture, core diameter, refractive index profile, wavelength, and laser linewidth cause the pulse to broaden. This poses a limitation on the overall bandwidth of the fiber as demonstrated in Figure 8-3.

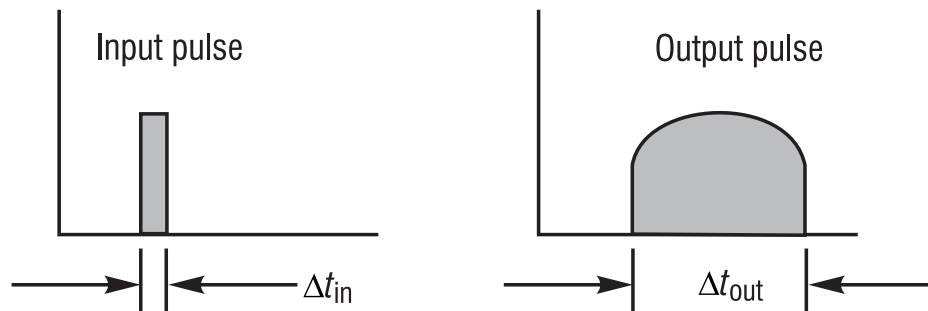


Figure 8-3 Pulse broadening caused by dispersion

Dispersion Δt can be determined from Equation 8-7.

$$\Delta t = (\Delta t_{\text{out}} - \Delta t_{\text{in}})^{1/2} \quad (8-7)$$

and is measured in time, typically nanoseconds or picoseconds. Total dispersion is a function of fiber length. The longer the fiber, the more the dispersion. Equation 8-8 gives the total dispersion per unit length.

$$\Delta t_{\text{total}} = L \times (\text{Dispersion/km}) \quad (8-8)$$

The overall effect of dispersion on the performance of a fiber optic system is known as *intersymbol interference* (Figure 8-4). Intersymbol interference occurs when the pulse spreading caused by dispersion causes the output pulses of a system to overlap, rendering them

undetectable. If an input pulse is caused to spread such that the rate of change of the input exceeds the dispersion limit of the fiber, the output data will become indiscernible.

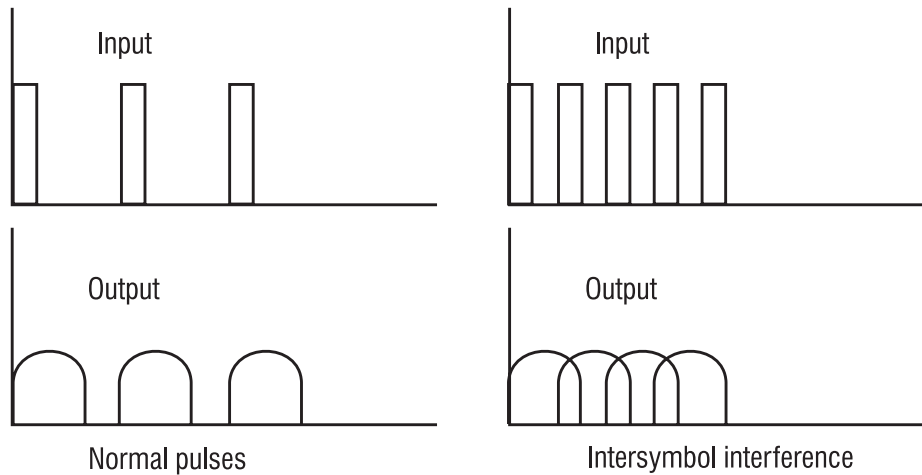


Figure 8-4 Intersymbol interference

Dispersion is generally divided into two categories: *modal dispersion* and *chromatic dispersion*.

Modal dispersion is defined as pulse spreading caused by the time delay between lower-order modes (modes or rays propagating straight through the fiber close to the optical axis) and higher-order modes (modes propagating at steeper angles). This is shown in Figure 8-5. Modal dispersion is problematic in multimode fiber, causing bandwidth limitation, but it is not a problem in single-mode fiber where only one mode is allowed to propagate.

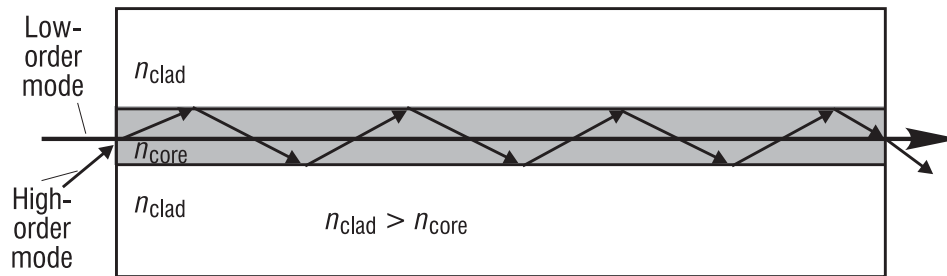


Figure 8-5 Mode propagation in an optical fiber

Chromatic dispersion is pulse spreading due to the fact that different wavelengths of light propagate at slightly different velocities through the fiber. All light sources, whether laser or LED, have finite linewidths, which means they emit more than one wavelength. Because the index of refraction of glass fiber is a wavelength-dependent quantity, different wavelengths propagate at different velocities. Chromatic dispersion is typically expressed in units of nanoseconds or picoseconds per (km-nm).

Chromatic dispersion consists of two parts: material dispersion and waveguide dispersion.

$$\Delta t_{\text{chromatic}} = \Delta t_{\text{material}} + \Delta t_{\text{waveguide}} \quad (8-9)$$

Material dispersion is due to the wavelength dependency on the index of refraction of glass. Waveguide dispersion is due to the physical structure of the waveguide. In a simple step-index-profile fiber, waveguide dispersion is not a major factor, but in fibers with more complex index profiles, waveguide dispersion can be more significant. Material dispersion and waveguide dispersion can have opposite signs depending on the transmission wavelength. In the case of a step-index single-mode fiber, these two effectively cancel each other at 1310 nm, yielding zero-dispersion. This makes very high-bandwidth communication possible at this wavelength. However, the drawback is that, even though dispersion is minimized at 1310 nm, attenuation is not. Glass fiber exhibits minimum attenuation at 1550 nm. Coupling that with the fact that erbium-doped fiber amplifiers (EDFA) operate in the 1550-nm range makes it obvious that, if the zero-dispersion property of 1310 nm could be shifted to coincide with the 1550-nm transmission window, high-bandwidth long-distance communication would be possible. With this in mind, *zero-dispersion-shifted fiber* was developed.

When considering the total dispersion from different causes, we can approximate the total dispersion by Δt_{tot} .

$$\Delta t_{\text{tot}} = \left[(\Delta t_1)^2 + (\Delta t_2)^2 + \dots + (\Delta t_n)^2 \right]^{1/2} \quad (8-10)$$

where Δt_n represents the dispersion due to the various components that make up the system. The transmission capacity of fiber is typically expressed in terms of *bandwidth* $\times$ *distance*. For example, the bandwidth $\times$ distance product for a typical 62.5/125- μm (core/cladding diameter) multimode fiber operating at 1310 nm might be expressed as 600 MHz $\cdot$ km. The approximate bandwidth of a fiber can be related to the total dispersion by the following relationship

$$\text{BW} = 0.35/\Delta t_{\text{total}} \quad (8-11)$$

Example 5

A 2-km-length multimode fiber has a modal dispersion of 1 ns/km and a chromatic dispersion of 100 ps/km $\cdot$ nm. If it is used with an LED of linewidth 40 nm, (a) what is the total dispersion? (b) Calculate the bandwidth (BW) of the fiber.

- a. $\Delta t_{\text{modal}} = 2 \text{ km} \times 1 \text{ ns/km} = 2 \text{ ns}$
 $\Delta t_{\text{chromatic}} = (2 \text{ km}) \times (100 \text{ ps/km} \cdot \text{nm}) \times (40 \text{ nm}) = 8000 \text{ ps} = 8 \text{ ns}$
 $\Delta t_{\text{total}} = \left((2 \text{ ns})^2 + (8 \text{ ns})^2 \right)^{1/2} = 8.24 \text{ ns}$
- b. $\text{BW} = 0.35/\Delta t_{\text{total}} = 0.35/8.24 \text{ ns} = 42.48 \text{ MHz}$

Expressed in terms of the product (BW $\cdot$ km), we get (BW $\cdot$ km) = (42.5 MHz)(2 km) $\simeq$ 85 MHz $\cdot$ km.

Dispersion-shifted fiber: By altering the design of the waveguide, we can increase the magnitude of the waveguide dispersion so as to shift the zero-dispersion wavelength to 1550 nm. This type of fiber has an index profile that resembles a “W” and hence is sometimes referred to as W-profile fiber (Figure 8-6). Although this type of fiber works well at the zero-dispersion wavelength, in systems in which multiple wavelengths are transmitted, such as in wavelength-division multiplexing, signals transmitted at different wavelengths around 1550 nm can interfere with one another, resulting in a phenomenon called *four-wave mixing*, which degrades system performance. However, if the waveguide structure of the fiber is modified so that the waveguide dispersion is further increased, the zero-dispersion point can be pushed past 1600 nm (outside the EDFA operating window). This means that the total chromatic dispersion can still be substantially lowered in the 1550-nm range without having to worry about performance problems. This type of fiber is known as *nonzero-dispersion-shifted fiber*. Figure 8-7 compares the material chromatic and wavelength dispersions for single-mode fiber and dispersion-shifted fiber.

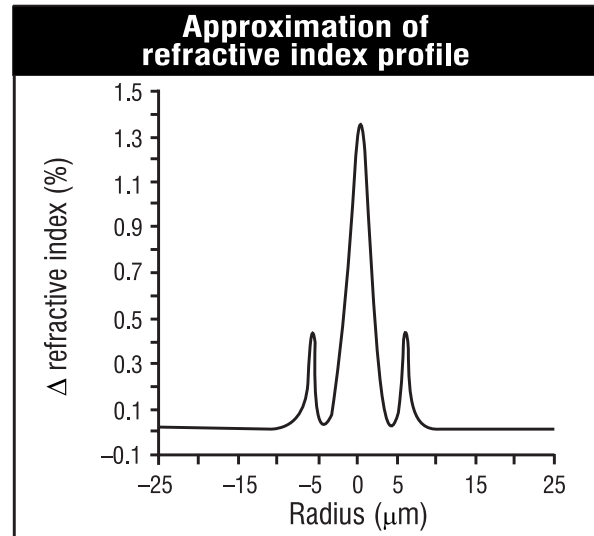


Figure 8-6 W-profile fiber

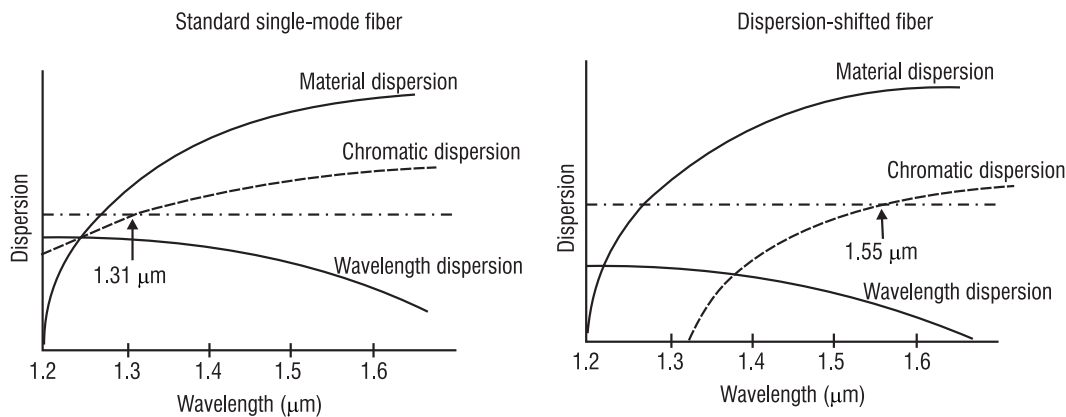


Figure 8-7 Single-mode versus dispersion-mode versus dispersion-shifted fiber

VII. ANALOG VERSUS DIGITAL SIGNALS

Information in a fiber optic system can be transmitted in one of two ways: *analog* or *digital* (see Figure 8-8). An analog signal is one that varies continuously with time. For example, when you speak into the telephone, your voice is converted to an analog voltage that varies continuously. The signal from your cable TV company is also analog. A digital signal is one that exists only at discrete levels. For example, in a computer, information is represented as zeros and ones (0 and

5 volts). In the case of the telephone, the analog voice signal emanating from your handset is sent through a pair of wires to a device called a concentrator, which is located either on a utility pole, in a small service box, or in a manhole. The concentrator converts the analog signal to a digital signal that is combined with many other telephone signals through a process called multiplexing. In telecommunication, most signals are digitized. An exception is cable TV, which still transmits video information in analog form. With the advent of digital and high-definition television (HDTV), cable TV will eventually also be transmitted digitally.

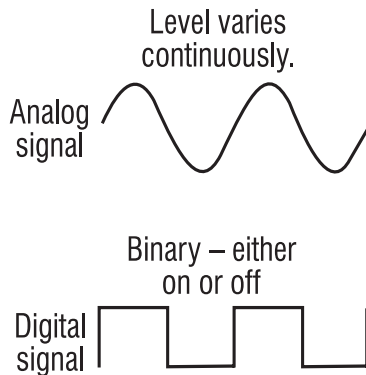


Figure 8-8 Analog and digital signals

Digital transmission has several advantages over analog transmission. First, it is easier to process electronically. No conversion is necessary. It is also less susceptible to noise because it operates with discrete signal levels. The signal is either on or off, which makes it harder to corrupt. Digital signals may also be encoded to detect and correct transmission errors.

VIII. PULSE CODE MODULATION

Pulse code modulation (PCM) is the process of converting an analog signal into a 2^n -digit binary code. Consider the block diagram shown in Figure 8-9. An analog signal is placed on the input of a *sample and hold*. The *sample and hold* circuit is used to “capture” the analog voltage long enough for the conversion to take place. The output of the sample and hold circuit is fed into the *analog-to-digital converter* (A/D). An A/D converter operates by taking periodic discrete samples of an analog signal at a specific point in time and converting it to a 2^n -bit binary number. For example, an 8-bit A/D converts an analog voltage into a binary number with 2^8 discrete levels (between 0 and 255). For an analog voltage to be successfully converted, it must be sampled at a rate at least *twice* its maximum frequency. This is known as the *Nyquist sampling rate*. An example of this is the process that takes place in the telephone system. A standard telephone has a bandwidth of 4 kHz. When you speak into the telephone, your 4-kHz bandwidth voice signal is sampled at twice the 4-kHz frequency or 8 kHz. Each sample is then converted to an 8-bit binary number. This occurs 8000 times per second. Thus, if we multiply

$$8 \text{ k samples/s} \times 8 \text{ bits/sample} = 64 \text{ kbits/s}$$

we get the standard bit rate for a single voice channel in the North American DS1 System, which is 64 kbits/s. The output of the A/D converter is then fed into a driver circuit that contains the appropriate circuitry to turn the light source on and off. The process of turning the light source on and off is known as modulation and will be discussed later in this module. The light then travels through the fiber and is received by a photodetector that converts the optical signal into an electrical current. A typical photodetector generates a current that is in the micro- or nanoamp range, so amplification and/or signal reshaping is often required. Once the digital signal has been reconstructed, it is converted back into an analog signal using a device called a *digital-to-analog converter* or DAC. A digital storage device or buffer may be used to temporarily store the digital codes during the conversion process. The DAC accepts an n -bit digital number and outputs a continuous series of discrete voltage “steps.” All that is needed to smooth the stair-step voltage out is a simple low-pass filter with its cutoff frequency set at the maximum signal frequency as shown in Figure 8-10.

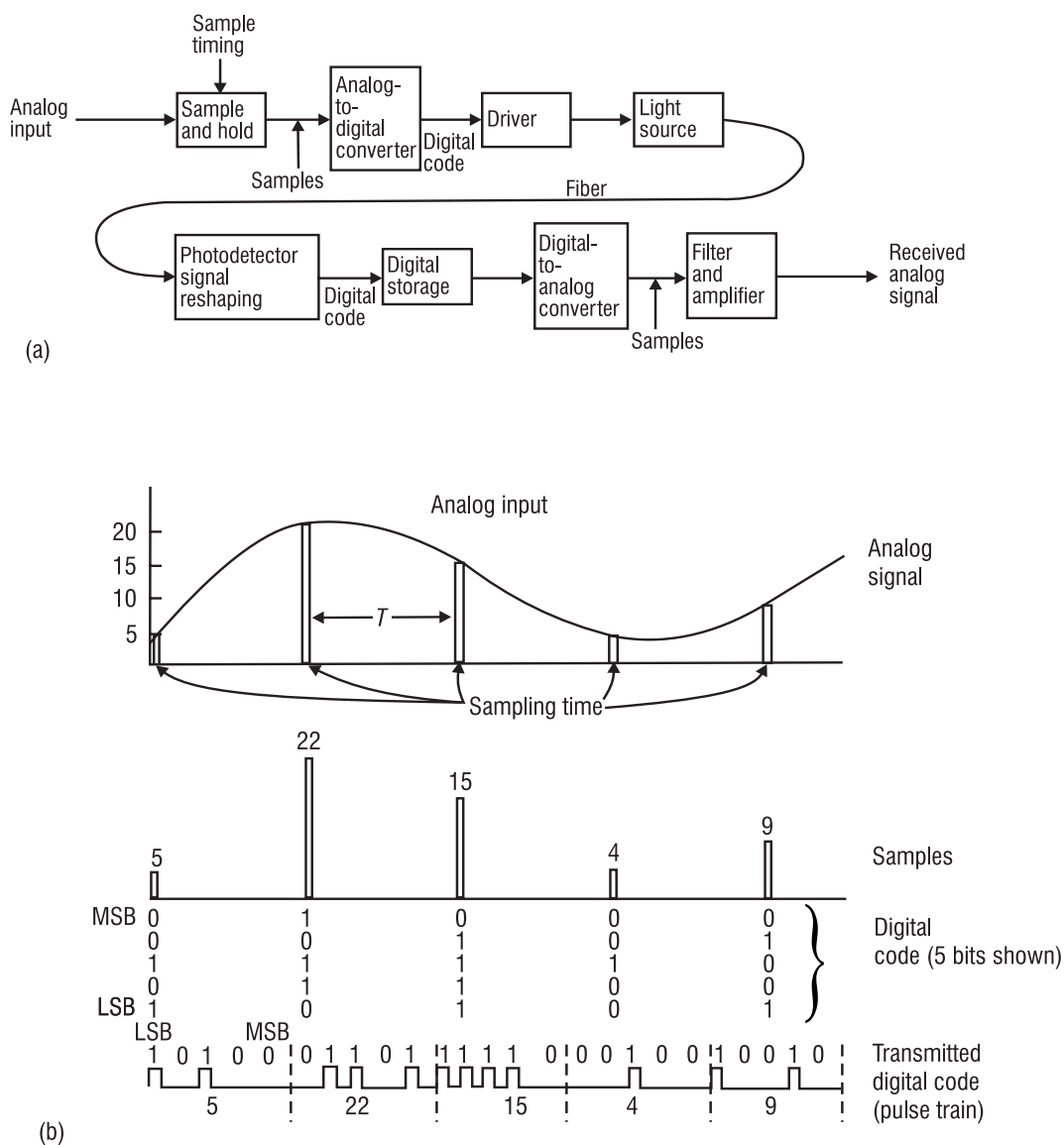


Figure 8-9 (a) Block diagram (b) Digital waveforms

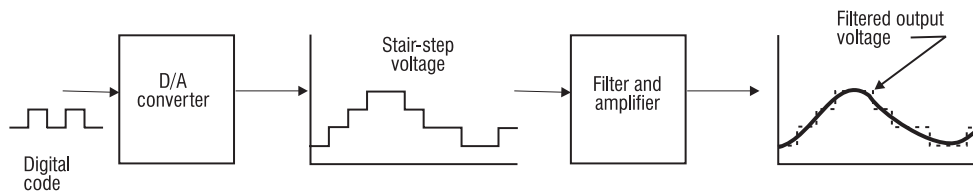


Figure 8-10 D/A output circuit

IX. DIGITAL ENCODING SCHEMES

Signal format is an important consideration in evaluating the performance of a fiber optic system. The signal format directly affects the detection of the transmitted signals. The accuracy of the reproduced signal depends on the intensity of the received signal, the speed and linearity of the receiver, and the noise levels of the transmitted and received signal. Many coding schemes are used in digital communication systems, each with its own benefits and drawbacks. The most common encoding schemes are the return-to-zero (RZ) and non-return-to-zero (NRZ). The NRZ encoding scheme, for example, requires only one transition per symbol, whereas RZ format requires two transitions for each data bit. This implies that the required bandwidth for RZ must be twice that of NRZ. This is not to say that one is better than the other. Depending on the application, any of the code formats may be more appropriate than the others. For example, in synchronous transmission systems in which large amounts of data are to be sent, clock synchronization between the transmitter and receiver must be ensured. In this case *Manchester encoding* is used. The transmitter clock is embedded in the data. The receiver clock is derived from the guaranteed transition in the middle of each bit. The various methods are illustrated in Figure 8-11.

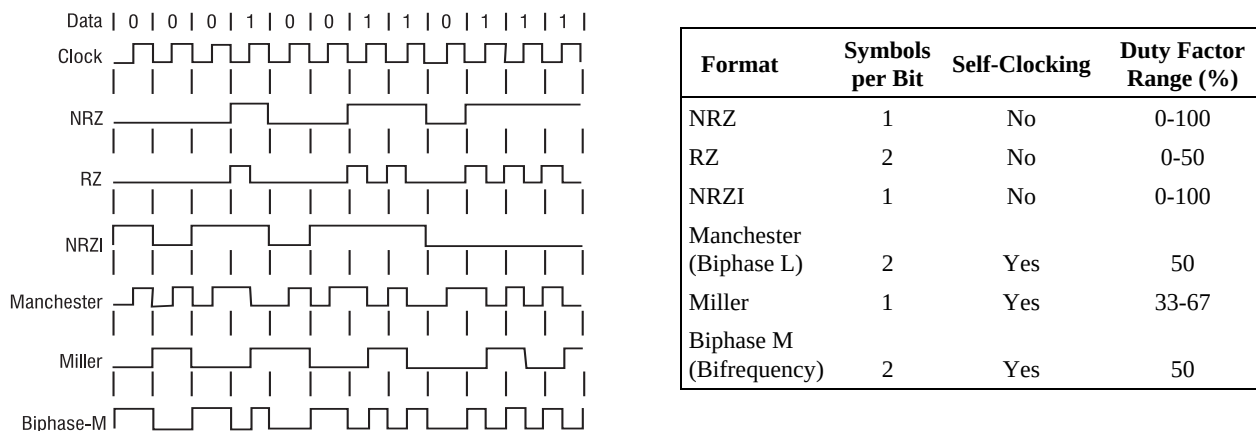


Figure 8-11 Different encoding schemes

Digital systems are analyzed on the basis of rise time rather than on bandwidth. The rise time of a signal is defined as the time required for the signal to change from 10% to 90% of its maximum value. The system rise time is determined by the data rate and code format. Depending on which code format is used, the number of transitions required to represent the

transmitted data may limit overall the data rate of the system. The system rise time depends on the combined rise time characteristics of the individual system components.

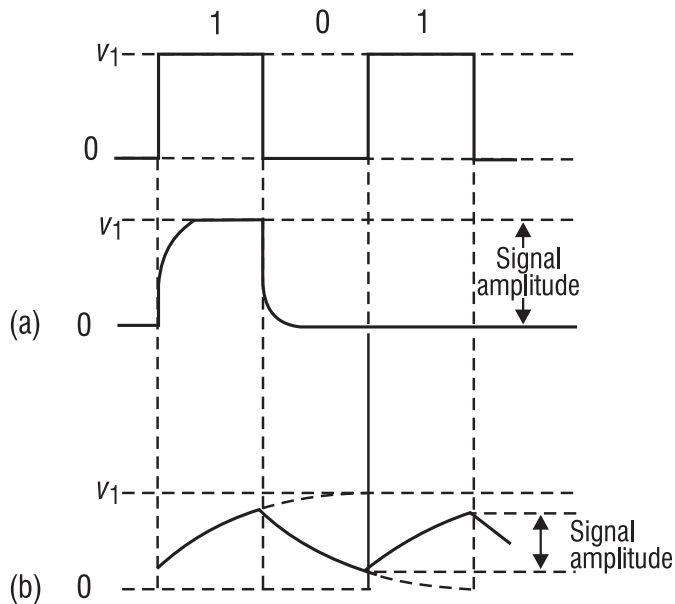
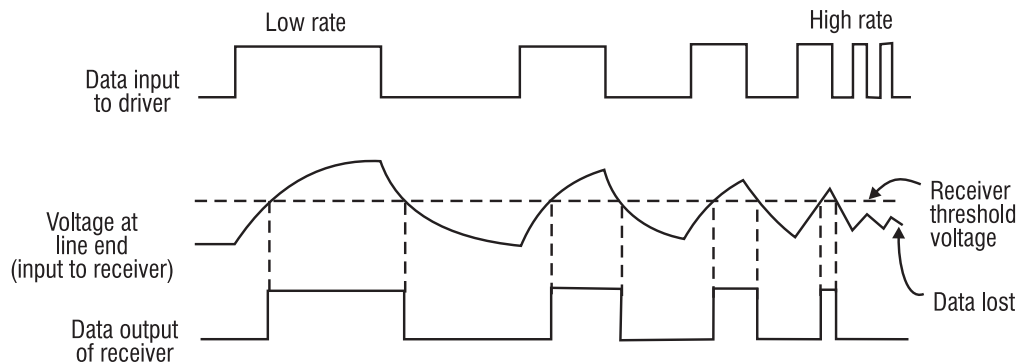


Figure 8-12 Effect of rise time: (a) Short rise time (b) Long rise time

The signal shown in Figure 8-12 (a) represents a signal with adequate rise time. Even though the pulses are somewhat rounded on the edges, the signal is still detectable. In Figure 8-12 (b), however, the transmitted signal takes too long to respond to the input signal. The effect is exaggerated in Figure 8-13, where, at high data rates, the rise time limitations cause the data to be distorted and thus lost.



Source: *The TTL Application Handbook*, August 1973f, p. 14-7.
Reprinted with permission of National Semiconductor.

Figure 8-13 Distortion of data bits by varying data rates

To avoid this distortion, an acceptable criterion is to require that a system have a rise time t_s of no more than 70% of the pulse width T_p ;

$$t_s \leq (0.7 \times T_p) \quad (8-12)$$

For an RZ, T_p takes half the bit time T so that

$$t_s \leq (0.7 \times T)/2 \quad (8-13)$$

or

$$t_s \leq 0.35/B_r \quad (8-14)$$

where $B_r = 1/T$ is the system bit rate.

For an NRZ format, $T_p = T$ and thus

$$t_s \leq 0.7/B_r \quad (8-15)$$

$\therefore$ RZ transmission requires a larger-bandwidth system.

Figure 8-14 shows transmitted (a) RZ and (c) NRZ pulse trains and the effects of system rise time on (b) format RZ and (d) format NRZ.

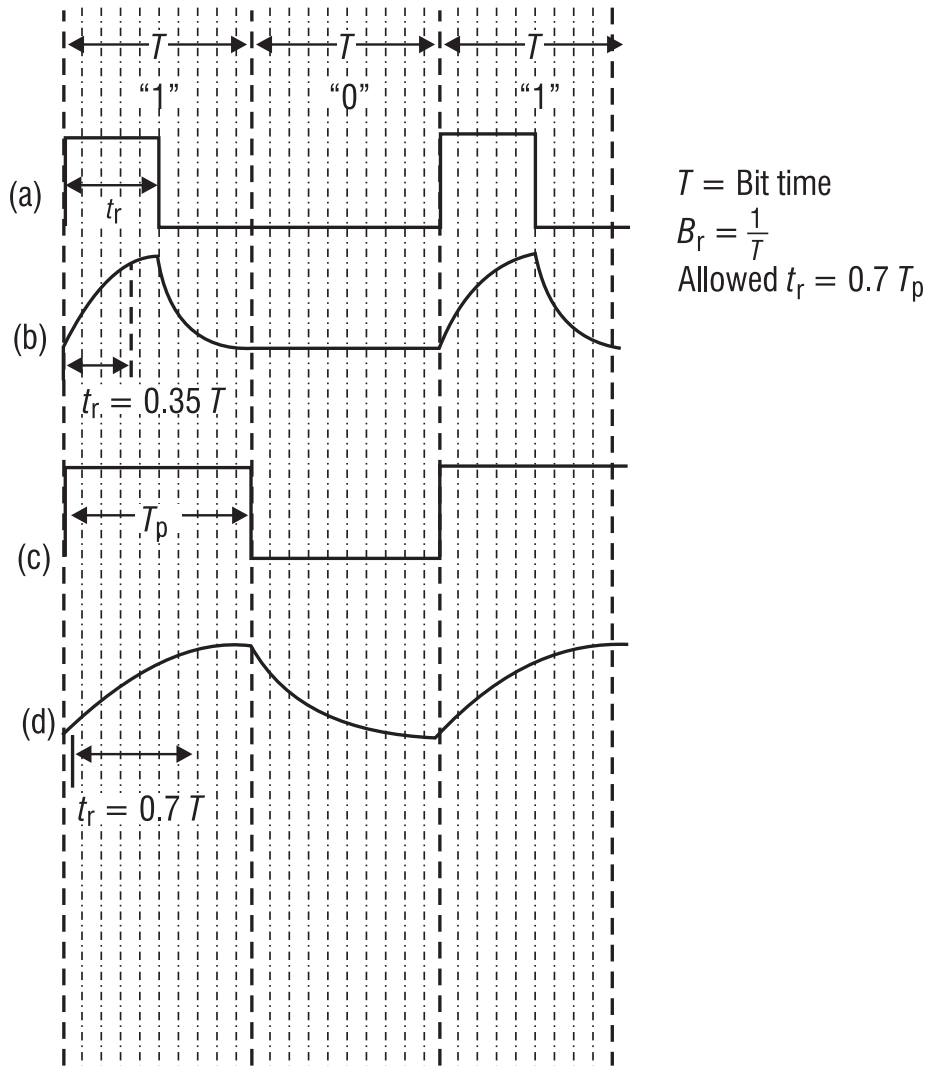


Figure 8-14 Effects of system rise time for RZ format and NRZ format:

- a) Transmitted RZ pulse train
- b) Received RZ signal with allowable t_r .
- c) Transmitted NRZ pulse train
- d) Received NRZ pulse train with allowable t_r

X. MULTIPLEXING

The purpose of multiplexing is to share the bandwidth of a single transmission channel among several users. Two multiplexing methods are commonly used in fiber optics:

1. Time-division multiplexing (TDM)
2. Wavelength-division multiplexing (WDM)

A. Time-Division Multiplexing (TDM)

In time-division multiplexing, time on the information channel, or fiber, is shared among the many data sources. The multiplexer MUX can be described as a type of “rotary switch,” which rotates at a very high speed, individually connecting each input to the communication channel for a fixed period of time. The process is reversed on the output with a device known as a demultiplexer, or DEMUX. After each channel has been sequentially connected, the process repeats itself. One complete cycle is known as a *frame*. To ensure that each channel on the input is connected to its corresponding channel on the output, start and stop frames are added to synchronize the input with the output. TDM systems may send information using any of the digital modulation schemes described (analog multiplexing systems also exist). This is illustrated in Figure 8-15.

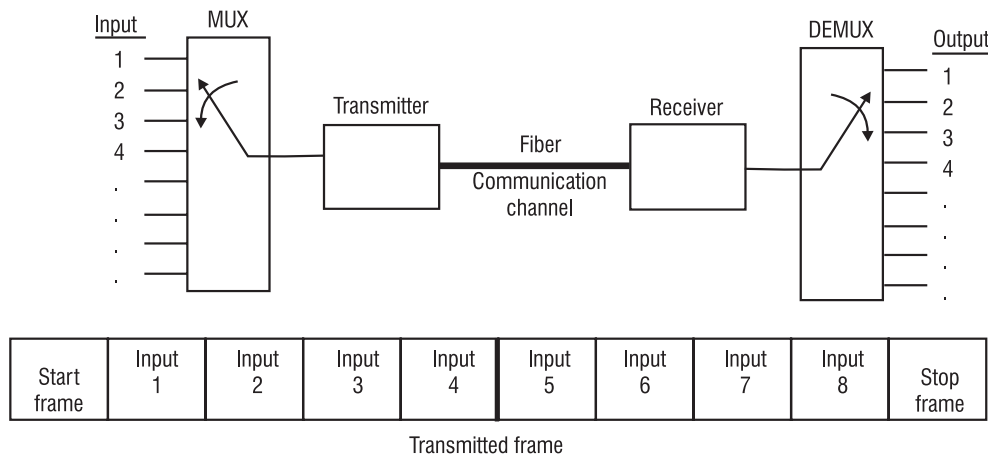


Figure 8-15 Time-division multiplexing system

The amount of data that can be transmitted using TDM is given by the MUX output rate and is defined by Equation 8-16.

$$\text{MUX output rate} = N \times \text{Maximum input rate} \quad (8-16)$$

where N is the number of input channels and the maximum input rate is the highest data rate in bits/second of the various inputs. The bandwidth of the communication channel must be *at least equal* to the MUX output rate. Another parameter commonly used in describing the information capacity of a TDM system is the *channel-switching rate*. This is equal to the number of inputs visited per second by the MUX and is defined as

$$\text{Channel switching rate} = \text{Input data rate} \times \text{Number of channels} \quad (8-17)$$

Example 6

A digital MUX operates with 8 sources. The rate of data in each source is 1000 bytes/s. Assume that 8-bits-per-byte data is transmitted byte by byte.

1. What is the data rate of the MUX output?
2. What is the channel switching rate?

Solution:

1. The data rate of each input channel is (8×1000) bits/s. The output data rate from Equation 8-16 is then:

$$\begin{aligned}\text{Output rate} &= N \times \text{Input rate} \\ &= 8 \times (8 \times 1000) = 64 \text{ kbits/s}\end{aligned}$$

2. Each channel must have access to the MUX 1000 times each second, transmitting 1 byte at a time. From Equation 8-17, the channel switching rate is

$$8 \times 1000 = 8,000 \text{ channels/s}$$

The Digital Telephone Hierarchy

The North American digital telephone hierarchy defines how the low-data-rate telephone signals are multiplexed together onto higher-speed lines. The system uses pulse code modulation (PCM) in conjunction with time-division multiplexing to achieve this. The basic digital multiplexing standard established in the United States is called *the Bell System Level 1 PCM Standard* or the *Bell T1 Standard*. This is the standard used for multiplexing 24 separate 64-kbps ($8 \text{ bits/sample} \times 8000 \text{ samples/s}$) voice channels together. Each 64-kbps voice channel is designated as *digital signaling level 0* or *DS-0*. Each frame in the 24-channel multiplexer consists of

$$8 \text{ bits/channel} \times 24 \text{ channels} + 1 \text{ framing bit} = 193 \text{ bits}$$

The total data rate when transmitting 24 channels is determined by:

$$193 \text{ bits/frame} \times 8000 \text{ frames/s} = 1.544 \text{ Mbps} = \text{T1 designation}$$

If four T1 lines are multiplexed together, we get

$$4 \times 24 \text{ channels} = 96 \text{ channels} = \text{T2 designation}$$

Multiplexing seven T2 lines together we get

$$7 \times 96 = 672 \text{ channels} = \text{T3 designation}$$

Figure 8-16 shows how the multiplexing takes place.

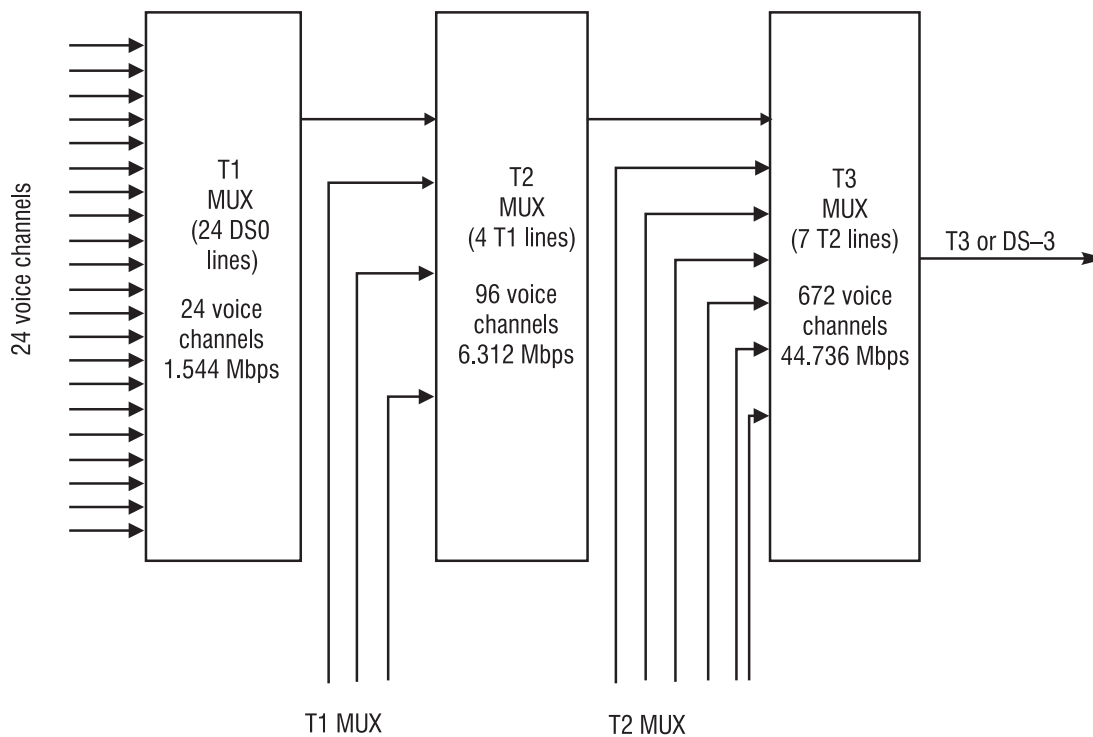


Figure 8-16 The North American digital telephone hierarchy

SONET

Fiber optics use *Synchronous Optical Network (SONET)* standards. The initial SONET designation is OC-1 (*optical carrier-1*). This level is known as synchronous transport level 1 (STS-1). It has a synchronous frame structure at a speed of 51.840 Mbps. The synchronous frame structure makes it easy to extract individual DS1 signals without disassembling the entire frame. OC-1 picks up where the DS3 signal (28 DSI signals or 672 channels) leaves off. With SONET standards any of these 28 T1 systems can be stripped out of the OC-1 signal.

The North American SONET rate is OC-48, which is 48 times the 51.840-Mbps OC-1 rate, or approximately 2.5 billion bits per second (2.5 Gbps). OC-48 systems can transmit 48×672 channels or 32,256 channels, as seen in Table 8-2. One fiber optic strand can carry all 32,256 separate 64-kbps channels. The maximum data rate specified for the SONET standard is OC-192 or approximately 9.9538 Gbps. At this data rate, 129,024 separate voice channels can be transmitted through a single fiber. Even though OC-192 is the maximum data rate specified by SONET, recent developments in technology allow for transmission as high as 40 Gbps. This, coupled with the availability of 32-channel wavelength-division multiplexers, has led to the development of systems capable of 1.2-terabit/s transmission. As can be seen, the data rates achievable through the use of fiber optics are dramatically greater than those achievable with copper. In addition, the distance between repeaters in a fiber optic system is considerably greater than that for copper, making fiber more reliable and, in most cases, more cost-effective.

Table 8-2 Digital Telephone Transmission Rates

Medium	Designation	Data Rate (Mbps)	Voice Channels	Repeater Spacing
Copper	DS-1	1.544	24	1-2 km
	DS-2	3.152	96	
	DS-3	44.736	672	
Fiber Optic	OC-1	51.84	672	50-100 km
	OC-3	155.52	2016	
	OC-12	622.08	8064	
	OC-18	933.12	12,096	
	OC-24	1244.16	16,128	
	OC-36	1866.24	24,192	
	OC-48	2488.32	32,256	
	OC-96	4976.64	64,512	
	OC-192	9953.28	129,024	

B. Wavelength-Division Multiplexing (WDM)

In wavelength-division multiplexing, each data channel is transmitted using a slightly different wavelength (different color). With use of a different wavelength for each channel, many channels can be transmitted through the same fiber without interference. This method is used to increase the capacity of existing fiber optic systems many times. Each WDM data channel may consist of a single data source or may be a combination of a single data source and a TDM (time-division multiplexing) and/or FDM (frequency-division multiplexing) signal. Dense wavelength-division multiplexing (DWDM) refers to the transmission of multiple closely spaced wavelengths through the same fiber. For any given wavelength λ and corresponding frequency f , the International Telecommunications Union (ITU) defines standard frequency spacing Δf as 100 GHz, which translates into a $\Delta\lambda$ of 0.8-nm wavelength spacing. This follows from the relationship $\Delta\lambda = \frac{\lambda \Delta f}{f}$. (See Table 8-3.) DWDM systems operate in the 1550-nm

window because of the low attenuation characteristics of glass at 1550 nm and the fact that erbium-doped fiber amplifiers (EDFA) operate in the 1530-nm–1570-nm range. Commercially available systems today can multiplex up to 128 individual wavelengths at 2.5 Gb/s or 32 individual wavelengths at 10 Gb/s (see Figure 8-17). Although the ITU grid specifies that each transmitted wavelength in a DWDM system is separated by 100 GHz, systems currently under development have been demonstrated that reduce the channel spacing to 50 GHz and below (< 0.4 nm). As the channel spacing decreases, the number of channels that can be transmitted increases, thus further increasing the transmission capacity of the system.

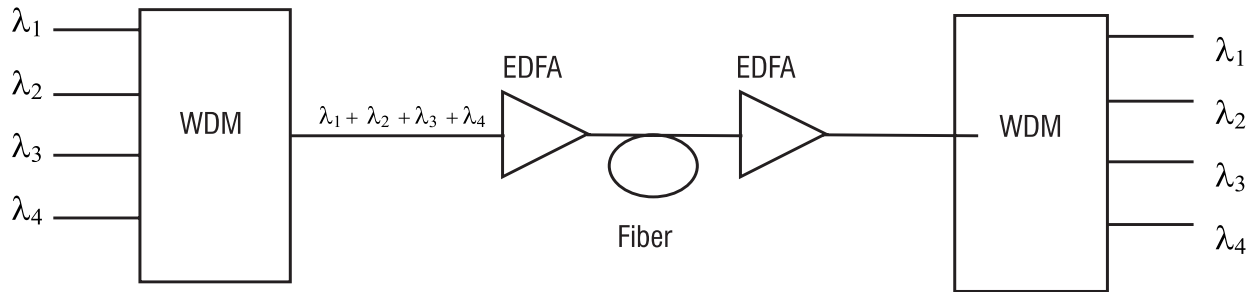


Figure 8-17 Wavelength-division multiplexing

Table 8-3 ITU GRID

Center Wavelength – nm (vacuum)	Optical Frequency (THz)		
1530.33	195.9	1546.92	193.8
1531.12	195.8	1547.72	193.7
1531.90	195.7	1548.51	193.6
1532.68	195.6	1549.32	193.5
1533.47	195.5	1550.12	193.4
1534.25	195.4	1550.92	193.3
1535.04	195.3	1551.72	193.2
1535.82	195.2	1552.52	193.1
1536.61	195.1	1553.33	193.0
1537.40	195.0	1554.13	192.9
1538.19	194.9	1554.93	192.8
1538.98	194.8	1555.75	192.7
1539.77	194.7	1556.55	192.6
1540.56	194.6	1557.36	192.5
1541.35	194.5	1558.17	192.4
1542.14	194.4	1558.98	192.3
1542.94	194.3	1559.79	192.2
1543.73	194.2	1560.61	192.1
1544.53	194.1	1561.42	192.0
1545.32	194.0	1562.23	191.9
1546.12	193.9	1563.05	191.8
		1563.86	191.7

XI. COMPONENTS—FIBER OPTIC CABLE

In most applications, optical fiber must be protected from the environment using a variety of different cabling types based on the type of environment in which the fiber will be used. Cabling provides the fiber with protection from the elements, added tensile strength for pulling, rigidity for bending, and durability. In general, fiber optic cable can be separated into two types: indoor and outdoor.

Indoor Cables

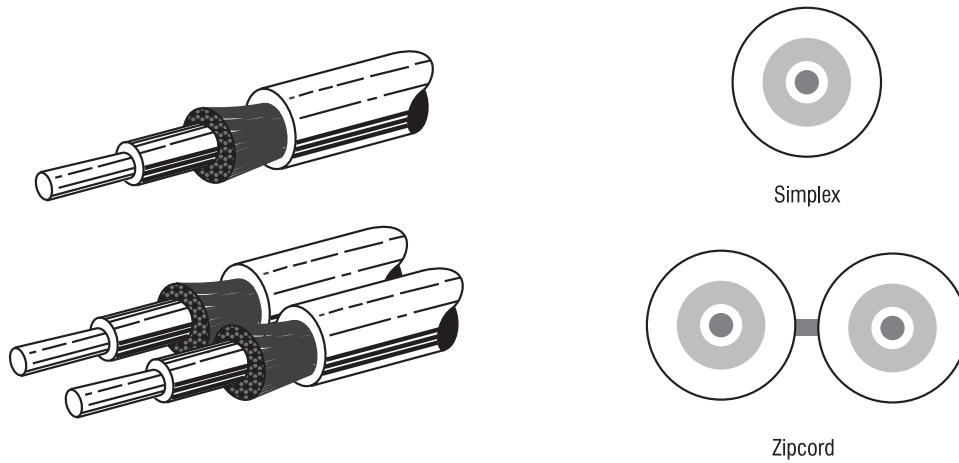
- *Simplex cable*—contains a single fiber for one-way communication
- *Duplex cable*—contains two fibers for two-way communication
- *Multifiber cable*—contains more than two fibers. Fibers are usually in pairs for duplex operation. A ten-fiber cable permits five duplex circuits.
- *Breakout cable*—typically has several individual simplex cables inside an outer jacket. The outer jacket includes a zipcord to allow easy access
- *Heavy-, light-, and plenum-duty and riser cable*
 - Heavy-duty cables have thicker jackets than light-duty cable, for rougher handling.
 - Plenum cables are jacketed with low-smoke and fire-retardant materials.
 - Riser cables run vertically between floors and must be engineered to prevent fires from spreading between floors.

Outdoor Cables

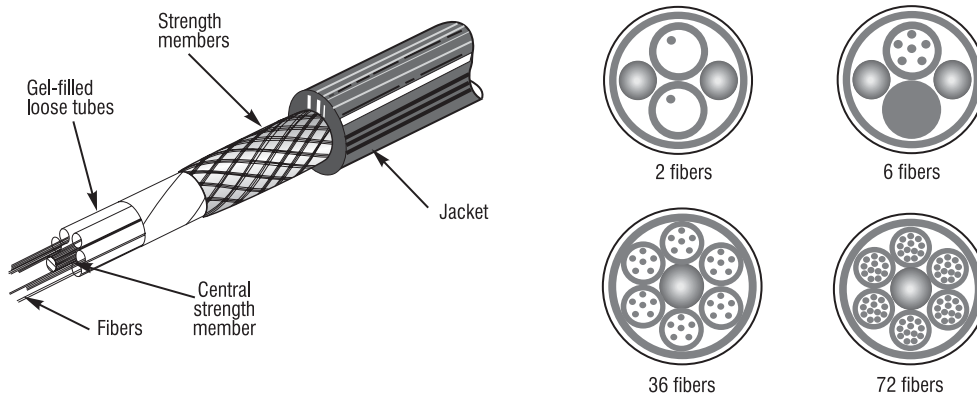
Outdoor cables must withstand harsher environmental conditions than indoor cables. Outdoor cables are used in applications such as:

- *Overhead*—cables strung from telephone lines
- *Direct burial*—cables placed directly in trenches
- *Indirect burial*—cables placed in conduits
- *Submarine*—underwater cables, including transoceanic applications

Sketches of indoor and outdoor cables are shown in Figure 8-18.



a) Indoor simplex and duplex cable (Courtesy of General Photonics)



b) Outdoor loose buffer cable (Courtesy of Siecor)

Figure 8-18 Indoor and outdoor cable

Cabling Example

Figure 8-19 shows an example of an interbuilding cabling scenario

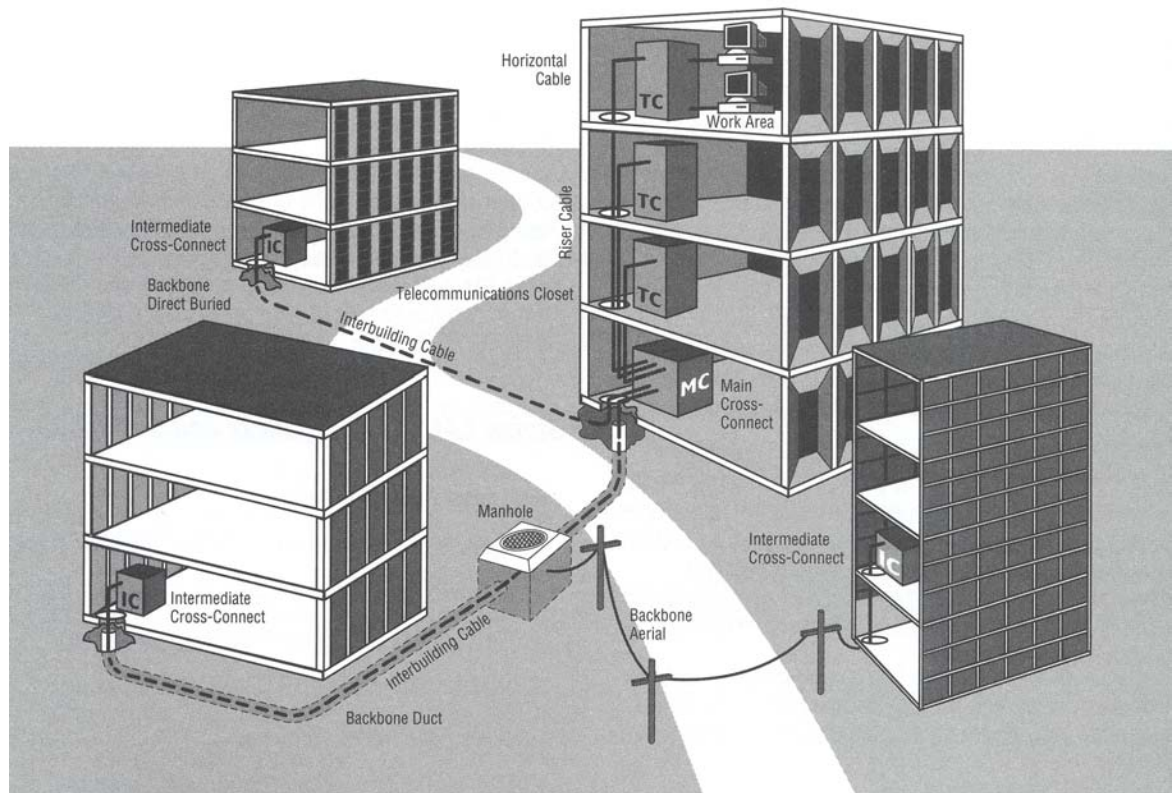


Figure 8-19 Interbuilding cabling scenario (Courtesy of Siecor)

XII. FIBER OPTIC SOURCES

Two basic light sources are used for fiber optics: laser diodes (LD) and light-emitting diodes (LED). Each device has its own advantages and disadvantages as listed in Table 8-4.

Table 8-4 LED Versus Laser

Characteristic	LED	Laser
Output power	Lower	Higher
Spectral width	Wider	Narrower
Numerical aperture	Larger	Smaller
Speed	Slower	Faster
Cost	Less	More
Ease of operation	Easier	More difficult

Fiber optic sources must operate in the low-loss transmission windows of glass fiber. LEDs are typically used at the 850-nm and 1310-nm transmission wavelengths, whereas lasers are primarily used at 1310 nm and 1550 nm.

LEDs are typically used in lower-data-rate, shorter-distance multimode systems because of their inherent bandwidth limitations and lower output power. They are used in applications in which data rates are in the hundreds of megahertz as opposed to GHz data rates associated with lasers. Two basic structures for LEDs are used in fiber optic systems: *surface-emitting* and *edge-emitting* as shown in Figure 8-20.

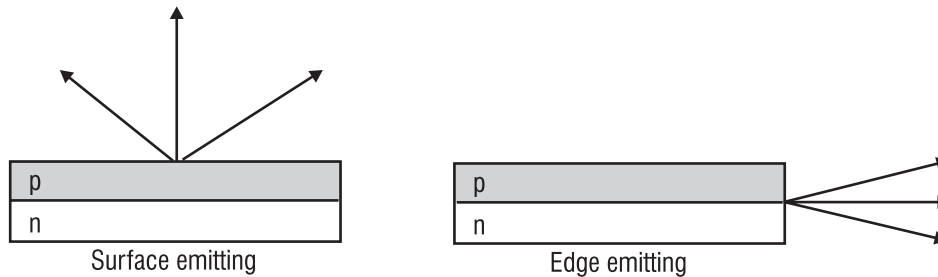
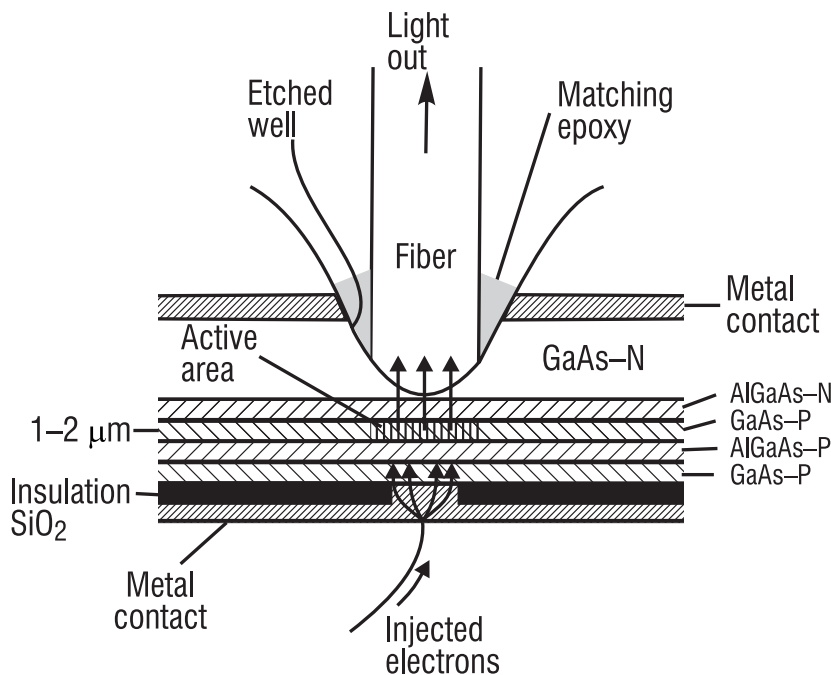


Figure 8-20 Surface-emitting versus edge-emitting diodes

In surface-emitting LEDs the radiation emanates from the surface. An example of this is the Burrus diode as shown in Figure 8-21. LEDs typically have large numerical apertures, which



Source: C. A. Burrus and B. I. Miller, "Small Area Double-Heterostructure Aluminum Gallium Arsenide Electroluminescent Diode Sources for Optical Fiber Transmission Lines," *Optical Communications* 4:307-69 (1971).

Figure 8-21 Burrus diode

makes light coupling into single-mode fiber difficult due to the fiber's small N.A. and core diameter. For this reason LEDs are most often used with multimode fiber. LEDs are used in lower-data-rate, shorter-distance multimode systems because of their inherent bandwidth limitations and lower output power. The output spectrum of a typical LED is about 40 nm, which limits its performance because of severe chromatic dispersion. LEDs operate in a more linear fashion than do laser diodes. This makes them more suitable for analog modulation. Figure 8-22 shows a graph of typical output power versus drive current for LEDs and laser diodes. Notice that the LED has a more linear output power, which makes it more suitable for analog modulation. Often these devices are pigtailed, having a fiber attached during the manufacturing process. Some LEDs are available with connector-ready housings that allow a connectorized fiber to be directly attached. They are also relatively inexpensive. Typical applications are local area networks, closed-circuit TV, and transmitting information in areas where EMI may be a problem.

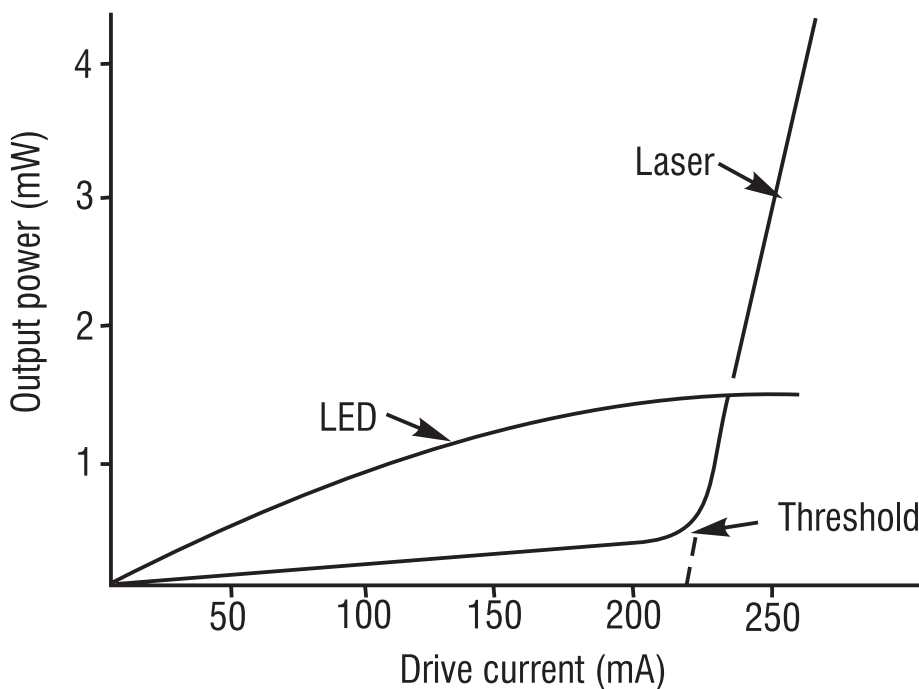


Figure 8-22 Drive current versus output power for LED and laser (Courtesy of AMP, Inc.)

Laser diodes (LD) are used in applications in which longer distances and higher data rates are required. Because an LD has a much higher output power than an LED, it is capable of transmitting information over longer distances. Consequently, and given the fact that the LD has a much narrower spectral width, it can provide high-bandwidth communication over long distances. The LD's smaller N.A. also allows it to be more effectively coupled with single-mode fiber. The difficulty with LDs is that they are inherently nonlinear, which makes analog transmission more difficult. They are also very sensitive to fluctuations in temperature and drive current, which causes their output wavelength to drift. In applications such as wavelength-division multiplexing in which several wavelengths are being transmitted down the same fiber, the stability of the source becomes critical. This usually requires complex circuitry and

feedback mechanisms to detect and correct for drifts in wavelength. The benefits, however, of high-speed transmission using LDs typically outweigh the drawbacks and added expense.

Laser diodes can be divided into two generic types depending on the method of confinement of the lasing mode in the lateral direction.

- *Gain-guided* laser diodes work by controlling the width of the drive-current distribution; this limits the area in which lasing action can occur. Because of different confinement mechanisms in the lateral and vertical directions, the emitted wavefront from these devices has a different curvature in the two perpendicular directions. This astigmatism in the output beam is one of the unique properties of laser-diode sources. Gain-guided injection laser diodes usually emit multiple longitudinal modes and sometimes multiple transverse modes. The optical spectrum of these devices ranges up to about 2 nm in width, thereby limiting their coherence length.
- *Index-guided* laser diodes use refractive index steps to confine the lasing mode in both the transverse and vertical directions. Index guiding also generally leads to both single transverse-mode and single longitudinal-mode behavior. Typical linewidths are on the order of 0.01 nm. Index-guided lasers tend to have less difference between the two perpendicular divergence angles than do gain-guided lasers.

Single-frequency laser diodes are another interesting member of the laser diode family. These devices are now available to meet the requirements for high-bandwidth communication. Other advantages of these structures are lower threshold currents and lower power requirements. One variety of this type of structure is the *distributed-feedback (DFB)* laser diode (Figure 8-23). With introduction of a corrugated structure into the cavity of the laser, only light of a very specific wavelength is diffracted and allowed to oscillate. This yields output wavelengths that are extremely narrow—a characteristic required for DWDM systems in which many closely spaced wavelengths are transmitted through the same fiber. Distributed-feedback lasers have been developed to emit light at fiber optic communication wavelengths between 1300 nm and 1550 nm.

Grating limits emission to one frequency.

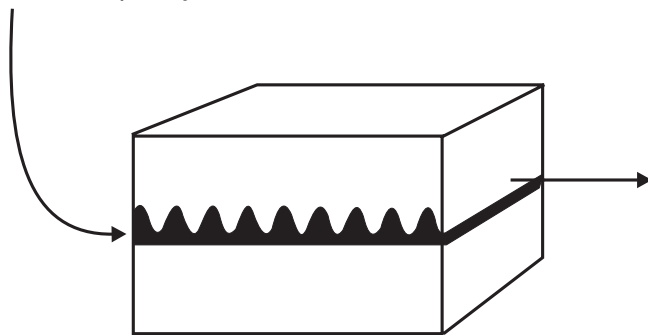


Figure 8-23 *Distributed-feedback laser*

XIII. PACKAGING

Laser diodes are available in a variety of packages. Most have monitoring photodiodes integrated with the packages. Because lasers inherently emit light from both ends of the cavity, a photodiode can be placed on one end to monitor and maintain the output power at a certain level. One of the most popular types of packages is the TO-can style (Figure 8-24) available in both 5.6-mm and 9-mm-diameter sizes. Either style can be purchased with connectorized fiber pigtailed for convenience. Devices used in telecommunication typically come in either 14-pin *butterfly* or *dual-in-line* (DIL) packages as shown in Figures 8-25 and 8-26. These devices typically include thermoelectric coolers (TEC) and mounting plates for heat-sinking.

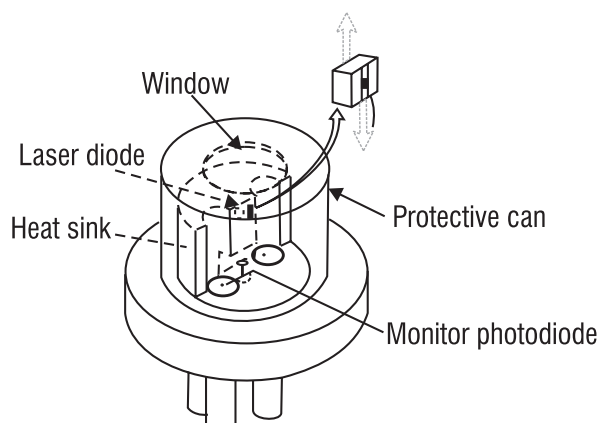


Figure 8-24 Laser diode in TO-can style package (Courtesy of Newport Corp.)

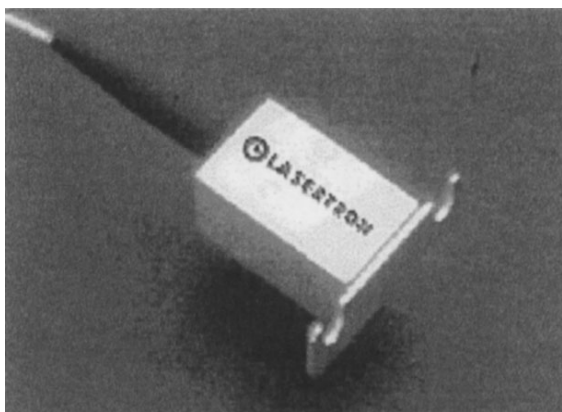


Figure 8-25 14-pin DIL package (Courtesy of Lasertron)

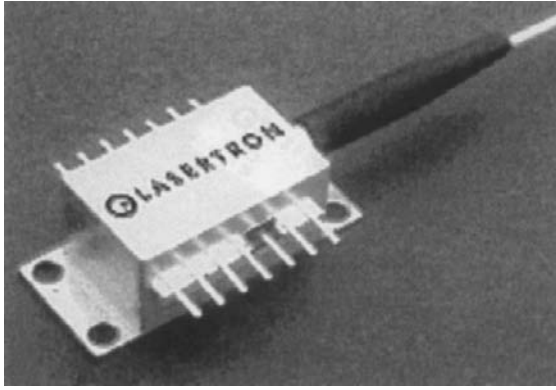


Figure 8-26 1550-nm DFB laser in butterfly package (Courtesy of Lasertron)

XIV. DIRECT VERSUS EXTERNAL MODULATION

Lasers and LEDs used in telecommunication applications are modulated using one of two methods: *direct modulation* or *external modulation*.

- In direct modulation (Figure 8-27), the output power of the device varies directly with the input drive current. Both LEDs and lasers can be directly modulated using analog and digital signals. The benefit of direct modulation is that it is simple and cheap. The disadvantage is that it is slower than indirect modulation with limits of less than approximately 3 GHz.

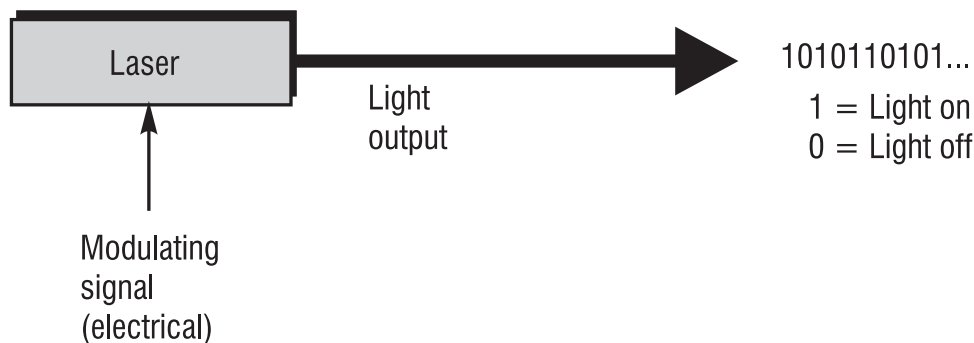


Figure 8-27 Direct modulation

- In external modulation (Figure 8-28), an external device is used to modulate the intensity or phase of the light source. The light source remains on while the external modulator acts like a “shutter” controlled by the information being transmitted. External modulation is typically used in high-speed applications such as long-haul telecommunication or cable TV head ends. The benefits of external modulation are that it is much faster and can be used with higher-power laser sources. The disadvantage is that it is more expensive and requires complex circuitry to handle the high frequency RF modulation signal.

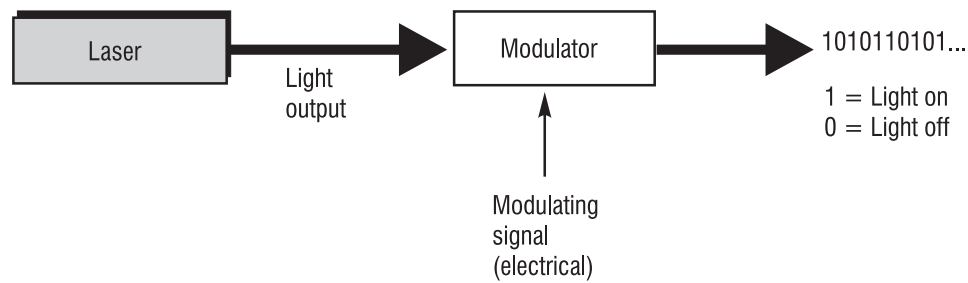


Figure 8-28 External modulation

External modulation is typically accomplished using an integrated optical modulator that incorporates a waveguide Mach-Zehnder interferometer fabricated on a slab of lithium niobate (LiNbO_3). The waveguide is created using a lithographic process similar to that used in the manufacturing of semiconductors. The waveguide region is slightly doped with impurities to increase the index of refraction so that the light is guided through the device (Figure 8-29).

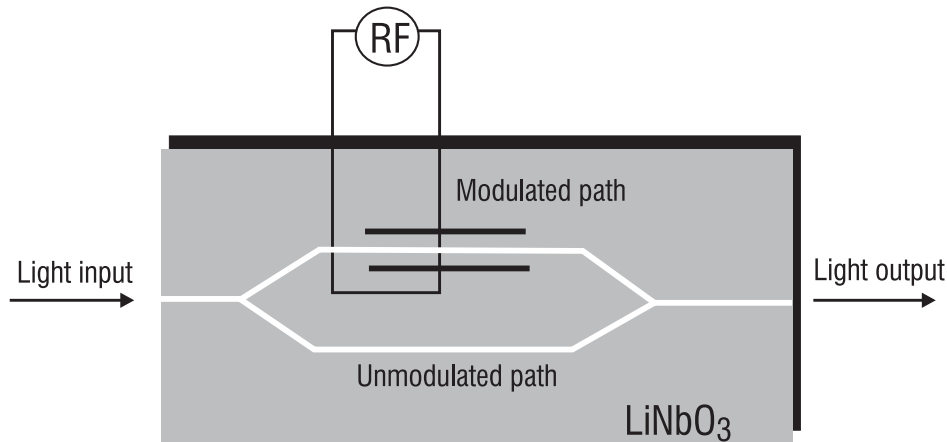


Figure 8-29 External modulation using Mach-Zehnder waveguide interferometer

Light entering the modulator (via fiber pigtail) is split into two paths. One path is unchanged or unmodulated. The other path has electrodes placed across it. Because LiNbO_3 is an *electro-optic material*, when a voltage is placed across the waveguide its index of refraction is changed, causing a phase delay proportional to the amplitude of the applied voltage. When the light is then recombined, the two waves interfere with one another. If the two waves are in phase, the interference is constructive and the output is on. If the two waves are out of phase, the interference is destructive and the waves cancel each other. The input voltage associated with a 180° phase shift is known as V_π . The induced phase shift can be calculated using:

$$\text{Phase shift} = \Delta\theta = 180^\circ \times V_{\text{in}}/V_\pi \quad (8-18)$$

where V_{in} is the voltage applied to the modulator. Lithium niobate modulators are well developed and used extensively in both CATV and telecommunication applications. Devices are available at both the 1310-nm and 1550-nm wavelengths.

XV. FIBER OPTIC DETECTORS

The purpose of a fiber optic detector is to convert light emanating from the optical fiber back into an electrical signal. The choice of a fiber optic detector depends on several factors including wavelength, responsivity, and speed or rise time. Figure 8-30 depicts the various types of detectors and their spectral responses.

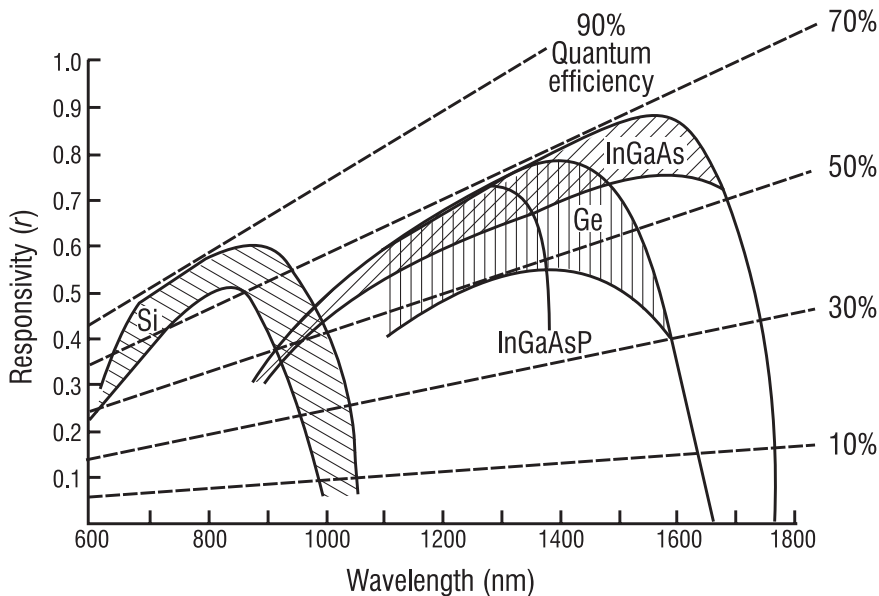


Figure 8-30 *Detector spectral response*

The process by which light is converted into an electrical signal is the opposite of the process that produces the light. Light striking the detector generates a small electrical current that is amplified by an external circuit. Absorbed photons excite electrons from the valence band to the conduction band, resulting in the creation of an electron-hole pair. Under the influence of a bias voltage these carriers move through the material and induce a current in the external circuit. For each electron-hole pair created, the result is an electron flowing in the circuit. Typical current levels are small and require some amplification as shown in Figure 8-31.

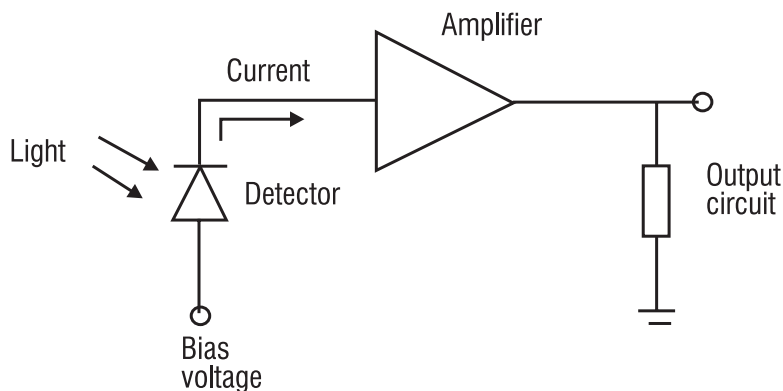


Figure 8-31 *Typical detector amplifier circuit*

The most commonly used photodetectors are the PIN and avalanche photodiodes (APD). The material composition of the device determines the wavelength sensitivity. In general, silicon devices are used for detection in the visible portion of the spectrum; InGaAs crystal are used in the near-infrared portion of the spectrum between 1000 nm and 1700 nm, and germanium PIN and APDs are used between 800 nm and 1500 nm. Table 8-5 gives some typical photodetector characteristics:

Table 8-5 Typical Photodetector Characteristics

Photodetector	Wavelength (nm)	Responsivity (A/W)	Dark Current (nA)	Rise Time (ns)
Silicon PN	550–850	0.4–0.7	1–5	5–10
Silicon PIN	850–950	0.6–0.8	10	0.070
InGaAs PIN	1310–1550	0.85	0.5–1.0	0.005–5
InGaAs APD	1310–1550	0.80	30	0.100
Germanium	1000–1500	0.70	1000	12

Some of the more important detector parameters listed below are defined and described in Module 1-6, *Optical Detectors and Human Vision*.

Responsivity—the ratio of the electrical power to the detector's output optical power

Quantum efficiency—the ratio of the number of electrons generated by the detector to the number of photons incident on the detector

$$\text{Quantum efficiency} = (\text{Number of electrons})/\text{Photon}$$

Dark current—the amount of current generated by the detector with no light applied. Dark current increases about 10% for each temperature increase of 1°C and is much more prominent in Ge and InGaAs at longer wavelengths than in silicon at shorter wavelengths.

Noise floor—minimum detectable power that a detector can handle. The noise floor is related to the dark current since the dark current will set the lower limit.

$$\text{Noise floor} = \text{Noise (A)}/\text{Responsivity (A/W)}$$

Response time—the time required for the detector to respond to an optical input. The response time is related to the bandwidth of the detector by

$$\text{BW} = 0.35/t_r$$

where t_r is the rise time of the device. The rise time is the time required for the detector to rise to a value equal to 63.2% of its final steady-state reading.

Noise equivalent power (NEP)—at a given modulation frequency, wavelength, and noise bandwidth, the incident radiant power that produces a signal-to-noise ratio of *one* at the output of the detector (Source: Electronic Industry Association—EIA)

XVI. FIBER OPTIC SYSTEM DESIGN CONSIDERATIONS

When designing a fiber optic communication system some of the following factors must be taken into consideration:

- Which modulation and multiplexing technique is best suited for the particular application?
- Is enough power available at the receiver (power budget)?
- Rise-time and bandwidth characteristics
- Noise effects on system bandwidth, data rate, and bit error rate
- Are erbium-doped fiber amplifiers required?
- What type of fiber is best suited for the application?
- Cost

A. Power Budget

The power arriving at the detector must be sufficient to allow clean detection with few errors. Clearly, the signal at the receiver must be larger than the noise. The power at the detector, P_r , must be above the threshold level or *receiver sensitivity* P_s .

$$P_r \geq P_s \quad (8-19)$$

The receiver sensitivity P_s is the signal power, in dBm, at the receiver that results in a particular bit error rate (BER). Typically the BER is chosen to be one error in 10^9 bits or 10^{-9} .

Example 7

A receiver has sensitivity P_s of -45 dBm and a BER of 10^{-9} . What is the minimum power that must be incident on the detector?

Solution: Use Equation 8-3 to find the source power in milliwatts, given the power sensitivity in dBm. Thus,

$$-45 \text{ dBm} = 10 \log \left(\frac{P}{1 \text{ mW}} \right)$$

so that $P = (1 \text{ mW}) \times 10^{-4.5} = 3.16 \times 10^{-5} \text{ mW} = 31.6 \text{ nanowatts}$
for a probability of error of 1 in 10^9 .

The received power at the detector is a function of:

1. Power emanating from the light source (laser diode or LED)—(P_L)
2. Source to fiber loss (L_{sf})
3. Fiber loss per km (F_L) for a length of fiber (L)
4. Connector or splice losses (L_{conn})
5. Fiber to detector loss (L_{fd})

The allocation of power loss among system components is the *power budget*. The *power margin* is the difference between the received power P_r and the receiver sensitivity P_s by some margin L_m .

$$L_m = P_r - P_s \quad (8-20)$$

where L_m is the loss margin in dB

P_r is the received power

P_s is the receiver sensitivity in dBm

If all of the loss mechanisms in the system are taken into consideration, the loss margin can be expressed as Equation 8-21.

$$L_m = P_L - L_{sf} - (F_L \times L) - L_{conn} - L_{fd} - P_s \quad (8-21)$$

All units are dB and dBm.

Example 8

A system has the following characteristics:

LED power (P_L) = 2 mW (3 dBm)

LED to fiber loss (L_{sf}) = 3 dB

Fiber loss per km (F_L) = 0.5 dB/km

Fiber length (L) = 40 km

Connector loss (L_{conn}) = 1 dB (one connector between two 20-m fiber lengths)

Fiber to detector loss (L_{fd}) = 3 dB

Receiver sensitivity (P_s) = -36 dBm

Find the loss margin.

Solution:

$$L_m = 3 \text{ dBm} - 3 \text{ dB} - (40 \text{ km} \times 0.5 \text{ dB/km}) - 1 \text{ dB} - 3 \text{ dB} - (-36 \text{ dBm}) = 12 \text{ dB}$$

This particular fiber optic loss budget is illustrated in Figure 8-32, with each loss graphically depicted.

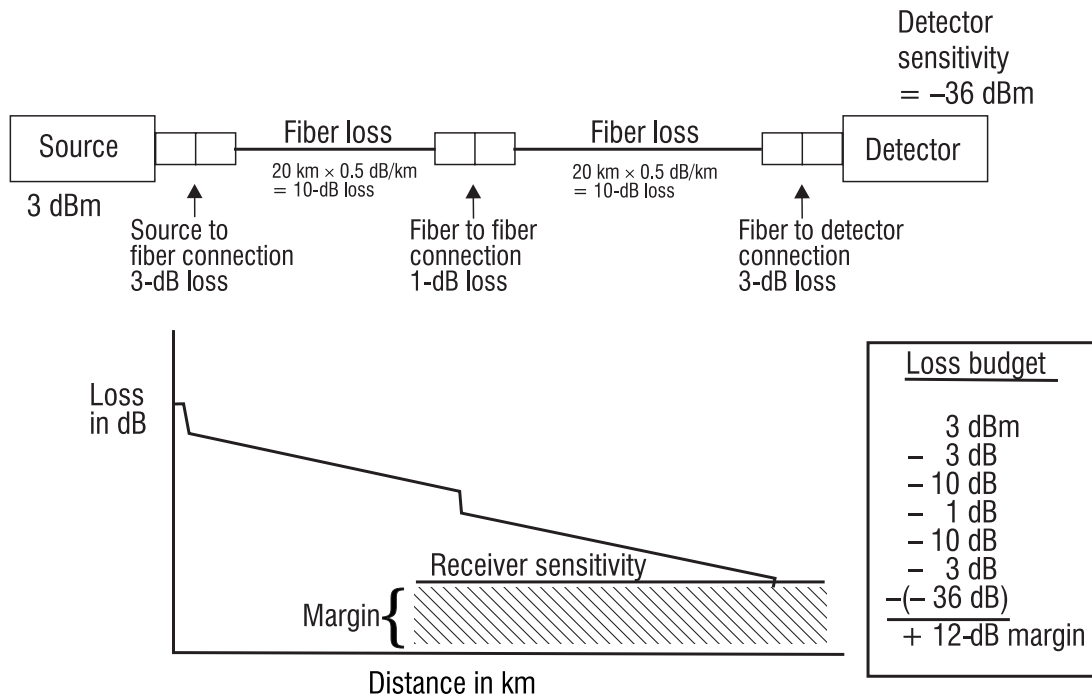


Figure 8-32 Fiber optic loss budget

B. Bandwidth and Rise Time Budgets

The transmission data rate of a digital fiber optic communication system is limited by the rise time of the various components, such as amplifiers and LEDs, and the dispersion of the fiber. The cumulative effect of all the components should not limit the bandwidth of the system. The rise time t_r and bandwidth BW are related by

$$BW = 0.35/t_r \quad (8-22)$$

This equation is used to determine the required system rise time. The appropriate components are then selected to meet the system rise time requirements. The relationship between total system rise time and component rise time is given by Equation 8-23

$$t_s = (t_{r1}^2 + t_{r2}^2 + t_{r3}^2 + \dots)^{1/2} \quad (8-23)$$

where t_s is the total system rise time and t_{r1} , t_{r2} , ... are the rise times associated with the various components.

To simplify matters, divide the system into five groups:

1. Transmitting circuits (t_{tc})
2. LED or laser (t_L)
3. Fiber dispersion (t_f)
4. Photodiode (t_{ph})
5. Receiver circuits (t_{rc})

The system rise time can then be expressed as

$$t_s = (t_{tc}^2 + t_L^2 + t_f^2 + t_{ph}^2 + t_{rc}^2)^{1/2} \quad (8-24)$$

The system bandwidth can then be calculated using Equation 8-25 from the total rise time t_s as given in Equation 8-24.

$$BW = 0.35/t_s \quad (8-25)$$

Electrical and Optical Bandwidth

- *Electrical bandwidth* (BW_{el}) is defined as the frequency at which the ratio *current out/current in* (I_{out}/I_{in}) drops to 0.707. (Analog systems are usually specified in terms of electrical bandwidth.)
- *Optical bandwidth* (BW_{opt}) is the frequency at which the ratio *power out/power in* (P_{out}/P_{in}) drops to 0.5.

Because P_{in} and P_{out} are directly proportional to I_{in} and I_{out} (not I_{in}^2 and I_{out}^2), the half-power point is equivalent to the half-current point. This results in a BW_{opt} that is larger than the BW_{el} as given in Equation 8-26.

$$BW_{el} = 0.707 \times BW_{opt} \quad (8-26)$$

Example 9

A 10-km fiber with a $BW \times \text{length}$ product of 1000 MHz $\times$ km (optical bandwidth) is used in a communication system. The rise times of the other components are $t_{tc} = 10$ ns, $t_L = 2$ ns, $t_{ph} = 3$ ns, and $t_{rc} = 12$ ns. Calculate the *electrical* BW for the system.

Solution:

Because we are looking for the electrical BW, first calculate the electrical BW of the 2-km fiber from the optical BW and then calculate the rise time $t_r = t_f$.

$$BW_{opt} = (1000 \text{ MHz} \times \text{km})/10 \text{ km} = 100 \text{ MHz}$$

$$BW_{el} = 0.707 \times 100 \text{ MHz} = 70.7 \text{ MHz}$$

The fiber rise time is

$$t_r = t_f = 0.35/(70.7 \text{ MHz}) = 4.95 \text{ ns}$$

The system rise time is

$$t_s = (10^2 + 2^2 + 4.95^2 + 3^2 + 12^2)^{1/2} = 16.8 \text{ ns}$$

System BW_{el} is

$$BW_{el} = 0.35/(16.8 \times 10^{-9}) = 20.8 \text{ Mhz}$$

C. Connectors

Many types of connectors are available for fiber optics, depending on the application. The most popular are:

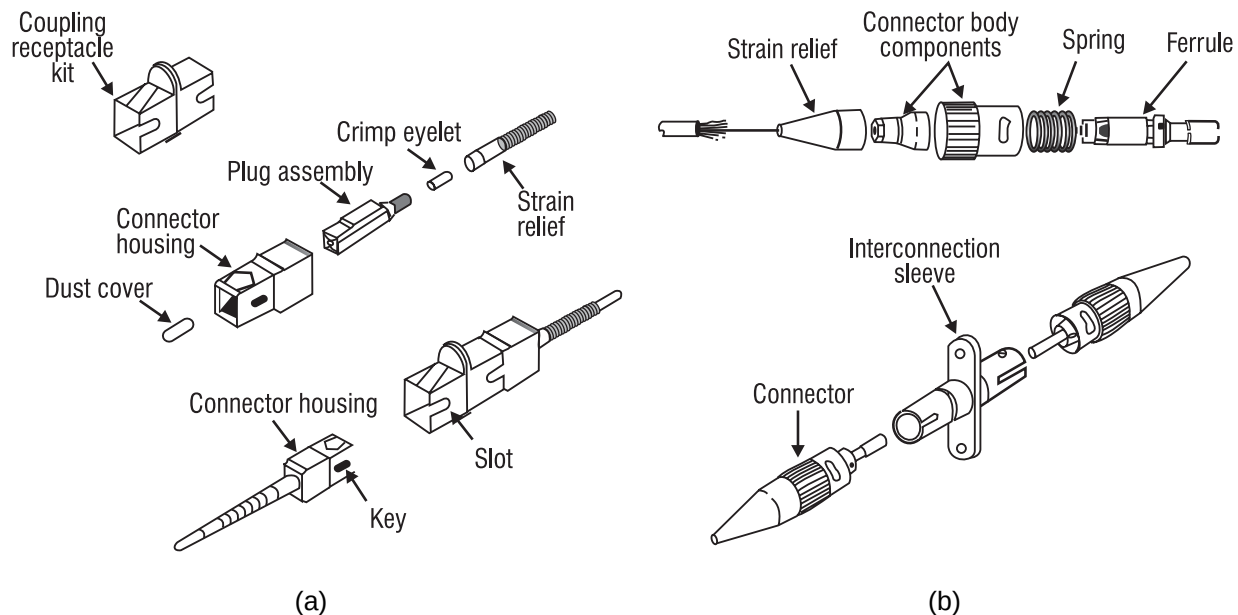
SC—snap-in single-fiber connector

ST and FC—twist-on single-fiber connector

FDDI—fiber distributed data interface connector

In the 1980s, there were many different types and manufacturers of connectors. Today, the industry has shifted to standardized connector types, with details specified by organizations such as the Telecommunications Industry Association, the International Electrotechnical Commission, and the Electronic Industry Association.

Snap-in connector (SC)—developed by Nippon Telegraph and Telephone of Japan. Like most fiber connectors, it is built around a cylindrical ferrule that holds the fiber, and it mates with an interconnection adapter or coupling receptacle. A push on the connector latches it into place, with no need to turn it in a tight space, so a simple tug will not unplug it. It has a square cross section that allows high packing density on patch panels and makes it easy to package in a polarized duplex form that ensures the fibers are matched to the proper fibers in the mated connector (Figure 8-33a).



Courtesy of Siecor, Inc.

Figure 8-33 (a) SC connector (b) ST connector

Twist-on single-fiber connectors (ST and FC)—long used in data communication; one of several fiber connectors that evolved from designs originally used for copper coaxial cables (see Figure 8-33b)

Duplex connectors—A duplex connector includes a pair of fibers and generally has an internal key so it can be mated in only one orientation. Polarizing the connector in this way is important

because most systems use separate fibers to carry signals in each direction, so it matters which fibers are connected. One simple type of duplex connector is a pair of SC connectors, mounted side by side in a single case. This takes advantage of their plug-in-lock design.

Other duplex connectors have been developed for specific types of networks, as part of comprehensive standards. One example is the fixed-shroud duplex (FSD) connector specified by the fiber distributed data interface (FDDI) standard (see Figure 8-34).

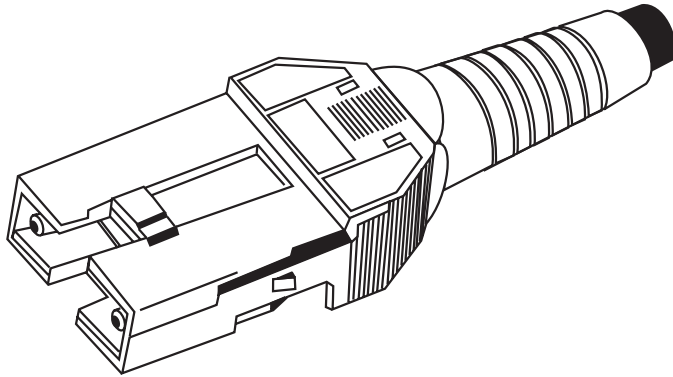


Figure 8-34 FDDI connector

D. Fiber Optic Couplers

A fiber optic coupler is a device used to connect a single (or multiple) fiber to many other separate fibers. There are two general categories of couplers:

- Star couplers (Figure 8-35a)
- T-couplers (Figure 8-35b)

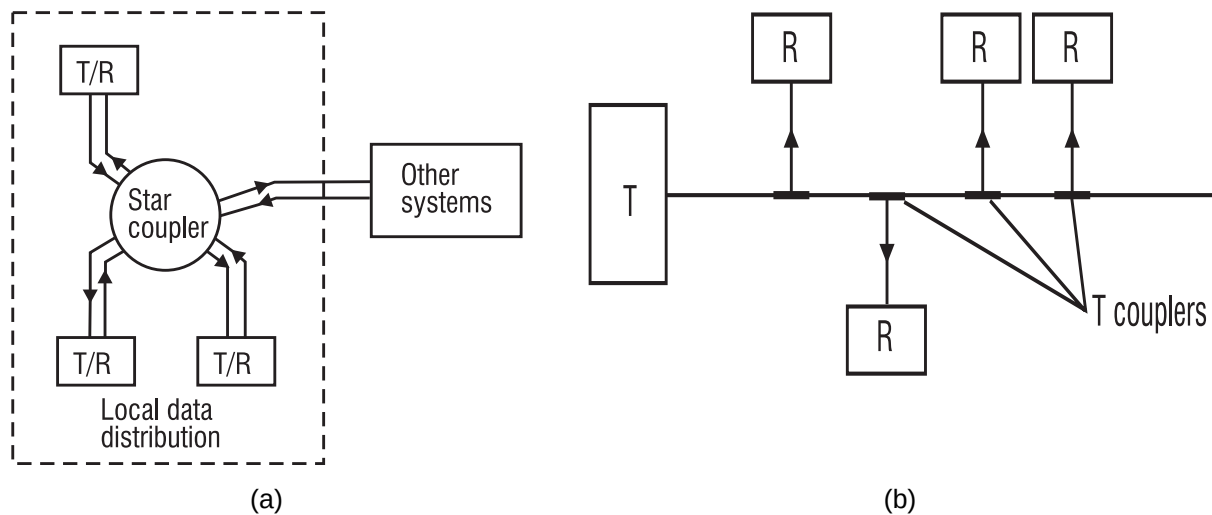


Figure 8-35 (a) Star coupler (b) T-coupler

Transmissive type

Optical signals sent into a mixing block are available at all output fibers (Figure 8-36). Power is distributed evenly. For an $n \times n$ star coupler (n -inputs and n -outputs), the power available at each output fiber is $1/n$ the power of any input fiber.

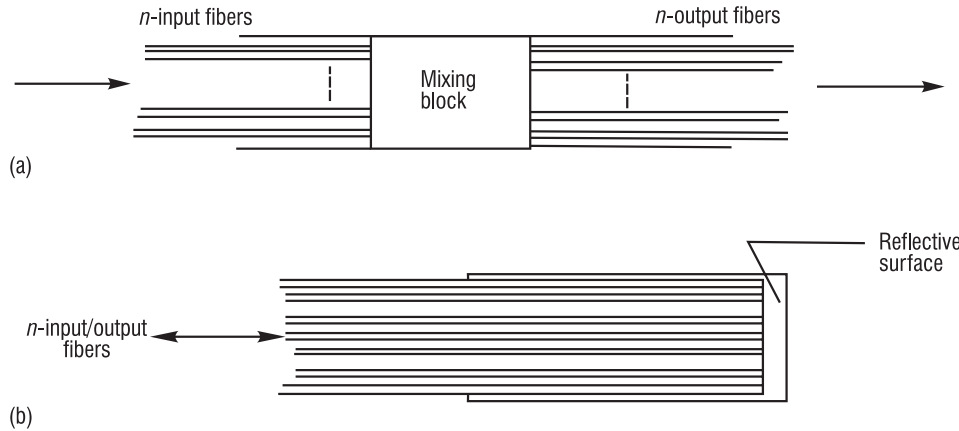


Figure 8-36 Star couplers (a) Transmissive (b) Reflective

The output power from a star coupler is simply

$$P_o = P_{in}/n \quad (8-27)$$

where n = number of output fibers.

The **power division** (power splitting ratio) in decibels is given by Equation 8-28.

$$PD_{st}(\text{dB}) = -10 \log(1/n) \quad (8-28)$$

The power division in decibels gives the number of decibels apparently lost in the coupler from single input fiber to single fiber output. **Excess power loss** (Loss_{ex}) is the power lost from input to total output, as given in Equation 8-29 or 8-30.

$$\text{Loss}_{ex} = \frac{P_{out}(\text{total})}{P_{in}} \quad (8-29)$$

$$\text{Loss}_{ex/\text{dB}} = -10 \log \frac{P_{out}(\text{total})}{P_{in}} \quad (8-30)$$

Example 10

An 8×8 star coupler is used in a fiber optic system to connect the signal from one computer to eight terminals. If the power at an input fiber to the star coupler is 0.5 mW, find (1) the power at each output fiber and (2) the power division in decibels.

Solution:

1. The 0.5-mW input is distributed to eight fibers. Each has $(0.50 \text{ mW})/8 = 0.0625 \text{ mW}$.
2. The power division, in decibels, from Equation 8-28 is

$$PD_{ST} = -10 \times \log(1/8) = 9.03 \text{ dB}$$

Example 11

A 10×10 star coupler is used to distribute the 3-dBm power of a laser diode to 10 fibers. The excess loss ($Loss_{ex}$) of the coupler is 2 dB. Find the power at each output fiber in dBm and μW .

Solution:

The power division in dB from Equation 8.28 is

$$PD_{st} = -10 \times \log(1/10) = 10 \text{ dB}$$

To find P_{out} for each fiber, subtract PD_{st} and $Loss_{ex}$ from P_{in} in dBm:

$$3 \text{ dBm} - 10 \text{ dB} - 2 \text{ dB} = -9 \text{ dBm}$$

To find P_{out} in watts we use Equation 8-3:

$$-9 = 10 \times \log(P_{out}/1 \text{ mW})$$

$$P_{out} = (1 \text{ mW})(10^{-0.9})$$

Solving, we get

$$P_{out} = 126 \mu\text{W}$$

An important characteristic of transmissive star couplers is *cross talk* or the amount of input information coupled into another input. Cross coupling is given in decibels and is typically greater than 40 dB.

The **reflective star coupler** has the same power division as the transmissive type, but cross talk is not an issue because power from any fiber is distributed to all others.

T-couplers

In Figure 8-37, power is launched into port 1 and is split between ports 2 and 3. The power split does not have to be equal. The power division is given in decibels or in percent. For example, and 80/20 split means 80% to port 2, 20% to port 3. In decibels, this corresponds to 0.97 dB for port 2 and 6.9 dB for port 3.



Figure 8-37 *T-coupler*

$$10 \log (P_2/P_1) = -0.97 \text{ dB}$$

$$10 \log (P_3/P_1) = -6.96 \text{ dB}$$

Directivity describes the transmission between the ports. For example, if $P_3/P_1 = 0.5$, P_3/P_2 does not necessarily equal 0.5. For a highly directive T-coupler, P_3/P_2 is very small. Typically, no power is expected to be transferred between any two ports on the same side of the coupler.

Another type of T-coupler uses a graded-index (GRIN) lens and a partially reflective surface to accomplish the coupling. The power division is a function of the reflecting mirror. This coupler is often used to monitor optical power in a fiber optic line.

E. Wavelength-Division Multiplexers

The couplers used for wavelength-division multiplexing (WDM) are designed specifically to make the coupling between ports a function of wavelength. The purpose of these couplers is to separate (or combine) signals transmitted at different wavelengths. Essentially, the transmitting coupler is a mixer and the receiving coupler is a wavelength filter. Wavelength-division multiplexers use several methods to separate different wavelengths depending on the spacing between the wavelengths. Separation of 1310 nm and 1550 nm is a simple operation and can be achieved with WDMs using bulk optical diffraction gratings. Wavelengths in the 1550-nm range that are spaced at greater than 1 to 2 nm can be resolved using WDMs that incorporate interference filters. An example of an 8-channel WDM using interference filters is given in Figure 8-38. Fiber Bragg gratings are typically used to separate very closely spaced wavelengths in a DWDM system (< 0.8 nm).



(Courtesy of DiCon, Inc.)

Figure 8-38 8-channel WDM

Erbium-doped fiber amplifiers (EDFA)—The EDFA is an optical amplifier used to boost the signal level in the 1530-nm to 1570-nm region of the spectrum. When it is pumped by an external laser source of either 980 nm or 1480 nm, signal gain can be as high as 30 dB (1000 times). Because EDFAs allow signals to be regenerated without having to be converted back to electrical signals, systems are faster and more reliable. When used in conjunction with wavelength-division multiplexing, fiber optic systems can transmit enormous amounts of information over long distances with very high reliability.

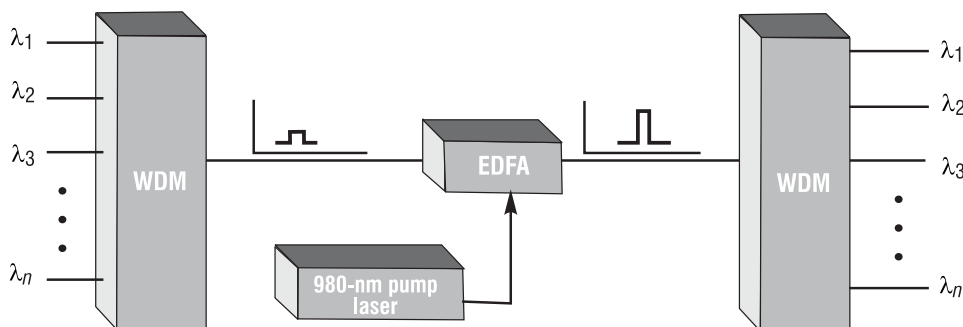


Figure 8-39 Wavelength-division multiplexing system using EDFAs

Fiber Bragg gratings—Fiber Bragg gratings are devices that are used for separating wavelengths through diffraction, similar to a diffraction grating (see Figure 8-40). They are of critical importance in DWDM systems in which multiple closely spaced wavelengths require

separation. Light entering the fiber Bragg grating is diffracted by the induced period variations in the index of refraction. By spacing the periodic variations at multiples of the half-wavelength of the desired signal, each variation reflects light with a 360° phase shift causing a constructive interference of a very specific wavelength while allowing others to pass. Fiber Bragg gratings

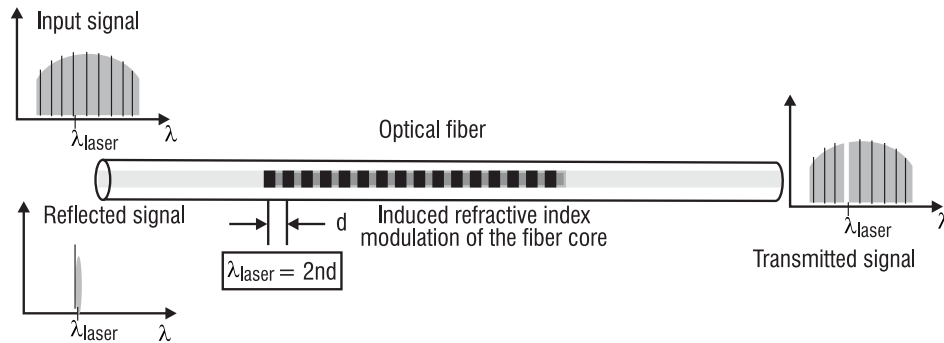
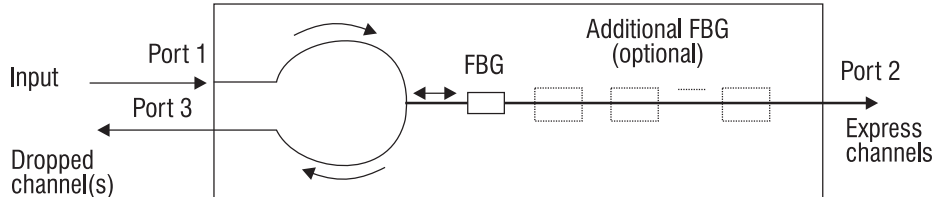


Figure 8-40 Fiber Bragg grating

are available with bandwidths ranging from 0.05 nm to >20 nm. Fiber Bragg gratings are typically used in conjunction with *circulators*, which are used to drop single or multiple narrow-band WDM channels and to pass other “express” channels (see Figure 8-41). Fiber Bragg gratings have emerged as a major factor, along with EDFAs, in increasing the capacity of next-generation high-bandwidth fiber optic systems.



Courtesy of JDS-Uniphase

Figure 8-41 Fiber optic circulator

Figure 8-42 depicts a typical scenario in which DWDM and EDFA technology is used to transmit a number of different channels of high-bandwidth information over a single fiber. As shown, n -individual wavelengths of light operating in accordance with the ITU grid are multiplexed together using a multichannel coupler/splitter or wavelength-division multiplexer. An *optical isolator* is used with each optical source to minimize troublesome back reflections. A tap coupler then removes 3% of the transmitted signal for wavelength and power monitoring. Upon traveling through a substantial length of fiber (50-100 Km), an EDFA is used to boost the signal strength. After a couple of stages of amplifications, an add/drop channel consisting of a fiber Bragg grating and circulator is introduced to extract and then reinject the signal operating at the λ_3 wavelength. After another stage of amplification via EDFA, a *broadband* WDM is used to combine a 1310-nm signal with the 1550-nm window signals. At the receiver end, another broadband WDM extracts the 1310-nm signal, leaving the 1550-nm window signals. The 1550-nm window signals are finally separated using a DWDM that employs an array of

fiber Bragg gratings, each tuned to the specific transmission wavelength. This system represents the current state of the art in high-bandwidth fiber optic data transmission.

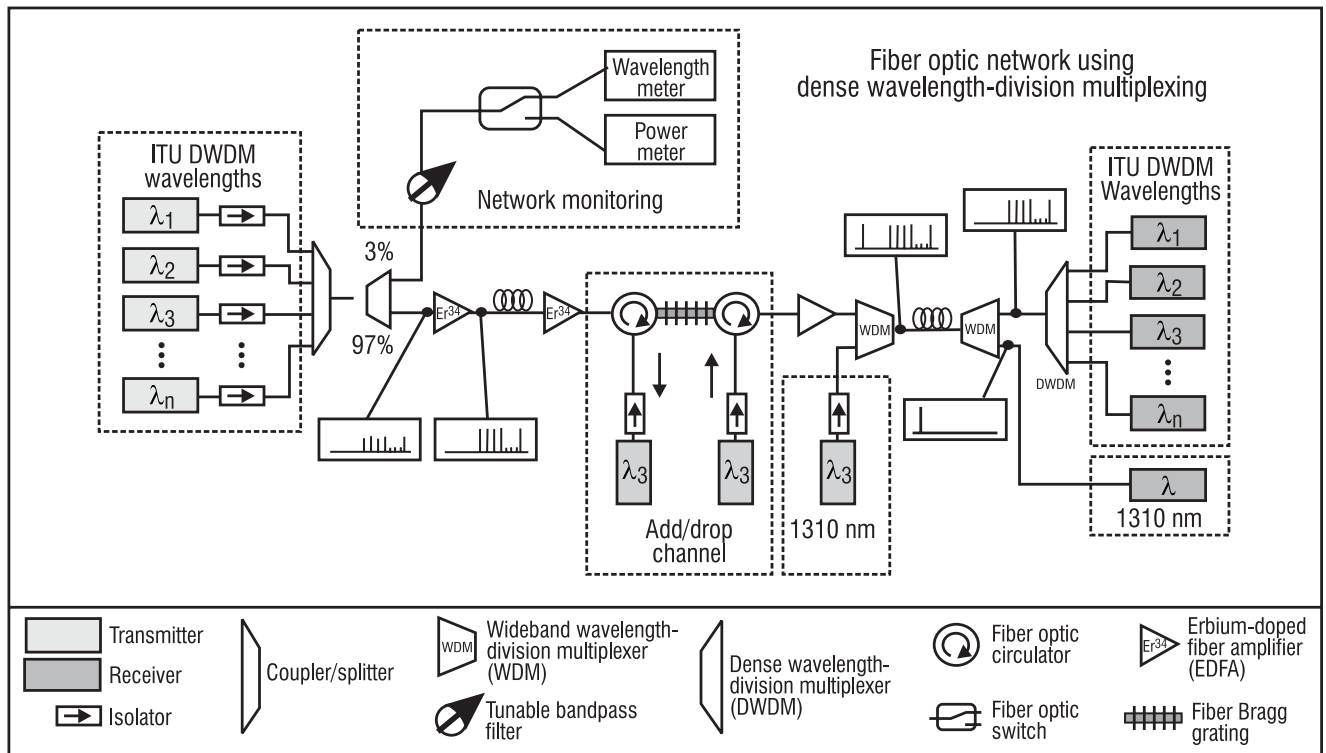


Figure 8-42 Typical DWDM transmission system (Courtesy of Newport Corporation)

What's ahead?

Over the past five years, major breakthroughs in technology have been the impetus for tremendous growth experienced by the fiber optic industry. The development of EDFAs, fiber Bragg gratings and DWDM, as well as advances in optical sources and detectors that operate in the 1550-nm range, have all contributed to advancing the fiber optics industry to one of the fastest growing and most important industries in telecommunication today. As the industry continues to grow, frustrating bottlenecks in the “information superhighway” will lessen, which will in turn usher in the next generation of services, such as telemedicine, Internet telephony, distance education, e-commerce, and high-speed data and video. More recent advances in EDFAs that operate at 1310-nm and 1590-nm technology will allow further enhancement in fiber optic systems. The future is bright. Just remember, the information superhighway is paved with glass!

Problem Exercises/Questions

1. A fiber of 1-km length has $P_{\text{in}} = 1 \text{ mW}$ and $P_{\text{out}} = 0.125 \text{ mW}$. Find the loss in dB/km.
2. A communication system uses 8 km of fiber that has a 0.8-dB/km loss characteristic. Find the output power if the input power is 20 mW.
3. A 5-km fiber optic system has an input power of 1 mW and a loss characteristic of 1.5 dB/km. Determine the output power.
4. What is the maximum core diameter for a fiber to operate in single mode at a wavelength of 1310 nm if the N.A. is 0.12?
5. A 1-km-length multimode fiber has a modal dispersion of 0.50 ns/km and a chromatic dispersion of 50 ps/km·nm. If it is used with an LED with a linewidth of 30 nm, (a) what is the total dispersion? (b) Calculate the bandwidth (BW) of the fiber.
6. A digital MUX operates with 16 sources. The rate of data in each source is 8000 bytes/second (assume 8 bits per byte). Data are transmitted byte by byte.
 - (a) What is the data rate of the MUX output?
 - (b) What is the channel switching rate?
7. A receiver has a sensitivity P_s of -40 dBm for a BER of 10^{-9} . What is the minimum power (in watts) that must be incident on the detector?
8. A system has the following characteristics:
 - LED power (P_L) = 1 mW (0 dBm)
 - LED to fiber loss (L_{sf}) = 3 dB
 - Fiber loss per km (F_L) = 0.2 dB/km
 - Fiber length (L) = 100 km
 - Connector loss (L_{conn}) = 3 dB (3 connectors spaced 25 km apart with 1 dB of loss each)
 - Fiber to detector loss (L_{fd}) = 1 dB
 - Receiver sensitivity (P_s) = -40 dBm

Find the loss margin and sketch the power budget curve.
9. A 5-km fiber with a $\text{BW} \times \text{length}$ product of $1200 \text{ MHz} \times \text{km}$ (optical bandwidth) is used in a communication system. The rise times of the other components are $t_{\text{tc}} = 5 \text{ ns}$, $t_L = 1 \text{ ns}$, $t_{\text{ph}} = 1.5 \text{ ns}$, and $t_{\text{rc}} = 5 \text{ ns}$. Calculate the electrical BW for the system.
10. A 4×4 star coupler is used in a fiber optic system to connect the signal from one computer to four terminals. If the power at an input fiber to the star coupler is 1 mW, find (a) the power at each output fiber and (b) the power division in decibels.
11. An 8×8 star coupler is used to distribute the $+3\text{-dBm}$ power of a laser diode to 8 fibers. The excess loss (Loss_{ex}) of the coupler is 1 dB. Find the power at each output fiber in dBm and μW .

Laboratory: Making a Fiber Optic Coupler

In this lab you will fabricate a 2×2 fiber optic coupler using 1-mm-diameter plastic fiber. The coupler can be used for a variety of applications including wavelength-division multiplexing and power splitting, which will be outlined in this lab.

Equipment List

The following equipment is needed to complete this laboratory.

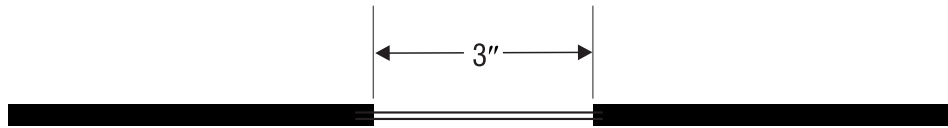
- 2 1-foot sections of 1-mm-diameter plastic-jacketed fiber (Part #2705FIBOPT)¹
- 1 razor blade
- 1 heat gun
- 1 4" piece of heat-shrink tubing
- 2 high-brightness LEDs (1 green and 1 red)
- 2 plastic fiber connectors (Part #2400228087-1)¹
- 2 plastic fiber LED mounts (Part #2400228040-1)¹
- 4 multimode ST-connectors for 1-mm fiber (Part #F1-0065)²
- 1 electronic breadboard with +5-volt supply
- 1 850-nm fiber optic source with ST adapter (Part #9050-0000)²
- 1 850-nm fiber optic detector with ST adapter (Part #F1-8513HH)²
- 1 low-cost diffraction grating (Part #J01-307)³
- 1 1-meter patch cord (terminated with ST connectors)
- 1 fiber optic termination kit (includes scissors, alcohol wipes, crimp tool, fiber-inspection microscope, razor blades, etc.)¹

(Notations 1, 2, 3: See sources in APPENDIX.)

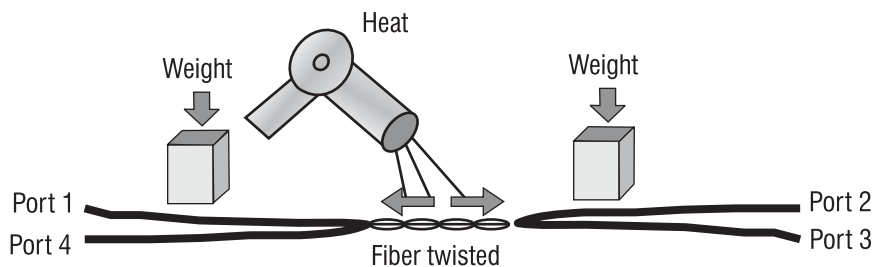
Procedure

PART I: Making a Fiber Optic Coupler

1. With the razor blade, carefully strip off approximately 3" of the fiber jacket in the middle of the fiber (see Figure 8-43).

**Figure 8-43**

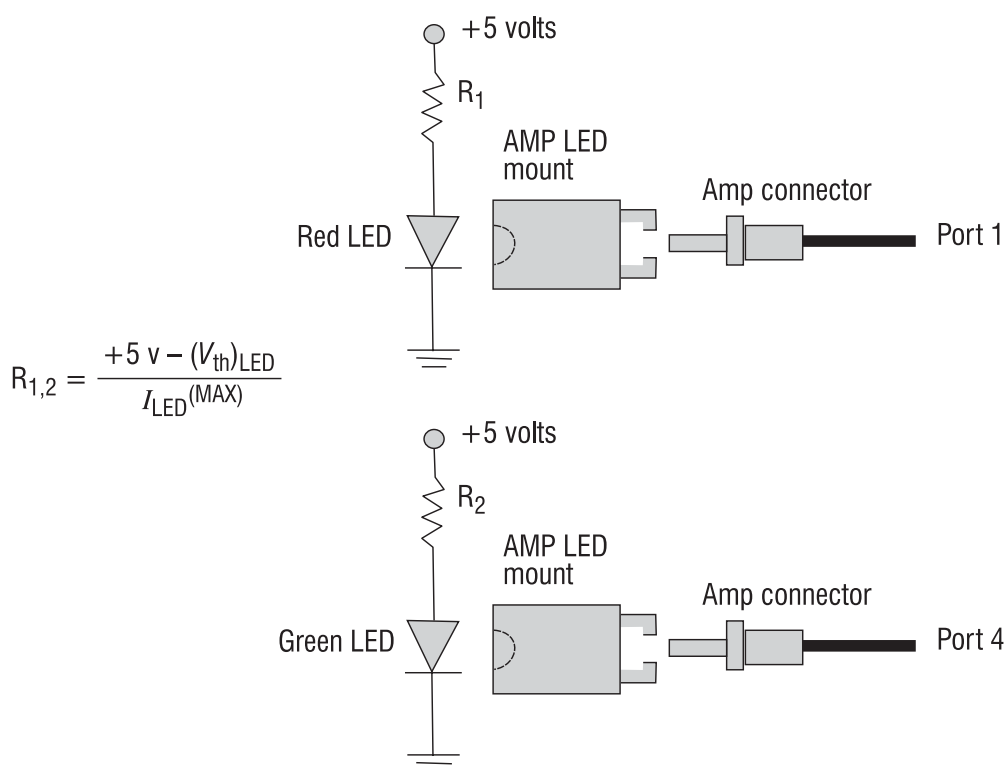
2. Where the fiber has been stripped, twist the two fibers together.
3. On each end of the stripped area, place a small weight (i.e., paperweight, book) to hold the fiber in place (see Figure 8-44).

**Figure 8-44**

4. Using the heat gun on the low setting, apply heat to the twisted area. Move the heat gun gently back and forth to uniformly melt the fiber. **CAUTION:** Do not hold the heat gun stationary because the fiber will melt quickly!
5. As the fiber is heated, you will notice that it will contract a bit. This is normal. When the contraction subsides, remove the heat gun and let the fiber cool for a minute.
6. With a laser pointer or fiber optic source, shine light into port 1 of the coupler. You should observe a fair amount of coupling (~20–30%) into port 3 of the coupler. If more coupling is needed, repeat the heating process until the desired coupling is obtained.

PART II: Wavelength-Division Multiplexing Demonstration

1. Apply the AMP plastic fiber connectors to the two input fibers (ports 1 and 4) according to manufacturer's specifications. Polish the ends if necessary. Also polish the ends of the unterminated fibers if necessary.
2. On the electronic breadboard, set up the circuit shown in Figure 8-45. Depending on the type of LED, you may have to use epoxy to secure the LED in the mount.

**Figure 8-45**

3. When the circuit is complete, connect the fibers to the LEDs and observe the output of port 2. The red and green colors will be mixed.
4. To separate the colors, observe the output of port 2 through the diffraction grating. You should observe a central bright spot (coming from the fiber) and two identical diffraction patterns—one on either side—with the red and the green separated (see Figure 8-46). To ensure that the two signals are indeed independent, turn off the LEDs one at a time and observe the output of port 2 through the diffraction grating.

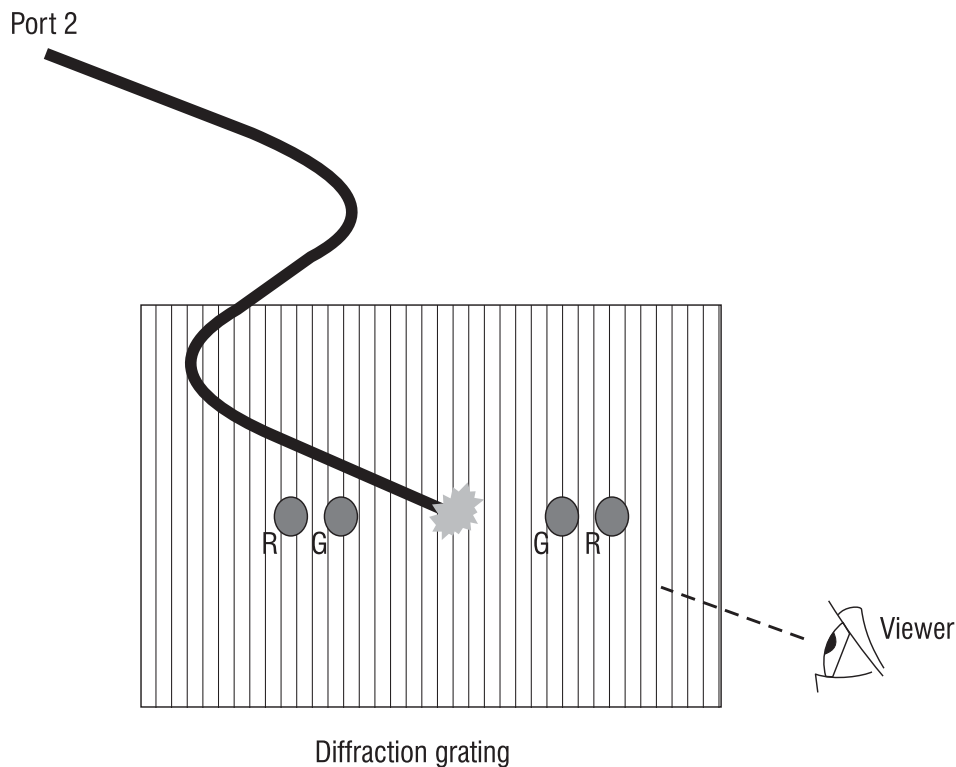


Figure 8-46

Part III: Measuring Coupler Loss

1. Repeat steps 1–6 (Part I) for fabrication of a 2×2 coupler.
2. “Connectorize” each port of the coupler using ST-multimode connectors and polish if necessary. (Instructions for termination are supplied with the connectors when purchased.)
3. Measure the output of your fiber optic source at the output of the patch cord. This will be the input power to the coupler. Record the power in Table 8.6.
4. Measure the output power at each of the ports and record in Table 8.6.
5. Calculate the throughput loss using the following equation:

$$L_{\text{th}} = -10 \log (P_2/P_1)$$

6. Calculate the tap loss using the following equation:

$$L_{\text{tap}} = -10 \log (P_3/P_1)$$

7. Calculate the directionality loss using the following equation:

$$L_{\text{dir}} = -10 \log (P_4/P_1)$$

8. Calculate the excess loss using the following equation:

$$L_{\text{ex}} = -10 \log (P_2 + P_3)/P_1$$

9. Repeat Part III – steps 3–9 using each port as the input. Record the results in Table 8.6.

Table 8.6

Input Port	Input Power (mW)	Throughput Loss (dB)	Directionality Loss (dB)	Tap Loss (dB)	Excess Loss (dB)
1					
2					
3					
4					

APPENDIX

- Items may be obtained through
Electronix Express
365 Blair Road
Avenel, NJ 07001
1-800-972-2225
- Items may be obtained through
Fiber Instrument Sales (FIS)
161 Clear Road
Oriskany, NY 13424
1-800-500-0347
- Items may be obtained through
Edmund Scientific, Inc.
101 East Gloucester Pike
Barrington, NJ 08007
856-573-6250

Photonic Devices for Imaging, Display, and Storage

Harley R. Myler
University of Central Florida
Orlando, Florida

Electronic and electro-optic devices are frequently used to display images obtained from the computer processing of data. Images, or *digital pictures*, are generally two-dimensional data structures that convey spatial information to the viewer. Images are collected through various means, from digital cameras to laser radar scanning systems and, once stored in a computer, can be manipulated mathematically to accomplish many different objectives. The improvement of images for viewing or analysis and computer interpretation of image content are among those objectives. This module explains the terminology associated with images, how images are acquired and stored, and how images are displayed.

Prerequisites

Before starting this module, you should have completed Modules 1-1, *Nature and Properties of Light*; 1-2, *Light Sources and Safety*; and 1-3, *Basic Geometrical Optics*. Specifically, you should have knowledge of fundamental optics to include lenses, apertures, and image formation. You will also require a basic knowledge of electronic device physics.

Objectives

When you complete this module you will be able to:

- Define imaging, pixel, quantization, sampling, and bandwidth.
- Explain the relationship between resolution and spatial frequency.
- Calculate the resolution of an imaging device.
- Describe the basic parts of a camera and explain how it is used to record images.
- List the different types of scanners, e.g., flying spot, flatbed, and drum, and explain their operation.
- Understand how computer files store images for archival purposes.
- Explain the difference between lossless and lossy image-compression schemes.
- Describe the structure, operation and capabilities of a CCD camera, a vidicon, and an image intensifier.
- Describe the structure, operation, and capabilities of a CRT display.
- List and define the phases of liquid crystal materials.
- Describe the parts and operation of a basic LCD.
- Explain the difference between active and passive LCD technologies.
- List and define the two addressing modes for passive LCDs.
- Define electroluminescence.
- Describe the operation of an electroluminescent display and explain how it differs from an LCD.
- Describe the operation of an LED display and explain how it differs from an LCD.

Scenario—recording and displaying images in the workplace

A company that develops and manufactures optical products using state-of-the-art technologies for the medical, display, instrument, and other industries has hired Jennifer Watley to assist optical product development engineers in the prototyping laboratory. During her interview for the position, Jennifer was told that the successful candidate needed a background in display technologies and image acquisition, which were areas of study that she had particularly enjoyed at school. Shortly after accepting the photonics technician position, Jennifer discovered that all of the fundamentals of displays and image storage that she learned are useful in her work. Her daily activities involve working closely with engineers and technicians to determine what data must be recorded from their experiments and the best ways to display and store the images produced. Jennifer has discovered that high tech companies like the one for which she works have sophisticated equipment and software that allow very-high-resolution images to be recorded, manipulated, and stored. Her background training has allowed her to quickly “come up to speed” on the latest technology and to understand the complex details of operating and maintaining imaging devices. Jennifer has just been told that her company is sending her to a special training course on a new high-resolution thermal imaging system that they will be purchasing for her lab. She finds the continual learning opportunities on high-tech state-of-the-art equipment one of the aspects of her job that she especially enjoys. Jennifer knows that her training and motivation are the reason her company continues to invest in her education.

Opening Demonstration

Using an optics bench and a simple lens system, image an object, such as a golf ball on tee, doll, or toy car, onto the CCD array of an inexpensive circuit board “pinhole” camera. You will have to remove the lens of the camera to expose the CCD “chip,” which should be about $\frac{1}{4}$ " square and located just behind the lens. Display the output of the camera on an LCD screen. Compare the object being imaged to the image displayed. Are any of the object’s details not displayed? For example, if your object is a doll’s head, does the doll have fine eyelashes that cannot be distinguished in the displayed picture? The ability to acquire and display fine detail is dependent on the *resolution* of the imaging system. The CCD camera documentation should state the resolution of the camera in lines. We will define resolution in terms of lines per length. If the CCD has 380 lines (very common) and is $\frac{1}{4}$ " square, the resolution will be 1520 lines per inch. If the image of the object has lines that are finer than that, they will not be observed at the output, provided that the optics can image the object to the size of the CCD array. The display also has a resolution associated with it, and, if the display resolution is less than that of the camera, the display will not be able to accurately reproduce what the camera acquires.

Basic Concepts

I. INTRODUCTORY CONCEPTS

A major outcome of photonics work is in the form of images, where an image is a two-dimensional structure—an array—that can represent many different things depending on how the image was made or acquired and processed. When an image is made available as the output of a photonic system, it must be displayed for observation and further study. The imaging display allows us to view images that have been produced as an outcome of a photonics investigation or process. Images can be static, such as photographs as shown in Figure 9-1, or dynamic, as in movies.



Figure 9-1 *Picture*

A movie is nothing more than a sequence of static images called *frames* that are displayed at a speed at which the human brain fuses them into a continuous sequence. The *flicker fusion rate*, as perceptual psychologists call it, is around 24 frames per second. In this module we are interested in the devices that are used to acquire and display both static and dynamic images. To discuss those devices, it is first necessary to define and explore some fundamental aspects common to all images produced by photonic systems. These elements of images have to do with sampling theory, which we discuss next.

A. Sampling Theory

Sampling theory is involved with the collection, analysis, and interpretation of data. The data in which we are interested are image data collected from photonic systems, and this includes images formed by scanning and staring devices. A *scanning device* has one sensor or a small array of sensors that *it moves* in order to collect an array of data. A *staring sensor* has as many sensor elements as data in the array that it records and so does not move. To discuss the concepts of sampling theory that are pertinent to imaging, we first have to discuss the

fundamental elements of an image and their relationship to each other and to the image overall. These elements are called *pixels*.

Pixels

A *pixel* is a conjunction of the words *picture element*, in which the term picture is synonymous with image. Imagine a set of white marbles set into holes on a wooden board. The holes have been drilled to form a square array, and so a view of the marbles would look something like Figure 9-2.

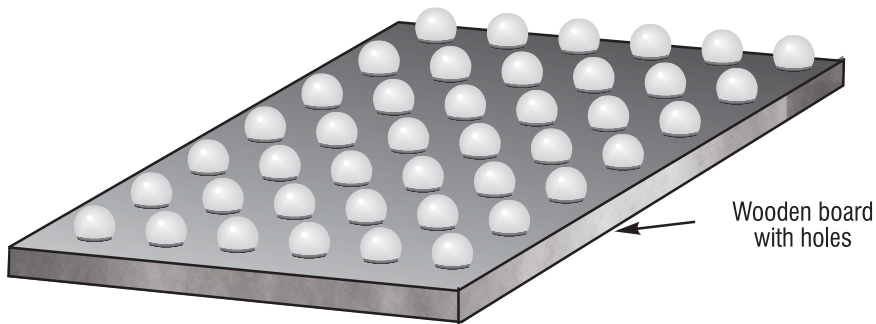


Figure 9-2 “Marble” array

Now imagine that we have 256 marbles and they are arranged as 16 rows by 16 columns. If we replace some of the white marbles with black ones, we can produce an image using the marbles, like that shown in Figure 9-3.

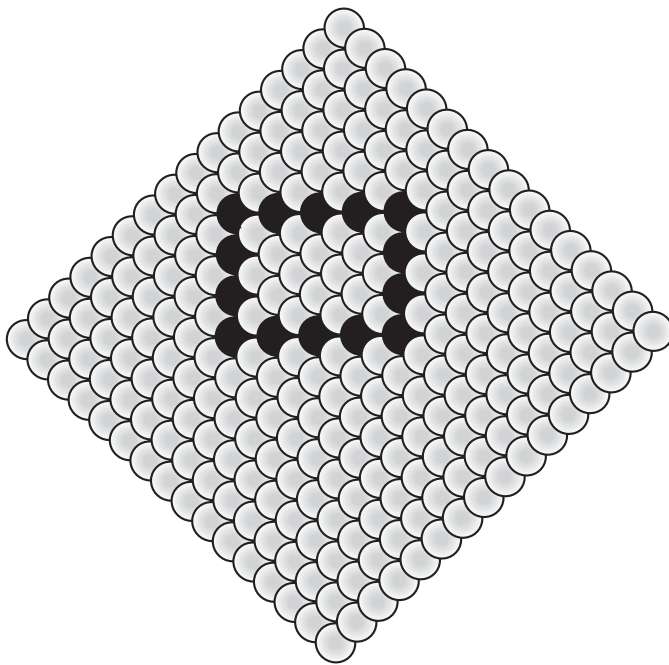


Figure 9-3 “Marble” array image

The pixels in the image are exemplified by the marbles. The image produced by the marbles is called a *binary* image, since each of the pixels (marbles) can be one of two values (black or white). Imagine that we have an array of 400 by 400 marbles and the marbles have a range of shades between white and black, a scale of gray levels. An image produced by a marble array of this sort would look something like Figure 9-4.

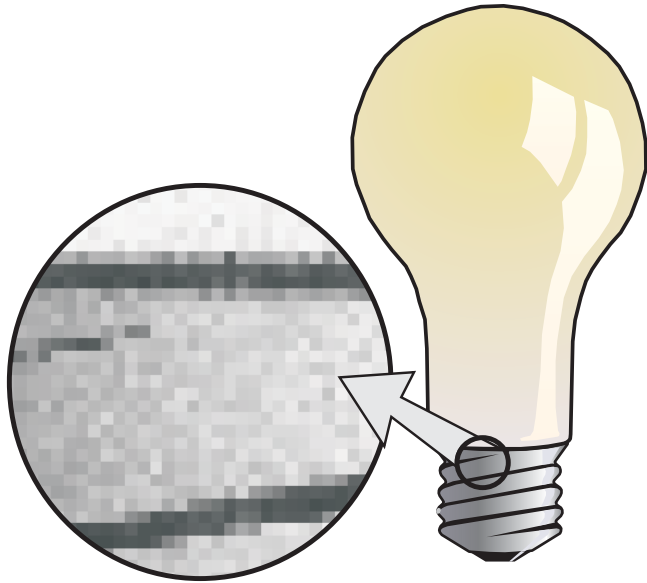


Figure 9-4 High resolution “marble” array image

What do we notice about this image? It is clearer and sharper than our 256-marble image, because it has a higher *resolution*. If we dispense with the marbles altogether and replace them with points of light on a rectangular screen, we have something close to the output of a black and white television. The range of grays from black to white is called the *grayscale*, and a black and white image of this sort is sometimes called a grayscale image. If the pixels can take on color values, we have a color image. Pixels are not restricted to visible light, but can be variations of ink on a printed page. Pixels can also represent signals that cannot be viewed directly by the human eye, such as the pixels in an infrared or laser radar image.

The values that a pixel can represent have to do with the *quantization* of the pixel, expressed in terms of *bits*. The images that we have been discussing are *digital images*, or images that take on discrete values. Each pixel is a discrete component of the image, and a fundamental assumption is that the image will be stored, manipulated, and displayed by a computer. As we saw with the digital binary image, the pixels may be only one of two (discrete) values. In a grayscale image, the pixel takes on a set of values that are typically defined by a power of 2, such as 4, 8, 16, 32, 64, 128, and 256. This is because each pixel is represented by a binary number in the computer. The more bits that represent the pixel, the more grayscale values it can take on.

The two most common image quantizations are 8-bit grayscale images and what are called 24-bit *truecolor* images. *To compute the quantization of a pixel, we simply raise 2 by the number of bits.* So, the 8-bit image will have $2^8 = 256$ gray levels. The 24-bit truecolor image is a bit (no pun intended) different. Here each pixel is actually three pixels, one each of red, green,

and blue. Color images are more complex than black and white images because they require a combination of the three primary colors, red, green, and blue. The three pixels of a color digital image are so small and closely spaced that their combination is fused by the eye into a single pixel whose color is the additive combination of three individual pixels. You can see this by looking at a color computer monitor with a magnifying glass. Each pixel is just like a grayscale pixel except that, instead of varying from black to white through shades of gray, the pixel varies from black to either red, green, or blue, with varying intensities of each color. The 24-bit truecolor image has 256 reds, 256 blues, and 256 greens. Collectively, they can produce $256 \times 256 \times 256$ colors, or 16,777,216 million colors, that is, 2^{24} colors. People can distinguish between only about 64 different shades of gray, but we do better with colors, where the number jumps into the millions.

We can conclude that the pixel is simply the fundamental *unit* of an image, like a brick is the unit of a wall or a grain is the unit of a sandpile. The pixel can take on various values that constitute the quantization of the image, like a brick or grain of sand can have a color or a shade of gray. We will now discuss how the image is defined in terms of resolution and spatial frequency.

Resolution and Spatial Frequency

Resolution has to do with the fineness of detail the image can represent, or the fineness of detail the camera can record or the display system display. The more pixels per unit area an image has, the higher the resolution. The term resolution comes from the word *resolve*, which the dictionary defines as “to break up into constituent parts.” There are a number of ways to define resolution in terms of imaging. One way is by using the following equation:

$$\text{Resolution} = \text{number of pixels/area.}$$

For an image that is 3×3 inches square and contains 900 by 900 pixels, the resolution is $(900 \times 900)/(3 \times 3) = 9000$ pixels per square inch. This sort of measure is difficult to visualize and so an alternative expression that is more common defines resolution in terms of *lines per distance*. When you purchase a CCD camera or a television, the resolution is stated in terms of *lines*. Many inexpensive CCD cameras have 380 lines; this means that the sensor array has 380 rows of sensors. If the CCD array—the electronic chip—is $\frac{1}{4}$ " wide, we can calculate:

$$\text{Resolution} = \text{number of lines/length} = 380 \text{ lines}/0.25 \text{ inch} = 1520 \text{ lines per inch.}$$

If we were to draw a one-inch square on a piece of paper and then draw lines on the square, we could draw lines as thin as $\frac{1}{1520}$ of an inch and the CCD could resolve them. This assumes, however, that a $\frac{1}{4}$ " subsquare is imaged onto the CCD array due to the size of the array. Also, we could resolve only 760 lines, which is one-half of 1520. The reason for this is that we need to have a black line then a white line, then a black, and so on. To display a line, we need a sensor for the line and a sensor for the space between that line and the next line. Hence, we need two sensors per line, which gives rise to the division by two of the resolution. In sampling theory, the *Nyquist rate* is the rate at which you must sample a signal in order to capture all of its frequencies, and this rate is twice the highest frequency that must be sampled! Frequency has to do with how often something changes in time. When something changes in distance, like pixels or lines per inch, we call this *spatial frequency*. Spatial frequency is a more specific term

than resolution, even though they share the same units. Typically, images have resolutions that are represented as pairs of numbers, indicating the number of rows and columns in pixels of the image—although the actual size in terms of length or area of the image is not mentioned. Common computer display resolutions are 512 by 512, 1024 by 1024, and 2048 by 2048 for square images and 640 by 480, 800 by 600, and 1024 by 768 for rectangular images. High definition television, or HDTV, is 1080 by 1920 pixels.

Figure 9-5 shows two grayscale images of the same object, both the same size. Figure 9-5a, however, is of far lower resolution than Figure 9-5b. The pixels of Figure 5a are much larger than those of Figure 5b, so there are fewer of them per unit area—hence a lower resolution. It is possible to see the square shape of the pixels in Figure 5a, whereas an individual pixel in Figure 5b cannot be distinguished without the aid of a magnifying glass.

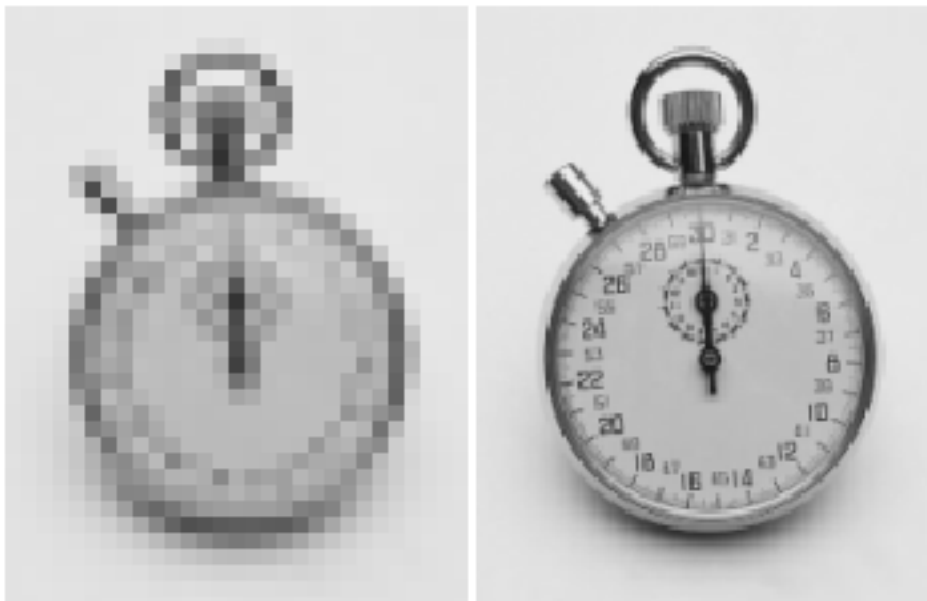


Figure 9-5 (a) low resolution (b) high resolution

The resolution of printed media is discussed in terms of *dots per inch* (DPI), where a dot is a spot of ink, like a black pixel. Resolution in office laser printers is typically 300 or 600 DPI. Printers of magazines and books use presses that can produce upward of 1200 DPI. The common newspaper is 72 DPI. It is possible to print images of grayscale or color using binary pixels. The two most common processes are *dithering* and *half-toning*, where patterns of ink dots are used to develop the illusion of a grayscale pixel. The next time you read a newspaper, examine the pictures with a magnifying lens or glass. The half-tone pixels are easily discernible.

It should be clear that, the more resolution an image has, the finer the lines that can be displayed or printed. In Figure 9-5b, it is very easy to distinguish the seconds tick-marks of the stopwatch dial, whereas in Figure 9-5a they cannot be identified. This has to do with the *spatial frequency*, which we defined earlier as a measure of lines per unit distance. We say *spatial frequency* because the frequency is two-dimensional. Consider Figure 9-6. Here we see eight rectangles that exhibit varying degrees and types of spatial frequency. Starting in the upper left frame and reading across, we have a frame with closely spaced vertical lines, followed by the same size

frame with fewer verticals. The first frame has a *higher spatial frequency* content than the second. The last two frames of the first row have high and low spatial frequency, respectively, but with horizontal lines. The second row at left shows a grid pattern where the vertical spatial frequency is equal to the horizontal. The frame next to it shows a square with very small dots. That square has high spatial frequency horizontally, vertically, and diagonally. The last two images show random textures. Can you tell which one has higher spatial frequency? If you said the first one, you were correct, because there is greater detail in the pattern. The spatial frequency is related directly to the resolution. In fact, the resolution of which a display is capable determines the maximum spatial frequency that can be displayed.

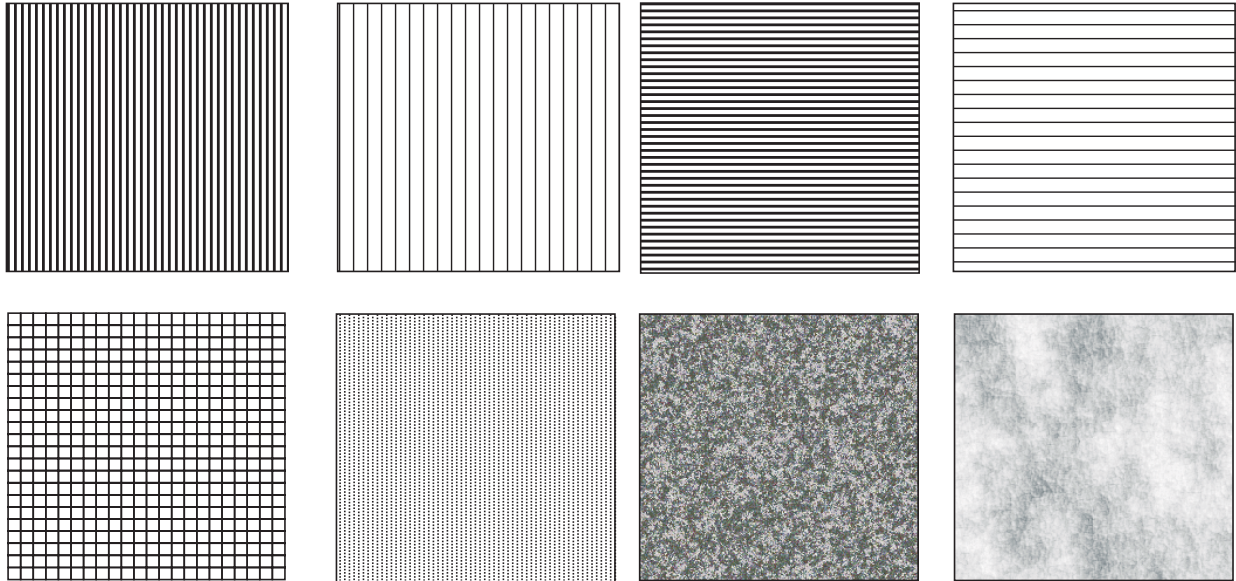


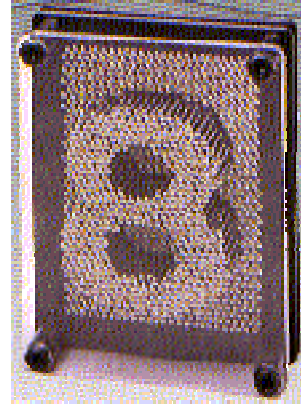
Figure 9-6 Examples of different spatial frequencies

As we saw earlier, if we have a resolution of 300 DPI, we can display a spatial frequency no greater than 150 lines per inch, or one-half the resolution. We can consider resolution to be a measure of sampling capability, where each pixel is a sample. The maximum number of lines that can be represented, i.e. the maximum spatial frequency, is measured in lines per unit distance and is just one-half the resolution. This can be expressed analytically as:

$$\text{Maximum spatial frequency (lines/distance)} = 1/2 \text{ resolution (lines/distance).}$$

LAB 1: Resolution

Using a “pin art” module (see Other Resources on where to obtain) produce reliefs of different objects (calculator, golf ball, stapler, etc.) like that shown in the illustration. Now imagine that the movable pins in the module are pixels. Instead of values of light intensity, the pin depth determines the pixel values of the pins. Use a ruler to determine the size of the pin area, then count the number of pins and calculate the resolution of the module. Find an object with physical features that are smaller than the resolution of the pin art. Can the module reproduce the object features?



Bandwidth

Bandwidth is defined as the amount of information that can be transmitted across a channel in a fixed amount of time. See Module 1-8, *Fiber Optic Telecommunication*. Bandwidth is a term from communication theory, but here we can consider an image to be a channel and can discuss *image bandwidth* in terms of spatial frequency. We replace time with distance and say that the bandwidth of an image is the maximum spatial frequency that the image can contain. This was computed earlier and can now be rewritten as the image bandwidth expression:

$$\text{Image bandwidth (lines/distance)} = \frac{1}{2} \text{ resolution (lines/distance)}.$$

Certainly, an image where all the pixels are of the same value does not convey much information. However, a very complex scene with a great amount of detail can reveal a good deal of information. Like we saw with the stopwatch in Figure 9-5, the image in 9-5b with the higher resolution and higher spatial frequency is the image with the greater bandwidth.

When images are transmitted over communication channels, the bandwidth of the channel will have a substantial effect on the speed at which images can be passed. Images are, by nature, high-information-content data structures. To illustrate this, consider the fact that a single frame of an HDTV image is roughly 6.2 billion pixels and to generate a 30-frame-per-second image sequence we need a channel capable of processing 186 billion pixels per second. With HDTV, it is necessary to reduce the amount of data (not information!) using compression techniques so that a channel of lesser bandwidth may be used. Nevertheless, a channel bandwidth of 8 MHz is needed. We can contrast this to analog TV, which requires only 4 MHz.

B. Imaging systems

Imaging systems include devices and methods used to capture and store images. These systems can capture images live from natural scenes or copy images that have been stored to a medium such as film or paper. *Imaging* refers to the methodologies associated with *capturing*, *storing*, and *computer-processing* images. Some clarification in terminology is now required. Images formed optically are referred to simply as images, and images that have been captured onto film are called *pictures*. These two terms are used interchangeably when the images are captured or

scanned electronically for input, storage, and processing by a computer. The formal terms are *digital images* and *digital pictures* to indicate that we are talking about computer imaging. However, we will remain with convention and allow the context of use to define the type of images being discussed. When the potential for confusion arises, we will specify computer image or picture to clarify.

Cameras

A camera is a photonic device that uses a lens system to form the image of a natural scene onto a recording medium. The first cameras employed various chemical compounds to record scenes onto paper, polymer sheets, or other inert substrates—called the film—such as glass or plastic. The basic parts of a camera are illustrated in Figure 9-7. The film responds to light through chemical change. Light-sensitive compounds change in varying degrees depending on the intensity and wavelength of the light. Other chemicals are used to develop the film and reveal the image as it was recorded to the film.

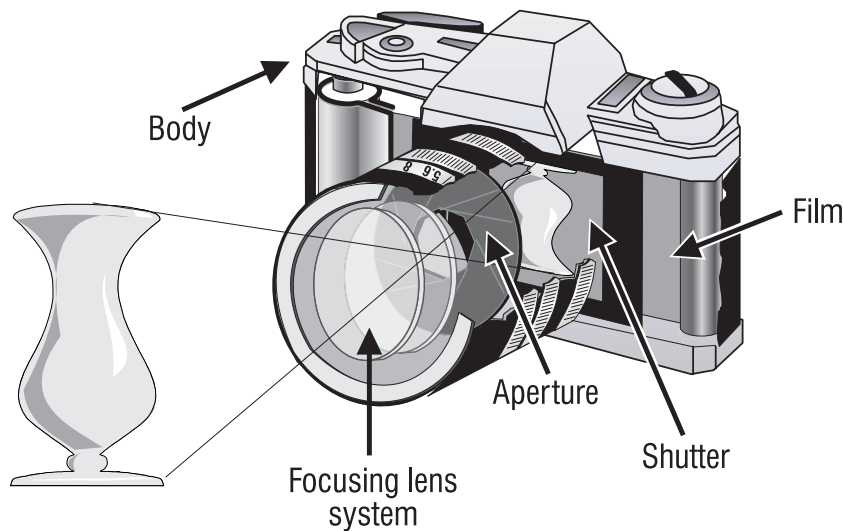


Figure 9-7 *Parts of a camera*

Light enters the camera through the lens system and is imaged onto the medium. The medium responds to the light proportionally, so some means must be provided to control the length of time that the image is allowed to illuminate the film. This time is controlled by the *shutter*, which is nothing more than a mechanical window that stays closed until it is time to expose the film. When that time arrives, the shutter opens for a preset amount of time, the *exposure time*. Exposure times vary depending on the lighting conditions, the camera optics, and the film. There are three critical aspects to camera photography: *f-stop*, *shutter speed*, and *film speed*.

The *f-stop* is the ratio of the focal length of the lens to the diameter of the lens opening, or aperture. The size of the aperture is controlled by turning a collar typically located at the base of the lens. Typical values may be *f* 2, *f* 2.8, *f* 5.6, and *f* 16 where the “*f*” denotes the *f-stop*, sometimes referred to as *f-numbers*. Smaller *f-numbers*—*f* 2, *f* 2.8, etc.—represent larger apertures, and higher *f-numbers*—*f* 11, *f* 16, *f* 22, etc.—give smaller apertures. The *f-stops* are

arranged so that each f-number allows exactly half as much light through as the previous one and twice as much light through as the next. For example:

f-number	1.4	2	2.8	4	5.6	8	11	16
relative brightness	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$

Shutter speed is a measure of how long the shutter remains open when the picture is taken. On manual cameras, it is usually set by means of a dial on the top of the camera or, less commonly, a ring around the base of the lens. Typical shutter speeds are $\frac{1}{1000}$, $\frac{1}{500}$, $\frac{1}{250}$, $\frac{1}{15}$ and 1 second. Note that, like f-stop, shutter speeds differ by factors of 2. This makes it easier for photographers to judge camera settings for the subjects or scenes they are trying to photograph. It is for this reason that much of photography is considered to be an art.

Film speed is a measure of how sensitive the film is to light and is measured by an International Standards Organization (ISO) number. High-sensitivity films are often called fast, and low-sensitivity films are called slow. Standard film speeds are ISO 100, ISO 200, and ISO 400. A film rated at ISO 200 needs half as much light to form the same image density as one rated at ISO 100. Note that ISO 400 is one stop faster than ISO 200 and two stops faster than ISO 100. Generally, if you need to take pictures in low light conditions, you need a faster film. You also need a fast film if your subject is in motion because, to freeze the motion, the shutter speed must be set very short. This, in turn, limits the total amount of light available to expose the film.

Automatic cameras have mechanisms, both electronic and mechanical, that simplify the setting of camera parameters. They also restrict the flexibility of the camera and sometimes limit the types of film and lenses that may be used. The photonics laboratory often includes cameras that are designed to interface with other components found on a well-equipped optical bench. A camera is called for when the optical signals produced are images and a permanent, hard-copy record is desired. Cameras are also useful for recording experimental setups and for recording the data from sequences of complex experiments. The photonics technician is typically responsible for camera setup and operation as well as for obtaining film and camera supplies. Operation of laboratory cameras used to record experimental data can be complex, and in all cases the instruction manuals must be consulted prior to use. Inexpensive automatic cameras are adequate for recording experimental setups for archive or documentation use. Inexpensive Polaroid cameras are also useful for these tasks because the film can develop itself quickly into a print or negative.

After images are acquired, the film is sent for processing into final form. This form may be slides, photographs, or negatives. Some film processing facilities offer the option of receiving film images in digital form on compact disks. Photonics experiments may require the use of specially prepared slides for use in later experiments (such as holography). Cameras have appeared recently that use electronic CCD arrays in place of the chemical film media. These cameras store images in electronic computer memories and the results are available nearly instantaneously. However, the availability of "hard copy," or printed results, is dependent on whether a printer is available to produce the prints. Printers capable of high-resolution, high-quality output comparable to that obtainable from chemical film technologies can be prohibitively expensive. CCD cameras are discussed in greater detail in Section II A.

LAB 2: Camera

Using an inexpensive camera whose back has been removed, identify the lens system, shutter, aperture, and film area. Mount the camera on an optics table in front of the CCD array of an inexpensive circuit board “pinhole” camera (with CCD lens removed). Connect the output of the CCD camera to a monitor. Set the shutter speed of the camera to *inf* (infinite) so that the shutter remains open. Try to image an object using the camera’s lens and aperture. After obtaining a suitable image on the monitor, adjust the aperture to f-stops above and below the current setting. Do you observe a doubling and a halving of the brightness of the image on the display? What effect does changing the f-stop have on the clarity and sharpness of the displayed image?

Scanners

Unlike cameras, which are 2-D devices that acquire an image all at once, scanners must capture images one pixel at a time. The very first electronic imaging systems used what is called a *flying-spot scanner*. Although many types of scanning devices exist, we are interested in only *imaging scanners*, like the flying spot, and so our discussion will be restricted to them. Essentially, a scanner converts a 2-D image or picture into a *digital* image or picture, i.e., an image in electronic form manipulated by a computer.

The flying-spot scanner is very simple in construction and consists of a photocell mounted to an x-y positioning system. The components of such a scanner are shown in Figure 9-8. The positioning system locates the photodiode so that it signals the upper left corner of the scene to be imaged. The intensity of the scene is recorded, and the positioning system moves the detector to the right. The output of the sensor is continuously recorded and the system continues in this fashion until it reaches the far limit of the first row of data points. The data points, of course, are the pixels of the image. The positioning system then moves the detector to the start of the second row, and the process repeats itself until the entire scene has been scanned.

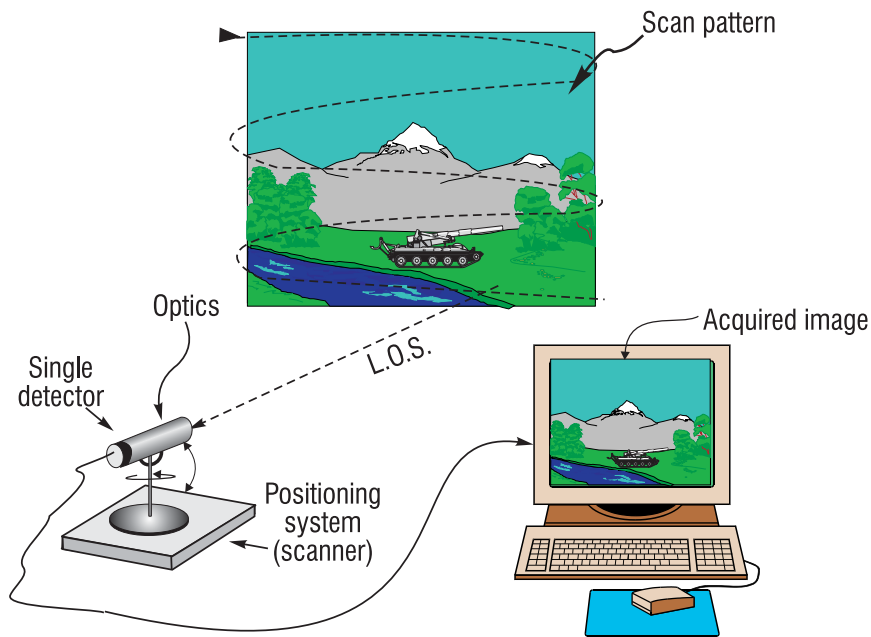


Figure 9-8 *Flying-spot scanner*

The resolution of the image produced by the flying-spot scanner is dependent on the quality and size of the detector, the optics, if any, of the detector, and the resolution and accuracy of the x - y positioning system. This type of scanner is very susceptible to noise produced by the positioning system. Modern scanners use nutating mirrors or detector arrays to minimize the mechanical noise induced by the scanning system. One of the most common scanners is the page scanner, such as the one shown in Figure 9-9. This scanner uses a linear array of detectors that capture one row of pixels at a time from the image. This type of scanner is less susceptible to noise than the flying spot.

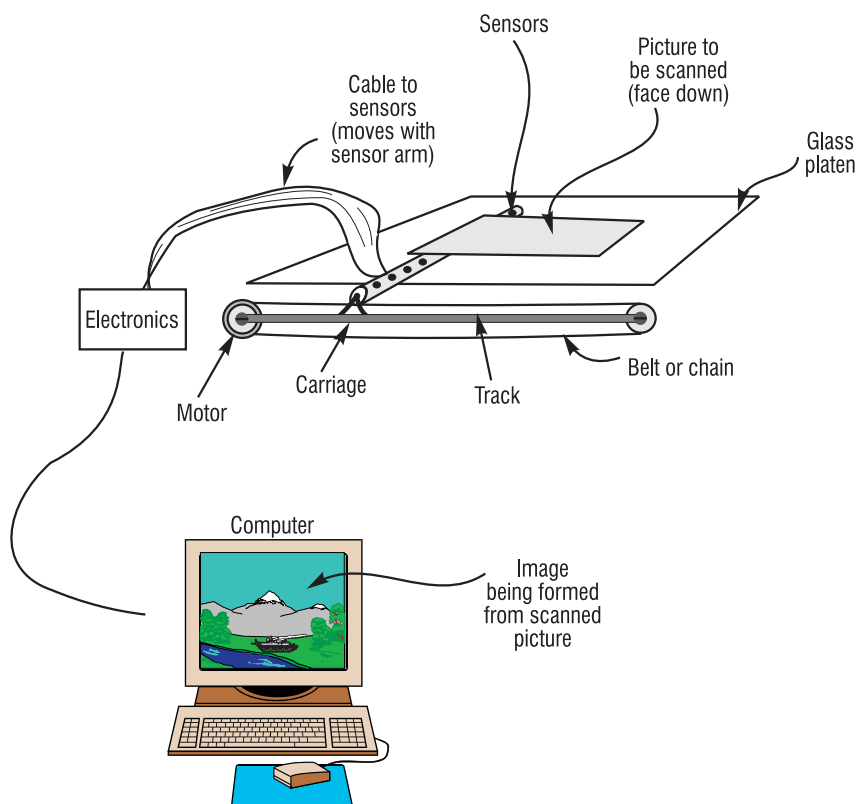
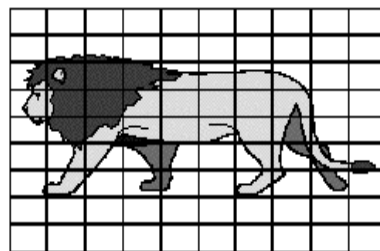


Figure 9-9 *Page scanner*

LAB 3: Scanners

Your instructor will have mounted a set of large posters or screens to the walls of the lab with images such as the lion in the graphic to the right. Set up a tripod-mounted light meter within viewing distance of one of the graphics. Use the viewfinder on the light meter to acquire image intensities from the superimposed grid squares on the poster. Move the tripod as necessary so that only one square is viewed at a time. Record the intensities on a sheet of graph paper so that grid square intensities from the poster match grid squares on the paper.

Now use the METIP Pixel Calculator program on one of the lab computers [see the Other Resources section at the end of this module] to enter the grid data into a blank grid image. What you have done is simulate a “flying-spot scanner.” How does the image that you scanned compare to the graphic in the poster? Calculate the resolution of the poster, your graph paper, and the pixel calculator. How do the resolutions differ? How does this difference relate to the subjective quality of each representation? How could the quality be improved?



Files

When images have been acquired, by either a camera or a scanner, they must be stored for easy access and cataloging. Print images are easily stored using standard office filing procedures, although care must be taken to ensure that proper environmental conditions—such as temperature and humidity—are maintained. Also, some print media are light sensitive even after processing, and these images should be filed in special opaque envelopes designed for photographic storage.

Images captured electronically may be filed on disk or tape for archival purposes. For many years magnetic tape was the medium of choice for archiving image files, but today the removable diskette has replaced tape systems. Two issues of great importance must be addressed when storing image data to files—the resolution of the image and the file storage format. Both of these issues will impact the size of the storage medium required to archive the images. Resolution affects the gross size of the image in terms of computer memory requirements.

A grayscale image with a matrix of 1024×1024 pixels that uses one byte per pixel will require 1 megabyte of storage if the image data are stored in raw form with no further processing. Table 9-1 shows a comparison between common digital storage media in terms of how many raw 1024×1024 images can be archived, the access rate of the device that reads and writes the medium, and the relative stability of the medium. This chart is crude because the capacities may change dramatically in a very short time.

Nevertheless, you can see how choice of medium for archiving images is no trivial task. If the images are data collected from rare events, highly stable media will be best, such as CD-ROM or DVD diskettes. If there is a lot of data, a tape medium may be the best choice. Computer memory has no longevity because the data are lost upon removal of power. Hard disks can be useful for image archiving if the data are used often, since they are the fastest method (short of computer memory) for storing and retrieving data. Hard-disk data are easily backed up using one of the less volatile media to ensure that a device failure does not cause loss of data.

Table 9-1 Storage Medium Capacities

Media	Number of Images*	Speed	Longevity
Computer memory	128	Very Fast	None
Hard disk	10,000	Fast	Low
CD-ROM	600	Medium	Extreme
DVD	5,000	Medium	Extreme
Tape	20,000	Slow	High

*1024-x-1024-pixel images (1 byte per pixel)

Like film media, digital media must be stored carefully according to the manufacturer instructions. Magnetic media such as tapes must be kept away from strong magnetic fields. Power supplies for high-powered laser systems can generate damaging electromagnetic fields, and so caution must be exercised when using or transporting magnetic computer media near these devices. This type of negative influence is called electromagnetic interference (EMI) or radio-frequency interference (RFI), and commercial equipment must be shielded from these fields. However, some special-purpose laboratory equipment may not be adequately shielded and so caution must be exercised.

Digital images are stored as computer *files*. Files are groupings of data that are kept on digital storage media like those described above. When the digital image is stored directly as pixels, the file that holds the data is called a *raw image file*. These files can be quite large and the storage of a large number of them can be difficult and time consuming. Some files use a *compression scheme* to reduce the amount of data that must be stored, but one must be very careful in the selection of the technique used. Compression schemes take advantage of the repetition in pixel characteristics, e.g., color, brightness, etc., that occurs in large collections of data or in the features that make up the image.

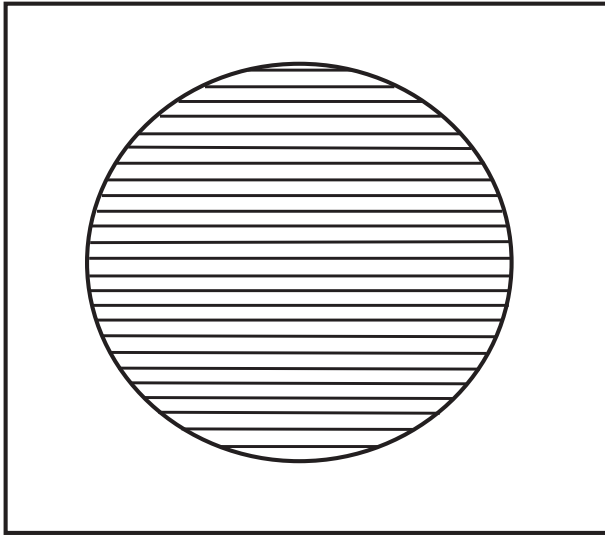


Figure 9-10 *Simple graphic image*

Imagine an image such as that shown in Figure 9-10. This could be an image requiring a large amount of data storage, or we could just store the phrase “A large circle with horizontal lines through it.” Likewise, we could have an image that used only a few different pixel values. For this sort of image, we could make up very short codes to represent the pixel data. These shortened codes could be stored along with a code key to allow us to reconstruct the original pixel values when we want the image restored. In both cases, we could store a very small *encoded* image or description rather than the original large one. With the latter approach, we could get back our original image with no degradation or loss of original pixels. But not so for the first approach, where how the reconstructed image would look would depend completely on the artistic and drawing skill of the individual tasked with recreating “a large circle with horizontal lines through it.” The point of this discussion is that compression schemes may be *lossless* or *lossy*. In almost all cases the lossy schemes give the best data reduction. This is why we say that you must be cautious in using image file formats, since some may give excellent compression but the cost of the compression will be paid in data loss. Table 9-2 lists a few popular file formats and the types of compression schemes that they employ.

Table 9-2 File Formats

Format	Compression Scheme
raw	None—pixel values stored directly.
GIF ¹	Lossless for grayscale images, can be lossy for color images since color imagery is forced to an indexed color map.
JPEG ²	Inherently lossy, JPEG removes redundant data based on a model of human color perception. JPEG can be used to store a lossless image, but the file size will be greater than that of a raw image!
TIFF ³	Compression scheme selectable; most schemes used are lossless.

¹Graphic interchange format ²Joint photographic experts group ³Tagged interchange file format

LAB 4: Image Files

Using a commercial image-processing software such as *Image Pro Plus* by Media Cybernetics [see Other Resources at end of module], display a grayscale image file that your instructor will specify. This image file will have 8-bit pixels at a resolution of 256×256 pixels. Since a byte is 8 bits, this file will require $256 \times 256 = 65,536$ bytes if stored as “raw” pixel values, i.e., no compression. Store the file as GIF and uncompressed TIFF, and observe the file sizes. The GIF image file should be smaller than the TIFF. The TIFF file will be larger than the “raw” file because TIFF adds information to the file besides just the pixel data values. Get a different 256×256 image and store it as GIF. Although the raw files of the two images will be the same size, the GIF image files may be different due to the information content of the images. A very complex image will not compress as well as a simple image. Save an image in JPEG format and set the mode to maximum compression. Now reload the JPEG file that you just saved. You should see a noticeable decrease in the quality of the image.

II. IMAGING DEVICES

Various optical and electro-optical devices comprise an important aspect of photonics work. They are employed by imaging systems as components responsible for converting images into electronic signals for display—or for further conversion into digital images for use in computers. Many different imaging devices are available. The three most popular ones are discussed here.

A. CCD Cameras

Charge-coupled device (CCD) cameras are the most pervasive of today’s imaging devices. A CCD camera uses an *array* of light-sensitive cells formed in silicon. The cells can be thought of as miniature capacitors, where each capacitor is a pixel in the image created by the array. When the array is exposed to light, the capacitors charge up proportional to the intensity of the light falling on the array. This charge, which is discriminated by values of voltage, is then read off of the array, converted to a digital signal, and transferred into a computer. The signal can also be sent directly to a CRT device (see Section III A) for viewing. The CCD array can

replace the film medium in a camera, where the camera optics form an image on the array instead of on the usual film.

CCD cameras have electronic shutters that control the integration time at each pixel. This means that the charging time of the capacitor is controlled. If the capacitor is allowed to charge a long time, the pixel will have more time to charge. Exposure in the camera is also dependent on the device physics used in the CCD array construction. Most cameras have automatic exposure controls that operate to the limits of the camera specifications and are nothing more than automatic gain controls that limit the electrical signal output by the camera or hold the signal to a constant value. More expensive cameras allow adjustments to be made under computer control and are equipped with special interface units.

Figure 9-11 shows the basics of how a CCD array works.

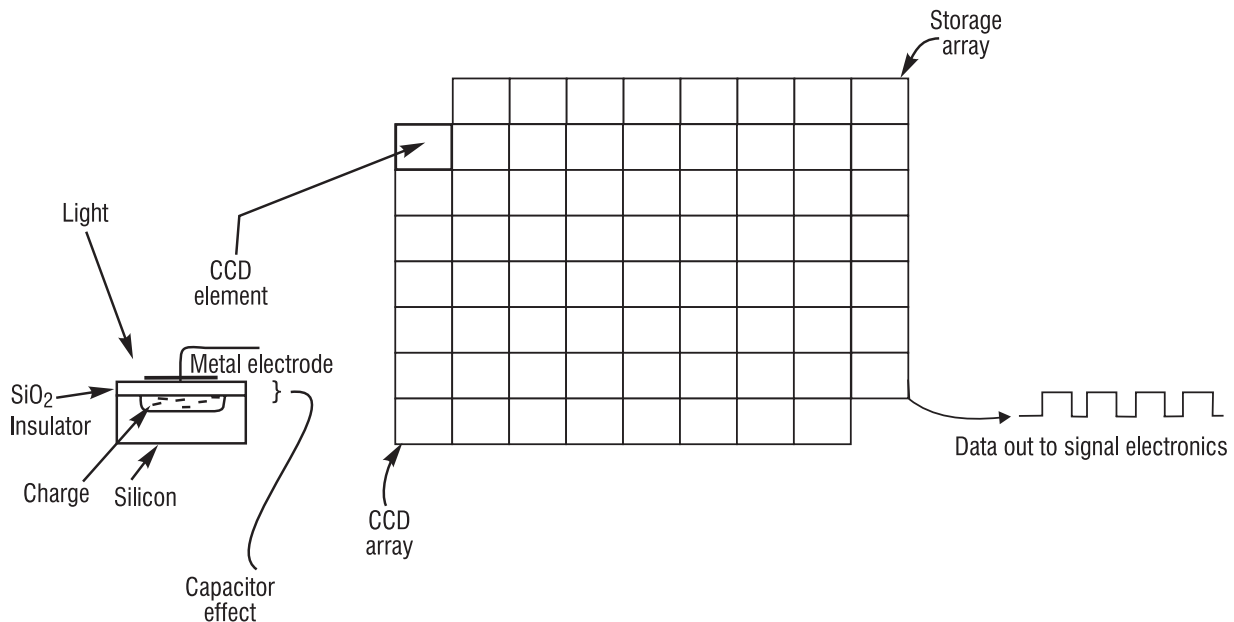


Figure 9-11 *CCD array*

Note that the image data in the electrical signal output from the array must leave the array serially, in single file. This process is illustrated in Figure 9-12. The pixel values exit the array in synchronization to an electronic clock signal. The top row of pixel values (voltages) is clocked out of the chip, then the pixel values are shifted up (copied) by one row and these are then clocked out, and the process repeats until all pixels have been read out. You may wonder about whether or not a shutter mechanism is required. The answer is yes. The shutter action is accomplished electronically by a signal that copies the charge from all of the pixels on the primary CCD array to another CCD array—called the storage array—behind the imaging array. This second array is used to transfer the pixels out of the device.

Color CCD arrays require filters to control the wavelength of light that strikes the individual elements of the array. For this reason, the color CCD is more complex physically, electronically, and optically—although the fundamental operation of the array is similar.

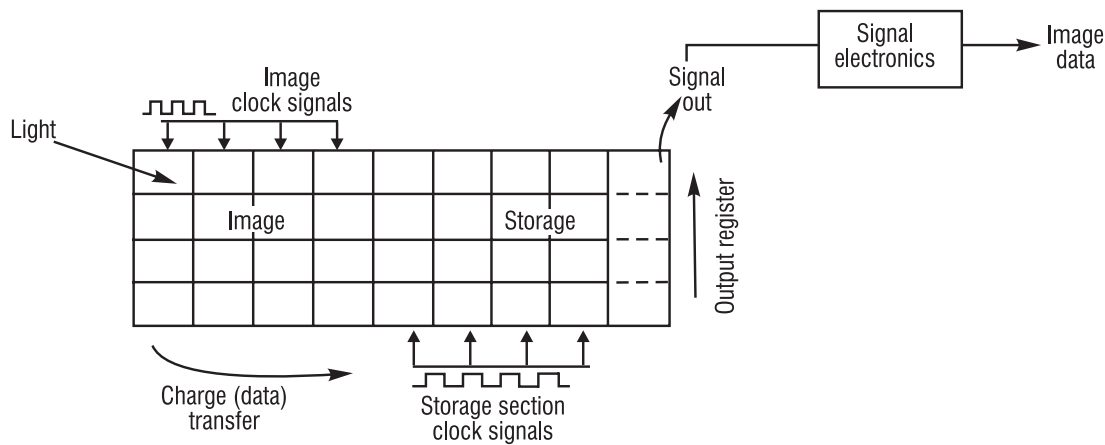


Figure 9-12 *CCD pixel-data transfer*

Structure

The structure of a CCD camera is identical to that of a film camera with two exceptions: 1) the CCD camera replaces the film medium with a CCD array, as discussed above, and 2) the shutter action is performed electronically. CCD cameras have widespread use and can be made very inexpensively and with very high resolution. Small, inexpensive cameras are used heavily by the security and surveillance industry. All modern camcorders contain CCD arrays. Almost all modern astrophysical recording is done using high-resolution CCD arrays. Finally, medical imaging in hospitals is converting from chemical film processes to CCD-captured digital imagery.

Capabilities

A number of specifications are used to characterize the capabilities of CCD cameras, but the primary ones are resolution in terms of pixel count and geometry, array size, pixel quantization, and frame rate. Arrays are now available with resolutions of 4096 by 4096 pixels, although combined array cameras of 8192 by 8192 pixels formed from sixteen 2048-by-2048 arrays were available at the time of this writing. Odd geometries are also prevalent, such as 1317 by 1035 or 2112 by 2070 pixel arrays. Electronic shutter speeds vary and range from less than one frame per second to over 100 frames per second. Pixel quantization is defined in terms of bits, as we mentioned earlier. Typically, the CCD pixel employs 8-bit quantization yielding a grayscale of 2^8 or 256 levels. Laboratory imaging generally requires 12-bit quantization or 4096 levels of grayscale.

There are rules of thumb to be applied when discussing CCD capabilities. As resolution of the array increases, frame rate decreases. This is due to the fact that the pixels must be removed from the array serially, so the more pixels acquired the longer it takes to get them out of the array. If the images produced by the CCD are to be used for human viewing, 8-bit quantization is usually adequate. However, for computer image processing the rule is that the greater the quantization the better the processing.

B. Vidicons

The *vidicon* is an electron tube that has been designed to capture images and convert them to electrical signals. The signals produced by a vidicon are similar to those generated by a CCD array, but they are produced from an electron beam as opposed to a silicon charge system. The vidicon tube was invented in 1951 and has seen little change in basic design since that time. It was used extensively in the broadcast (TV) industry until the advent of the CCD. Likewise, the CCD camera is slowly replacing the vidicon in both scientific and medical imaging as CCD technology improves and vidicon technology ages. The two greatest drawbacks to the vidicon are the delicacy of the vacuum tube and the high voltage requirements for the electron scanning beam.

Structure

The components of the vidicon tube are diagrammed in Figure 9-13. A cylindrical glass tube is constructed with a flat glass plate window at one end. Inside the tube and behind this window is a *photoconductive target*. The material from which this target is made has the property that electrical resistance varies according to the intensity of illumination that strikes it. An external lens system is used to image a scene onto the window of the vidicon, which then charges the target material to varying intensities depending on the illumination. An electrode is placed at the other end of the tube. The electrode produces an electron beam when the glass tube is sealed under a vacuum. The beam is focused and deflected by a set of coils that surround the tube. The beam is scanned across the target, and an electrical signal is developed that is then output from the target. The voltage of this signal is proportional to the resistance of the target at the position of the beam and to the light intensity striking the target at that location. The signal output is then synchronized to the electron beam and output as the electronic image.

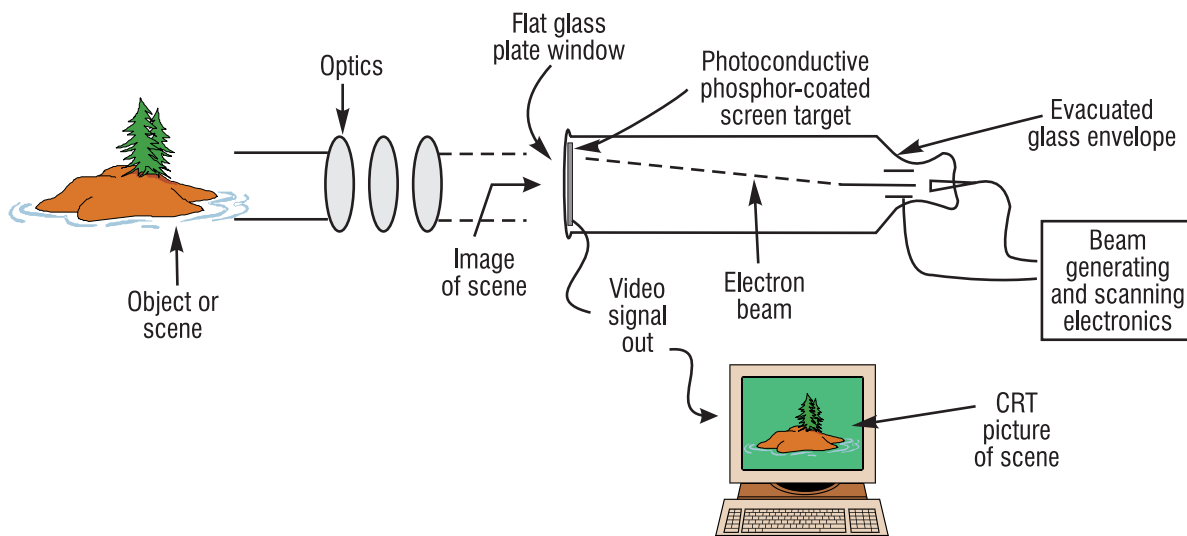


Figure 9-13 Vidicon tube

Capabilities

As discussed earlier, vidicons are becoming obsolete as the capabilities of CCD arrays are improved. At one time they were the system of choice for medical and scientific imagery because of their high speed and high resolution. But now the CCD array camera has exceeded these capabilities. So companies have formed that offer CCD replacement units for vidicon tube cameras.

The delicacy of the glass vidicon tubes and the complexity of the electronics required to drive them have contributed to their demise. The resolution of the vidicon is dependent on the construction of the glass envelope, the type of photoconductive material used, and the complexity of the electronics. If the electronics are susceptible to noise or suffer from thermal drift, the scanning beam will not accurately track the image formed by the phosphors and final image quality will be degraded.

C. Image Intensifiers

The *image intensifier* is a vacuum-tube device (similar to the vidicon) that accepts an image at one end and produces an image of higher intensity at the other. The image intensifier can be considered an image amplifier that uses energy to achieve the necessary amplification. These devices are used in low-light-level situations such as night vision and astronomy. Image intensifiers were first developed to work with vidicons and other early electronic camera technologies and were often manufactured as components of these cameras. Modern image intensifiers are used as adapter to cameras or by themselves as night-vision devices under starlight conditions.

Structure

The image-intensifier tube has a *photocathode* screen at one end and a phosphor screen at the other end as shown in Figure 9-14. The photocathode is a material, such as gallium arsenide, that emits electrons when exposed to light. An optical system is used to image a scene onto the photocathode, and electrons are emitted in proportion to the amount of light imaged. The electrons are accelerated by an electric field, which produces a gain in terms of number of electrons at the phosphor screen. When the electrons strike the phosphor screen, the screen emits visible light and produces an intensified image on the photocathode screen of the scene formed at the input end of the tube.

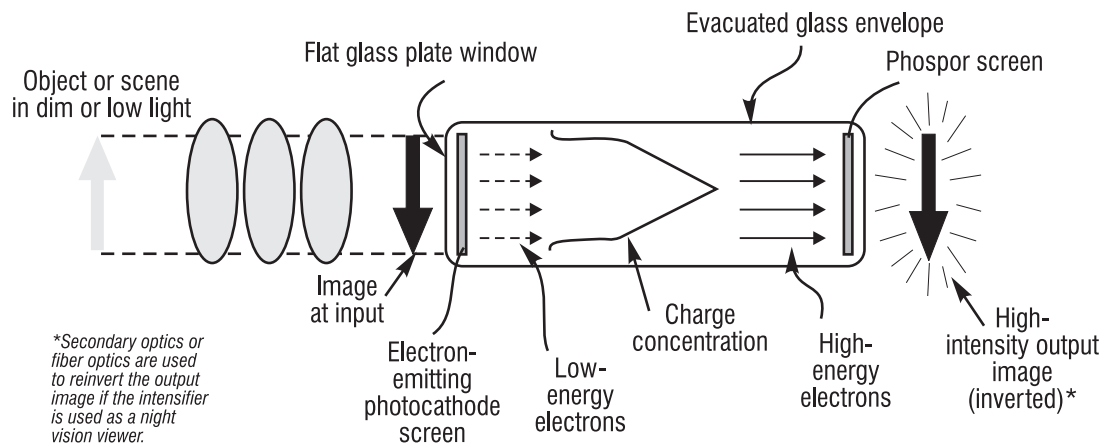


Figure 9-14 Image intensifier

Capabilities

Early image intensifiers designed in the 1960s could amplify light by 1000 times. Current intensifiers can amplify upward of 50,000 times. These systems can develop the illusion of daylight when used for night vision under starlight-only conditions. An image intensifier amplifies light, so it will not operate when no light is falling on the input screen.

LAB 5: Image Intensifier

Extinguish the room lights and use an image intensifier to experience “night vision.” Turn the lights back on and set up a CCD circuit board “pinhole” camera to image an object on the bench. Display the object on a monitor. Mount the image intensifier in front of the CCD camera and again extinguish the room lights. The CCD camera is now imaging the output of the intensifier and thus is able to operate in the darkened room.

III. DISPLAY DEVICES

Display devices are the complement of imaging devices and are used to output images for viewing. Our discussion of displays is restricted to a few purely electronic display technologies that output images, although other technologies—such as printers and electromechanical systems—exist that can display imagery. These devices use various approaches to take an image signal from an imaging device, or from a computer, and display the pixel data so that the image can be observed. All display devices have a common capability and that is to output a pixel value visually.

A. Introduction to Cathode-Ray Tubes

The cathode-ray tube (CRT) is the oldest electronic display technology and—unlike vidicons and flying-spot scanners—is in little danger of becoming technologically extinct in the near term. The CRT is a vacuum tube, much like a combination of a vidicon and an image intensifier. The image signal to be displayed is input as a serial sequence of pixel data, and these data are displayed through the use of luminous phosphors at the viewing end of the tube.

Construction

Figure 9-15 shows a schematic drawing of how a basic CRT is constructed. The principal component parts are labeled. At one end is the *electron gun*. This unit produces a stream of electrons that are modulated in intensity by the image signal as it is accessed. The electrons are accelerated toward the opposite end of the tube by electromagnetic coils. The coils also play a role in focusing the beam, which also passes through opposing x-y charged plates that are used to electronically sweep the beam across a phosphor-coated screen at the far end. The electronics required to synchronize the beam to the input signal stream and to perform the scanning operation are quite sophisticated. At the end of the scan, the electronics must cause the beam to shut down while it is sent back to the start point of the scan.

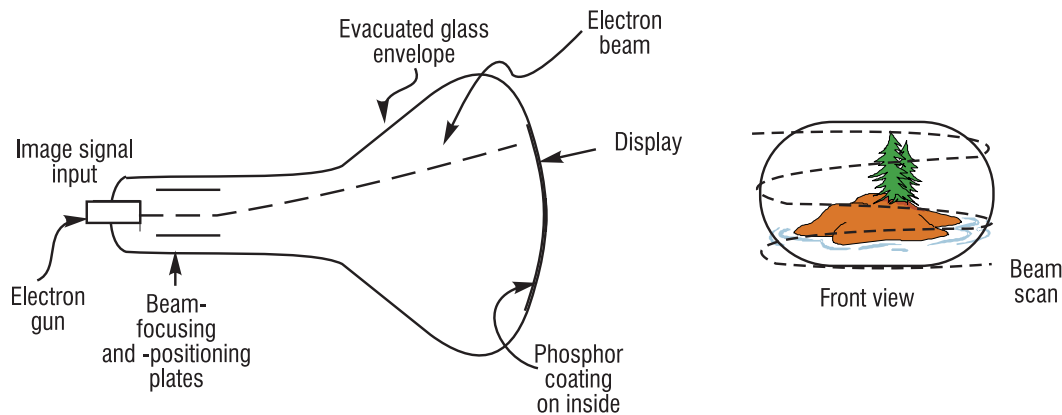


Figure 9-15 Cathode-ray tube (CRT)

Capabilities

CRT displays are capable of very-high-resolution output and can display very large images. They can also handle very high frame rates and can output black and white (intensity only) images or color images. Modern CRT displays are equipped with electronic controls so that they can automatically adjust to various resolutions and color requirements of images output from computers or received as television signals.

B. Flat-Panel Liquid-Crystal Displays

Liquid crystals are substances whose material state is somewhere between crystals and liquids. They appear to be like a gel and have very interesting and useful properties. Liquid crystals are light-polarizing substances, and their polarization can be modified by the application of an

electric field. This property makes them useful in display technologies. The advantages of LCD displays include low cost, small size, and low power consumption. Their primary disadvantage is that they modulate light and so require external light sources to operate. The light source used to illuminate a large LCD display, such as those used for computer or television screens, is typically supplied from *behind* the display itself and is called a *backlight*.

Liquid-Crystal Theory

Liquid crystals are classified into three different phases, called *nematic*, *smectic*, and *cholesteric*. In the *nematic phase*, individual molecules are threadlike and longitudinally arranged. In the *smectic phase*, molecules organize themselves into layers. In the *cholesteric phase*, molecules in the different layers orient themselves at a slight angle relative to each other (rather than parallel as in the nematic phase) and take on a helical orientation. The three phases are illustrated in Figure 9-16. These phases are important to the use of liquid crystals as a display methodology. Liquid crystals have birefringent properties, which means that they can polarize light. However, the polarization can be switched using an electric field. This property gives rise to their use in displays.

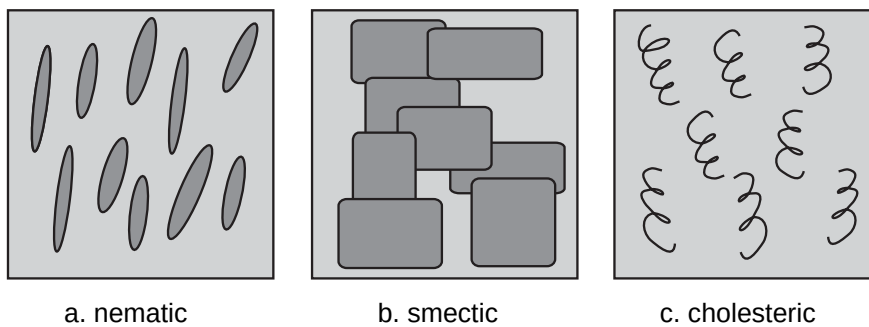


Figure 9-16 Liquid-crystal phases

In Figure 9-17 we see how a simple liquid-crystal display is constructed. Such a display might be used in a digital watch or the display of a small computer controller in an appliance or instrument. The liquid-crystal material is placed between two glass plates, one of which is coated with a transparent metal-oxide film. This film is one electrode of the display. The other electrode is formed into shapes, patterns, or symbols, and a separate wire lead is attached to each. The front of the display has a polarizer. When the display is inactive, the light passing through the assembly remains randomly polarized and nothing is observed. However, when one or more of the electrodes is energized by the application of a small electric field, the crystals align themselves such that they produce a polarization that is perpendicular to the polarizer layer. This behavior causes the symbol or pattern defined by the electrode to be darkened, as shown in Figure 9-18.

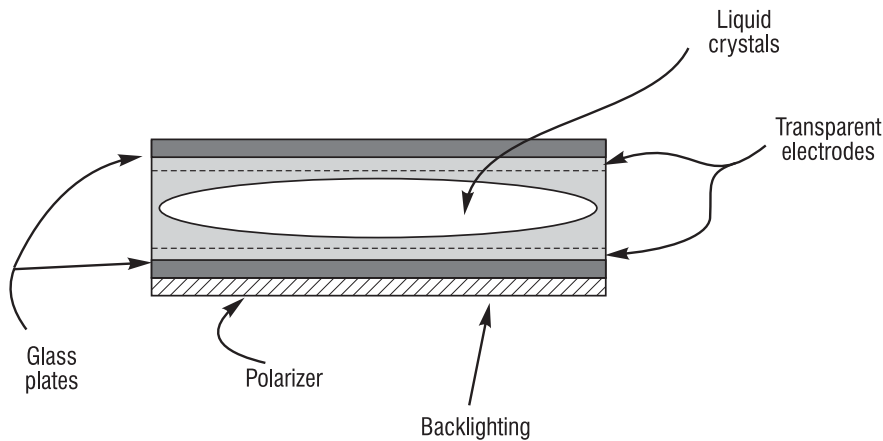


Figure 9-17 Simple LCD construction

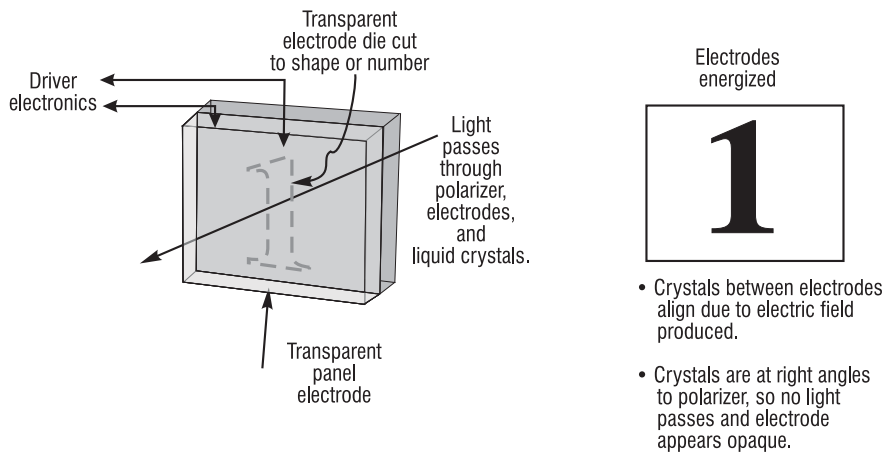


Figure 9-18 LCD operation

The LCD can produce a grayscale effect depending on the amount of charge placed on the electrodes. The strength of the charge controls the degree of alignment of the crystals, hence the amount of light blockage by the polarization effect. Color can also be incorporated into these displays by the addition of red, green, and blue filters.

LCDs capable of displaying images, where a pixel array is required, are called *matrix LCDs* since the image array is a matrix of pixels. These displays are considerably more complex than the simple displays described above. However, they represent the future of imaging display technology and will eventually displace CRTs, as indicated in Figure 9-19. This figure shows the displacement of CRT technology by LCD in terms of sales volume.

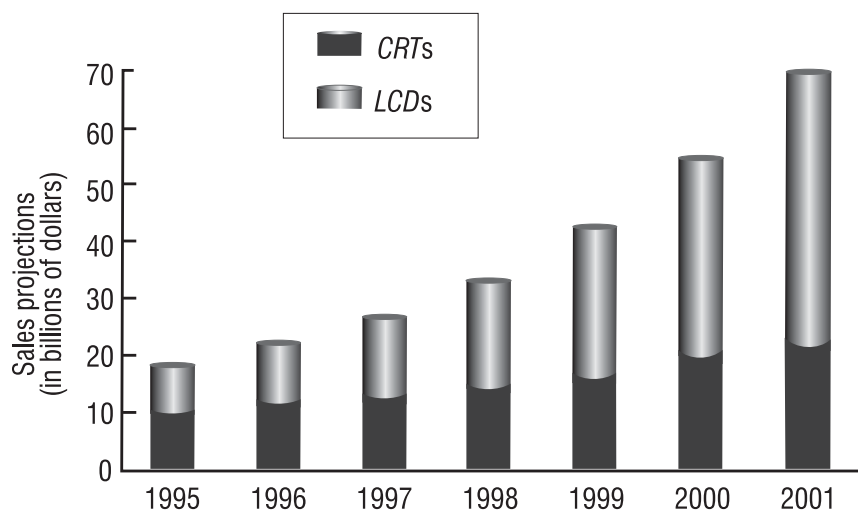


Figure 9-19 *Worldwide market for CRT versus LCD displays*

The matrix display must address a two-dimensional array of pixels. Unlike the scanning beam of the CRT that traces back and forth across the screen, the pixels in a matrix display must be individually addressed. There are two basic matrix LCD types, passive and active, and the differences between the two are discussed next.

Passive-Matrix Liquid-Crystal Displays

A passive-matrix LCD incorporates a grid arrangement of electrodes where all the pixels in a column are connected and all the pixels in a row are connected. To address a single pixel, the column/row of that pixel is energized. This is illustrated in Figure 9-20.

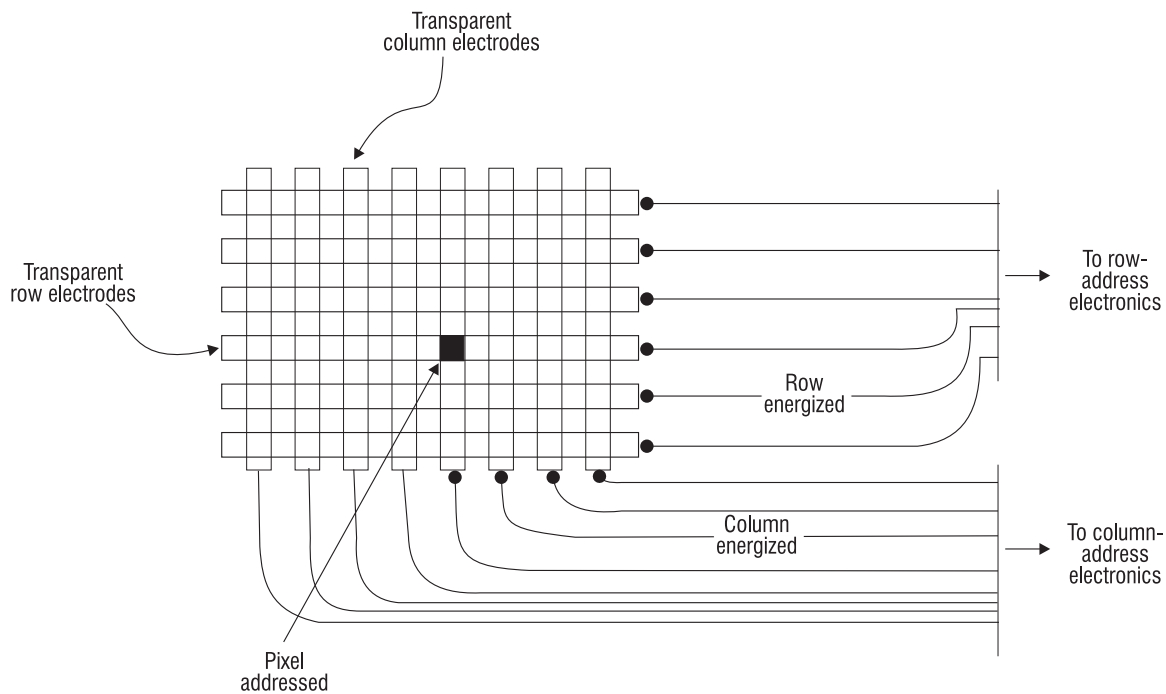


Figure 9-20 LCD row-column addressing

The display is updated using a scanning process and so is very slow—not fast enough to display movies. The row-column grid can allow leakage of charge, so pixels can appear smeared or fuzzy. The passive display is very inexpensive and is used primarily for imaging displays where high resolution and speed are not required. Passive LCDs can take advantage of dual-scan refresh, where the screen is divided into two sections that are refreshed simultaneously. Dual-scan displays are not as sharp or bright as active-matrix displays, but they consume less power.

Active-Matrix Liquid-Crystal Displays

Thin-film transistors (TFT) are switching transistors that can be fabricated onto the glass layer of an LCD panel. The transistors may then be used to directly switch the field of the LCD pixel, providing a substantial increase in display-refresh speed and in sharpness and clarity. The use of the TFT to actively switch the LCD pixels gives rise to the terminology *active-matrix* LCD. These displays are substantially more expensive than those of the passive type, due to their complexity. But the higher resolution and speed have placed them in great demand for high-end portable computers. In time, the active-matrix LCD will dominate the LCD market and ultimately eclipse that of the CRT.

C. Flat-Panel Electroluminescent Displays

Electroluminescent displays (ELD) are very similar to LCDs with the primary difference being that they generate light through the process of electroluminescence. When a voltage is applied directly to certain phosphors, they emit light in the same way that they do when struck with an electron beam in a CRT. ELDs have more limited applications than LCDs due to their higher cost of construction and the fact that full-color systems have not been developed.

D. Flat-Panel LED Displays

LED stands for *light emitting diode*, and these electronic components are true photonic devices. LEDs are now available in almost any visible wavelength. Single, high-power, high-output LED units are replacing tungsten-filament light bulbs in many applications. An LED display is formed from an array of LEDs where each LED serves as a pixel. The modulation of current to the LED varies the light output and thus produces varying intensity levels. If clusters of red, green, and blue LEDs are used at each pixel position, a color image can be produced. LED displays are simpler in construction than either LCD or ELD systems. However, they are as yet incapable of the resolution of these displays and are very expensive to produce. Like the ELD, however, they produce light and so do not require an external light source for viewing. The LED display does require a greater amount of power to operate than does an LCD or an ELD.

Looking Toward the Future

Image-processing and -display technology is changing daily, and these changes will impact the photonics technician of the future. One major aspect of imaging that is guaranteed to change is the resolution of displays. The two fields of work most likely to be impacted by this will be medical and military imaging. Both of these areas use photonics heavily and both are increasingly interested in capturing, displaying, and analyzing images with high information content. Systems of the future will have resolutions that exceed the capability of human sight and will place great demands on the computing systems that analyze these images. The installation and maintenance of these systems will be demanding for the photonics technician as well. In the future, display systems will place more reliance on optics and it will not be uncommon to see 3-D holographic displays and head-mounted displays of great resolution and speed so that the wearer is immersed in the image. Much of this virtual-reality work relies heavily on display technology, and the photonics technician will play a critical role in that field.

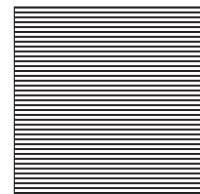
Laboratory

In this laboratory you will determine the resolution of a CCD camera.

Equipment List

The following equipment is needed to complete this laboratory:

- 1 Logitech QuickCam™ CCD camera
- 1 computer to support QuickCam™
- 1 table-tripod or holder for camera
- 1 ruler marked in tenths of an inch or millimeters
- 75-LPI test target →



(Note: You may use the target on this page for this experiment.)

Procedure

1. Verify that the target has a resolution of 75 lines/inch. Do this by first measuring the target with the ruler. The target should be 1" square. Now count the lines in the target, which should be 75. **Remember to include the white lines when counting.**
1. Connect the camera to the video monitor. Set the camera so that the target is imaged on the monitor with the lines on the target oriented vertically. Adjust the focus so that the lines are clearly distinguishable.
2. Move the camera (or the target) back until the lines are no longer clear, but keep the camera in focus. The goal is to stop moving the camera at just the point before the lines are no longer distinguishable. This is the point at which the camera sensors can just *resolve* the lines and we are at the resolution limit. The camera now has one line per pixel imaged onto the CCD array.
3. Measure the width in inches of the target appearing on the monitor and measure the width of the image displayed. The software that accompanies the QuickCam™ allows you to display the image in a window. This window should be at maximum size.
4. Compute the number of (horizontal) pixels of the camera using the following formula:

$$\text{Number of pixels} = (75 \text{ lines/size of target in window}) \times \text{size of window}.$$
5. To determine the resolution of the camera in lines per inch (LPI), we need to know the size of the CCD array. These data should be listed in the camera. The specifications will also list the number of pixels in the CCD, so that you can compare your experimental result to the actual specifications of the array.
6. Now turn the target so that the lines are horizontal and repeat the experiment to determine the vertical pixels. Many CCDs (like the QuickCam™) have different numbers of horizontal and vertical pixels.

Student Project (optional)

Choose a recently unveiled display technology that has not been covered in the course (examples: Seiko Chip-on-Glass© or Phillips 3D-LCD). Obtain vendor literature and possibly the demo kits. Report on how the new technology is improving on current state-of-the-art devices. Determine the resolution of these products and draw a cost-per-pixel bar chart.

Other Resources

Associated Available Software

- *METIP Pixel Calculator*, University of Washington Math Experiences Through Image Processing Program—<http://www.cs.washington.edu/research/metip/metip.html>
- *Image Pro Plus* by Media Cybernetics—<http://www.mediacy.com/ippage.htm>

- MATLAB Image Processing Toolkit by Math Works, Inc.—
<http://www.mathworks.com/products/image/>

References

Textbooks

- Andrews, Harry C. *Computer Techniques in Image Processing*. New York: Academic Press, 1970.
- Myler, H. R. *Fundamentals of Machine Vision*. Bellingham: SPIE Press, 1998.
- Myler, H. R., and A. R. Weeks. *Computer Imaging Recipes in C*. Englewood Cliffs: Prentice Hall, 1993.
- Ono, Y. A. *Electroluminescent Displays*. World Scientific, 1995.
- Russ, John C. *The Image Processing Handbook*. Boca Raton: CRC Press, 1992.
- Tsukada, Toshihisa. *TFT/LCD: liquid-crystal displays addressed by thin-film transistors*. Amsterdam: Gordon and Breach, 1996.

Articles

- Musa, S. “Active-Matrix Liquid-Crystal Displays,” *Scientific American*, November 1997.

Equipment

LCD/LED Displays

- Digi-Key Corporation, 701 Brooks Avenue South, Thief River Falls, MN 56701.
- Jameco Electronics, 1355 Shoreway Road, Belmont, CA 94002.

Image Intensifiers

- ITT Industries Night Vision, 7671 Enon Drive, Roanoke, VA 24019.

“Pin Art”

- Action Products International, Inc. , 390 North Orange Avenue, Orlando, FL 32801

CCD “Pinhole” Cameras

- Hosfelt Electronics, 2700 Sunset Boulevard, Steubenville, OH 43952.

Logitech QuickCam Camera

- Logitech, Inc., 6505 Kaiser Drive, Fremont, CA 94555.

Problem Exercises

1. A CCD sensor array is 8.5 mm by 8.5 mm with 300 rows by 300 columns of sensors. What is the resolution of this array in lines per inch?
2. A computer monitor can produce 4096 colors using red, green, and blue pixels. Can you explain this in terms of bits per pixel?
3. A computer monitor is 15" by 15" with a resolution of 1024 by 1024 pixels. If a binary image is displayed on the monitor, what will be the result if the image is printed on an 8.5" by 11" sheet of paper by a laser printer capable of 300 DPI?
4. Photographs taken during a recent experiment have turned out to be too light. What action should be taken to correct the problem?
5. List advantages and disadvantages between flying-spot scanners and flatbed scanners.
6. A high-resolution CCD camera produces RGB color images that are 1024×1024 pixels with a bit-depth of 24 bits. Assuming that the images are stored in raw format, how many of these images could a CD-ROM hold?
7. You have been given a floppy disk for analysis. The disk contains image files produced by a high-resolution camera. Upon listing the disk files you note that they are in JPEG format. Does this present a problem?
8. You have just finished installing a new CCD camera into an imaging system in the lab, BUT the system cannot maintain the frame rate that it had before. What could be the problem?
9. An experiment that you are conducting generates very-low-intensity images and your CCD camera set at the widest possible aperture still produces unacceptably dark images. What might correct this problem?
10. An LCD display has a dead row of pixels. What could explain this?
11. An LCD has a resolution of 640 by 480 pixels. Will an image from a CCD camera with 380 lines of resolution be displayed at full resolution?
12. An LCD display is used to display images from a CCD camera. You notice that fine detail is missing from objects in the image. What could explain this?

Basic Principles and Applications of Holography

Tung H. Jeong
Lake Forest College
Lake Forest, Illinois

Holography is a much broader field than most people have perceived. Recording and displaying truly three-dimensional images are only small parts of it. Holographic optical elements (HOE) can perform the functions of mirrors, lenses, gratings, or combinations of them, and they are used in myriad technical devices. Holographic interferometry measures microscopic displacements on the surface of an object and small changes in index of refraction of transparent objects like plasma and heat waves. Future photonic devices such as electro-optical chips will undoubtedly incorporate micro-lasers and HOEs for optical computations, free-space interconnects, and massive analog and digital memory systems.

Prerequisites

To understand the concepts presented here, you will need to have studied Module 1-1, *Nature and Properties of Light*; Module 1-2, *Light Sources and Safety*; Module 1-3, *Basic Geometrical Optics*; and Module 1-4, *Basic Physical Optics*. In this module, a physical model will be developed so that all major features of basic holography can be explained through visualization and without the use of mathematics. Basic trigonometry will be helpful, but not required, for quantitative understanding.

Objectives

At the conclusion of the module, the student is expected to:

- Understand that there are two basic types of holograms—transmission and reflection.
- Know how to illuminate and view the two types of holograms.
- Be able to visualize the interference pattern between two point sources of light of the same wavelength and how this pattern changes depending on the relative position of the sources.
- Understand that making a hologram requires the recording of interference patterns between light from a fixed point source and light from each point on an object.
- Explain why the images of holograms are truly three dimensional.
- Know how to reconstruct the real and the virtual images from a transmission hologram.
- Understand why a transmission hologram can be broken into pieces and each piece can recreate the entire recorded image.
- Be able to make a reflection hologram and explain why the image can be viewed with incandescent light.
- Be able to make a transmission hologram with more than one channel of information and explain why.
- Understand the concept of coherence length as applied to laser light and holography.
- Define holographic interferometry.
- Know how to make a holographic diffraction grating with very high dispersive power.

Scenario—A day in the life of a holographer

Maria works for a company that has received a new contract from a software firm for producing 1,000,000 small holograms to be used as security seals on all its products. Having been trained as a holography technician, Maria is responsible for making the master hologram by using the three-dimensional model of a logo provided by the customer. She sets up the basic split-beam, focused-image, reflection-hologram configuration on the optical table and aligns the components using techniques learned in the holography course she completed in school. She produces an excellent holographic image of the logo that seems to stick out from the surface when illuminated with any point source of white light. She then hands the image to her friend, another technician in the replication department, who treats the hologram as if it were an original object and makes a million copies of it using rolls of photopolymers. All this is in a day's work for Maria and her friend.

Opening Demonstration

In a darkened classroom, demonstrate the properties of a reflection and a transmission hologram:

1. Use a point source of incandescent light—such as a “spotlight” available in hardware stores—and illuminate a typical white-light reflection hologram. Demonstrate the full parallax of the virtual image by allowing students to look up and down and left to right.

Now hold the emulsion side of the hologram on top of a cup of hot water (tea, coffee) and allow the steam to “swell” the gelatin. Observe the image again and notice that the color has been shifted toward red. This demonstrates that reflection holograms, although recorded on “black-and-white” photographic emulsion, can recreate three-dimensional images in color. This shows the essence of Lippmann photography (Nobel Prize in physics, 1908) and Bragg diffraction (Nobel Prize in physics, 1915).

2. Use a two-channel transmission hologram. Project a real image onto a screen using a *collimated* beam of laser light (or a laser pointer). Move the laser beam around the hologram and show that each spot recreates a distinctly different view of the object. Thus, if the hologram were broken into pieces, each piece would show a complete view of the recorded object.

Now change the angle of incidence on the same hologram and note that a completely different view appears—with all the same characteristics of the first image.

With use of a diode laser without the collimator, the diverging light can now be used to show the virtual image of each of the two channels in the hologram. Students must come up individually and look through the hologram as if it were a window with the object(s) on the other side. Point out that this demonstrates the enormous information-storage capability of holograms.

Basic Concepts

I. TYPES OF HOLOGRAMS

A *hologram* is a recording in a two- or three-dimensional medium of the *interference pattern* formed when a point source of light (*the reference beam*) of fixed wavelength encounters light of the same fixed wavelength arriving from an object (*the object beam*). When the hologram is illuminated by the reference beam alone, the *diffraction pattern* recreates the wave fronts of light from the original object. Thus, the viewer sees an image indistinguishable from the original object.

There are many types of holograms, and there are varying ways of classifying them. For our purpose, we can divide them into two types: *reflection holograms* and *transmission holograms*.

A. The reflection hologram

The reflection hologram, in which a truly three-dimensional image is seen near its surface, is the most common type shown in galleries. The hologram is illuminated by a “spot” of white incandescent light, held at a specific angle and distance and located on the viewer’s side of the hologram. Thus, the image consists of light **reflected** by the hologram. Recently, these holograms have been made and displayed in color—their images optically indistinguishable from the original objects. If a mirror is the object, the holographic image of the mirror reflects white light; if a diamond is the object, the holographic image of the diamond is seen to “sparkle.”

Although mass-produced holograms such as the eagle on the VISA card are viewed with reflected light, they are actually transmission holograms “mirrorized” with a layer of aluminum on the back.

B. Transmission holograms

The typical transmission hologram is viewed with laser light, usually of the same type used to make the recording. This light is directed from behind the hologram and the image is transmitted to the observer’s side. The *virtual image* can be very sharp and deep. For example, through a small hologram, a full-size room with people in it can be seen as if the hologram were a window. If this hologram is broken into small pieces (to be less wasteful, the hologram can be covered by a piece of paper with a hole in it), one can still see the entire scene through each piece. Depending on the location of the piece (hole), a different perspective is observed. Furthermore, if an undiverged laser beam is directed backward (relative to the direction of the reference beam) through the hologram, a real image can be projected onto a screen located at the original position of the object.

C. Hybrid holograms

Between the reflection and transmission types of holograms, many variations can be made.

- *Embossed holograms*: To mass produce cheap holograms for security application such as the eagle on VISA cards, a two-dimensional interference pattern is pressed onto thin plastic foils. The original hologram is usually recorded on a photosensitive material called *photoresist*. When developed, the hologram consists of grooves on the surface. A layer of nickel is deposited on this hologram and then peeled off, resulting in a metallic “shim.” More secondary shims can be produced from the first one. The shim is placed on a roller. Under high temperature and pressure, the shim presses (embosses) the hologram onto a roll of composite material similar to Mylar.
- *Integral holograms*: A transmission or reflection hologram can be made from a series of photographs (usually transparencies) of an object—which can be a live person, an outdoor scene, a computer graphic, or an X-ray picture. Usually, the object is “scanned” by a camera, thus recording many discrete views. Each view is shown on an LCD screen illuminated with laser light and is used as the object beam to record a hologram on a narrow vertical strip of holographic plate (*holoplate*). The next view is similarly recorded on an adjacent strip, until all the views are recorded. When viewing the finished composite hologram, the left and right eyes see images from different narrow holograms; thus, a stereoscopic image is observed. Recently, video cameras have been used for the original recording, which allows images to be manipulated through the use of computer software.
- *Holographic interferometry*: Microscopic changes on an object can be quantitatively measured by making two exposures on a changing object. The two images *interfere* with each other and fringes can be seen on the object that reveal the vector displacement. In *real-time holographic interferometry*, the virtual image of the object is compared directly with the real object. Even invisible objects, such as heat or shock waves, can be rendered visible. There are countless engineering applications in this field of *holometry*.
- *Multichannel holograms*: With changes in the angle of the viewing light on the same hologram, completely different scenes can be observed. This concept has enormous potential for massive computer memories.
- *Computer-generated holograms*: The mathematics of holography is now well understood. Essentially, there are three basic elements in holography: the light source, the hologram, and the image. If any two of the elements are predetermined, the third can be computed. For example, if we know that we have a parallel beam of light of certain wavelength and we have a “double-slit” system (a simple “hologram”), we can calculate the diffraction pattern. Also, knowing the diffraction pattern and the details of the double-slit system, we can calculate the wavelength of the light. Therefore, we can dream up any pattern we want to see. After we decide what wavelength we will use for observation, the hologram can be designed by a computer. This computer-generated holography (CGH) has become a sub-branch that is growing rapidly. For example, CGH is used to make holographic optical elements (HOE) for scanning, splitting, focusing, and, in general, controlling laser light in many optical devices such as a common CD player.

II. THE PHYSICAL BASIS OF HOLOGRAPHY

The Physical Model

To make holography accessible to a general audience with widely varied backgrounds, a physical model is useful. Just as chemists use sticks and balls to help them visualize the structure of molecules, our model will allow us to visualize, and thus “understand,” the physical characteristics of holograms without using advanced mathematics.

Two-Source Interference

In two dimensions (the plane of this paper), the pattern of waves from a stationary source generating waves at constant frequency (and wavelength) is a set of concentric circles (see Figure 3-2 [a] and [b] in Module 1-3, *Basic Geometrical Optics*, as an example of water waves). The distance between any two adjacent circles is one wavelength (λ). Let each circle represent the crest of a wave. Halfway between any two waves is the trough. The pattern shown in Figure 10-1 represents a snapshot of such a wave pattern.

Example 1

If we tried to visualize sound or light waves from a point source in space, what would an instantaneous pattern be?

Answer: The pattern would be a set of concentric *spheres*. The shortest distance between adjacent spheres is one wavelength.

To simulate the interference pattern caused by two point sources emitting waves at the same frequency and amplitude, let's make a transparent photocopy of the set of concentric circles shown in Figure 10-1. Then let's place the copy on top of Figure 10-1, move it around, and observe the results. A typical pattern is shown in Figure 10-2.

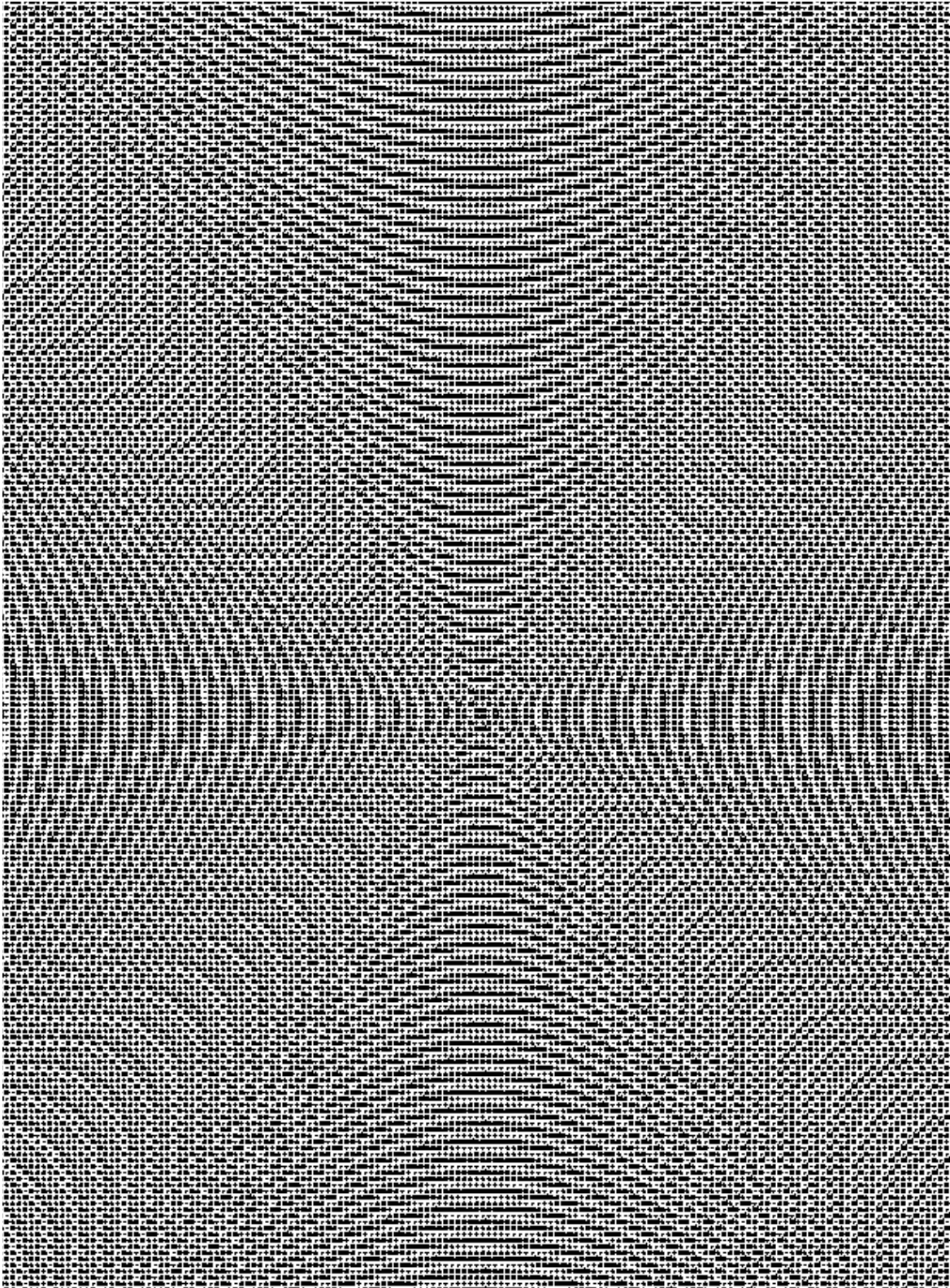


Figure 10-1 A two-dimensional “snapshot” of wave fronts from a constant-frequency point source at the center. The radial distance from one line to the next is one wavelength (λ).

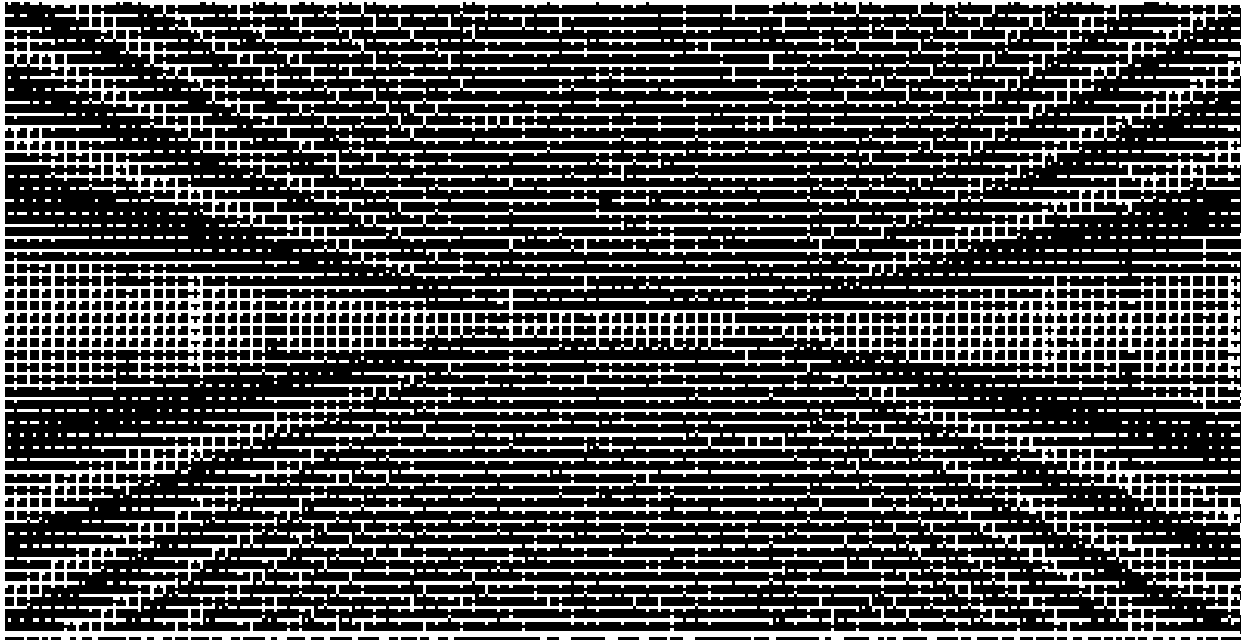


Figure 10-2 The two-dimensional interference pattern caused by two point sources of waves of constant frequency. Each set of waves is moving away from its source at the same constant speed. Nevertheless, the overall interference pattern remains constant in time. The two point sources creating this pattern can be seen near the center, along a horizontal direction, about four centimeters apart.

The *bright* (white) areas represent *constructive interference* because the crests from both sources—as well as the troughs from both sources—coincide, causing the waves to go up and down with twice the amplitude of each wave alone. The *dark* (black) areas represent *destructive interference* because the crest of one wave encounters the trough of the other wave, thus causing a cancellation of wave amplitude at that point. For water waves, the center of each dark area represents perfect and *permanent calm* in spite of the fact that waves from the two sources are passing through the dark area at all times. For sound waves, the same areas would represent regions of absolute silence.

A Trace of the Maxima from Two Point Sources

Figure 10-3 represents a trace of the constructive interference maxima—the white regions—observed in Figure 10-2. Here, S and S' denote the locations of the centers of the two sources. At precisely the midpoint between S and S' is a straight line OO' . At any point along this line, waves arriving from the two synchronized sources meet exactly in phase (or have zero phase difference), since they have traveled the same distance. This is called the *zeroth order* of constructive interference. For *all* points on the first curved line PP' (a hyperbola) at the right of the zeroth-order line, waves from S' travel a distance exactly one wavelength more than waves from S . Thus, this line represents the location of the *first order* of constructive interference. Similarly, the first curved line at the left of the zeroth order is also a first-order constructive interference pattern.

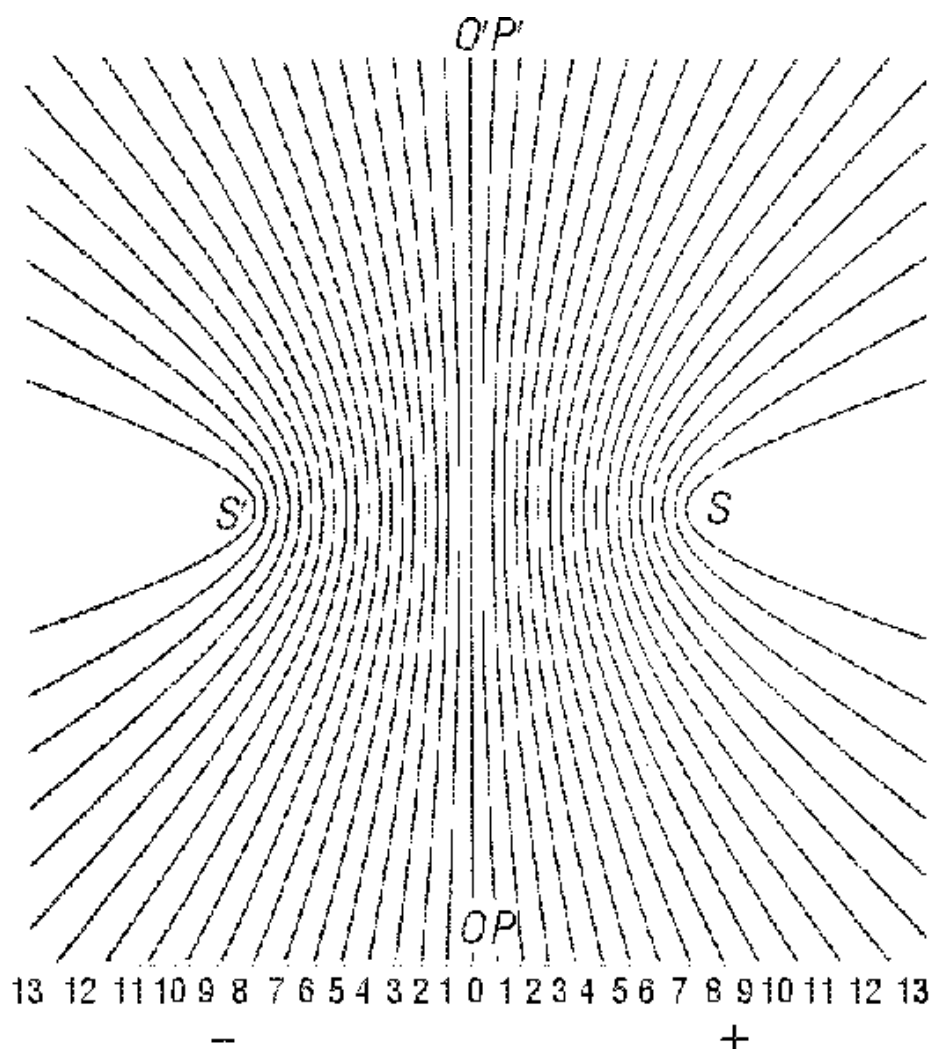


Figure 10-3 A computer trace of the locations of interference maxima on a plane containing the two point sources S and S'

Thus, the tenth curves at the left and right of the zeroth order are *tenth orders* of constructive interference. Any point on these curves has a difference in distance from S and S' equal to 10λ .

Exactly halfway between consecutive orders of interference maxima are hyperbolas (not drawn in Figure 10-3), which represent the *minima*, where the wave amplitude is always zero in spite of the fact that waves from two sources of disturbance are continuously passing through. In other words, waves meeting at any point along any line in Figure 10-3 are *in phase*. And waves meeting at any point along lines halfway between the constructive interference lines are *out of phase* and result in zero amplitude.

Now, suppose we imagine the interference pattern from S and S' as it exists in space, that is, in *three dimensions*. Figure 10-4 represents a cross section of this three-dimensional interference pattern. If the pattern were to spin around the x -axis, one would observe a set of *hyperboloidal surfaces*. The zeroth order ($m = 0$) is a flat plane, and all other orders ($m = 1$ and higher) are

smooth surfaces of varying curvatures. On the x - y and x - z planes are interference patterns like those shown in Figure 10-3, a set of hyperbolas. On the y - z plane, as shown in Figure 10-4, the pattern is a set of concentric circles.

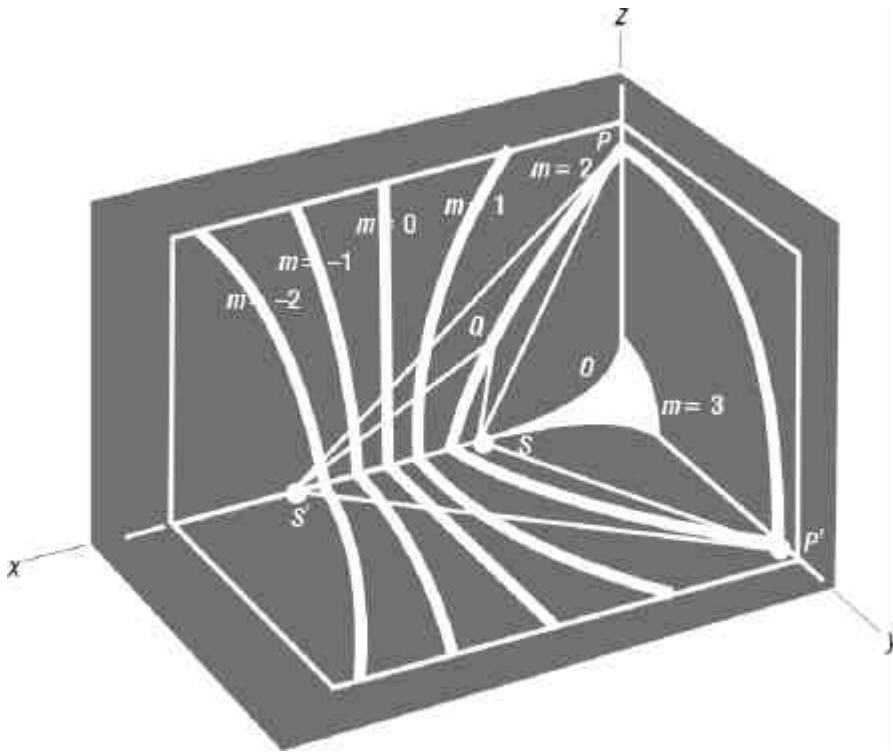


Figure 10-4 A three-dimensional interference pattern of waves from two point sources S and S' . The constructive interference order numbers, m , are indicated for the first several maxima.

Using the transparency of circles of Figure 10-1 that we made earlier, and placing it again on top of the original set (Figure 10-1), we can demonstrate how uniquely different interference patterns are formed—corresponding to each unique location of S versus S' . Observing the pattern along the axial direction (line through S and S') reveals the concept of *Michelson interferometry*. Far above and below S and S' , *Young's double-slit interference pattern* is recreated. Changing the distance between S and S' shows how the pattern changes correspondingly. For example, as the separation between S and S' increases, the interference fringes become more dense, i.e., more maxima and minima per unit distance. This is measured in terms of *spatial frequency*, or cycles per millimeter. Conversely, as the distance between S and S' decreases, the spatial frequency of the interference pattern decreases—less cycles per millimeter.

Example 2

Verify the statement in the above paragraph by moving the *transparency of circles* around on top of the page with circles (Figure 10-1).

The Physical Model

Some interesting characteristics of *hyperboloids* are represented in Figure 10-5. Think of the separate hyperboloids as the three-dimensional surface traced out when Figure 10-3 is rotated about an axis through the points SS' . Imagine that all the hyperboloidal surfaces are mirrors. Take the zeroth-order "mirror" OO' , which perpendicularly bisects the line SS' joining the two sources. In three dimensions, this is a flat mirror. Each ray from point S , striking the hyperboloidal surface (mirrors) at $m = 2$, $m = 1$, $m = 0$, $m = -1$, and $m = -2$, as shown, reflects from these surfaces (mirrors) in a direction such that the reflected ray appears to come from point S' . Two such rays from S are shown in Figure 10-5, one up and to the left (labeled a), the other down and to the left (labeled b). The reflected rays are labeled a_0 , a_1 , a_2 , a_{-1} , a_{-2} , and b_2 and b_1 , in accordance with the appropriate ray from S and the hyperboloid from which they reflect. Thus, S' is the *virtual image* of S , for any and all of the surfaces. The reverse is true for light from S' incident on the surfaces, for which S would then be the virtual image.

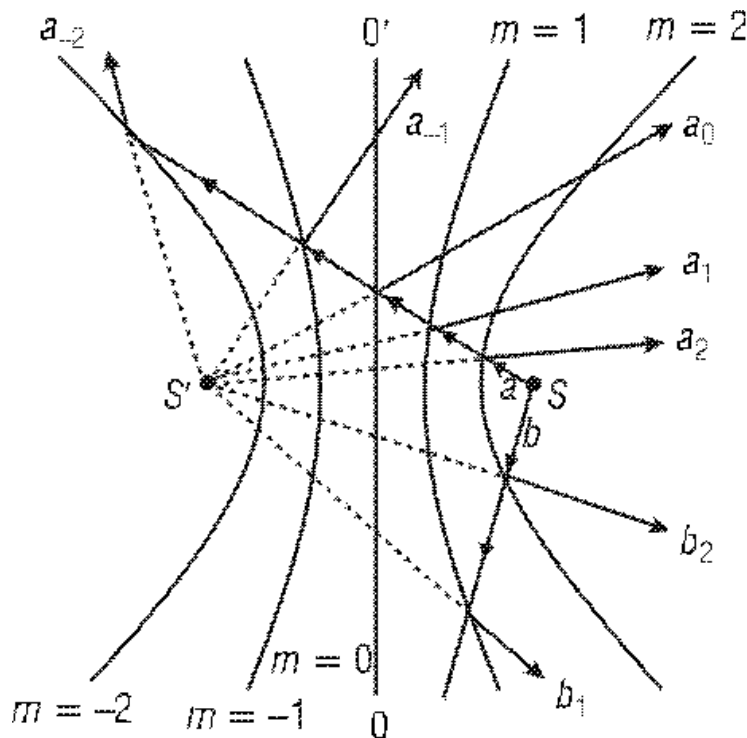


Figure 10-5 Light from S is reflected by any part of any hyperboloidal surface (mirror) in a direction such that it appears to be originating from S' .

With this physical model in mind, we are now ready to explain all the important characteristics of holograms recorded in a medium such as photographic emulsion that has a thickness of about 6 to 7 micrometers (μm).

Applications of the Model

Creating the Virtual Image

Figure 10-6 shows the optical case of two-beam interference in three-dimensional space.

Assume that the light from the two sources (S and S') is directed at a recording medium such as a silver halide photographic emulsion (“holoplate”) at a position as shown. The flat rectangle in the figure represents the top edge of the holoplate. Since the typical thickness of these emulsions is in the vicinity of 6 or 7 μm and the wavelength of laser light used to record holograms is 0.633 μm (HeNe laser), this thickness is approximately 10λ . *The interference pattern recorded inside the emulsion represents sections of hyperboloidal surfaces of many different orders of m .* These are, of course, sections of the hyperboloidal surfaces that we have been describing.

Observe carefully the orientation of the “mirrors” formed inside the emulsion. The “mirrors” on the left side lean toward the right, those on the right side lean toward the left, and those in the center are perpendicular to the plane of the holograms. In precise terms, *the plane of each “mirror” bisects the angle formed between rays from S and S' .*

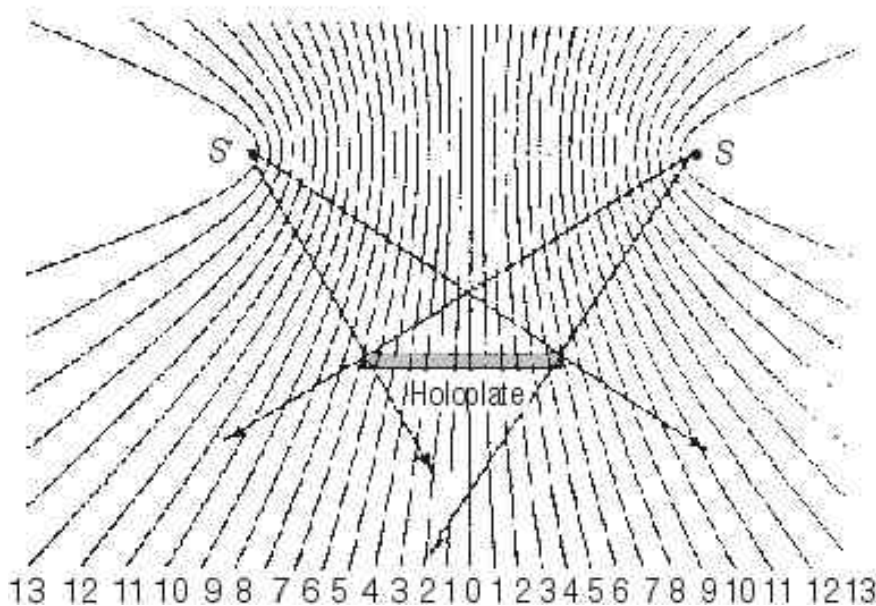


Figure 10-6 Light from S interferes with light from S' and produces a three-dimensional interference pattern inside a “thick” medium such as photographic emulsion (the “holoplate”).

The exposed and developed emulsion (holoplate) is called a *hologram*. Within the hologram, the recorded silver surfaces are partially reflecting—as well as partially transmitting and absorbing. If we replace the hologram in its original position during the recording, take away S' , and illuminate the hologram with S alone, as shown in Figure 10-7, all the reflected rays will appear to originate from S' . An observer would see these reflected rays as if they all came from S' . In other words, *the virtual image of S' has been created.*

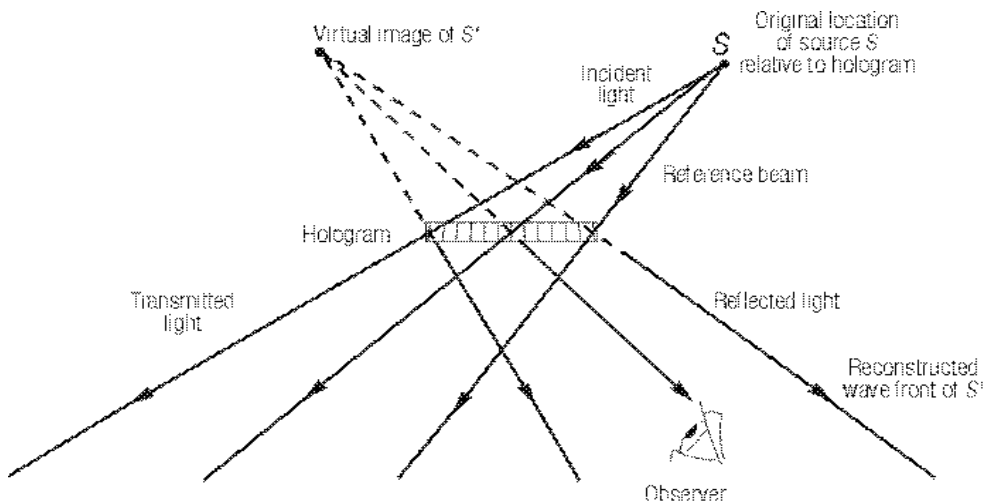


Figure 10-7 When the developed emulsion (hologram) is illuminated by S alone, the virtual image of S' is observed.

We can arbitrarily call the light from source S a *reference beam* and from source S' an *object beam*. If more than one point source is located in the vicinity of S' , each source will form a unique hyperboloid set with source S and the film will record all of them. When the processed hologram is illuminated, with source S only, each set will reflect light in such a way as to recreate the virtual image of all its object points.

If we replace point source S' with a three-dimensional scene (or object) illuminated by light having the same constant frequency as the reference beam, each point on its surface (S_1' and S_2' , for example) creates a unique hyperboloid set of patterns with S inside the emulsion. Thus, we have a hologram of a three-dimensional object (Figure 10-8). When the hologram is illuminated by S , each set of hyperboloidal mirrors recreates a virtual image of each point (S_1' , S_2' , etc.), so that a complete, three-dimensional virtual image of the object is reconstructed.

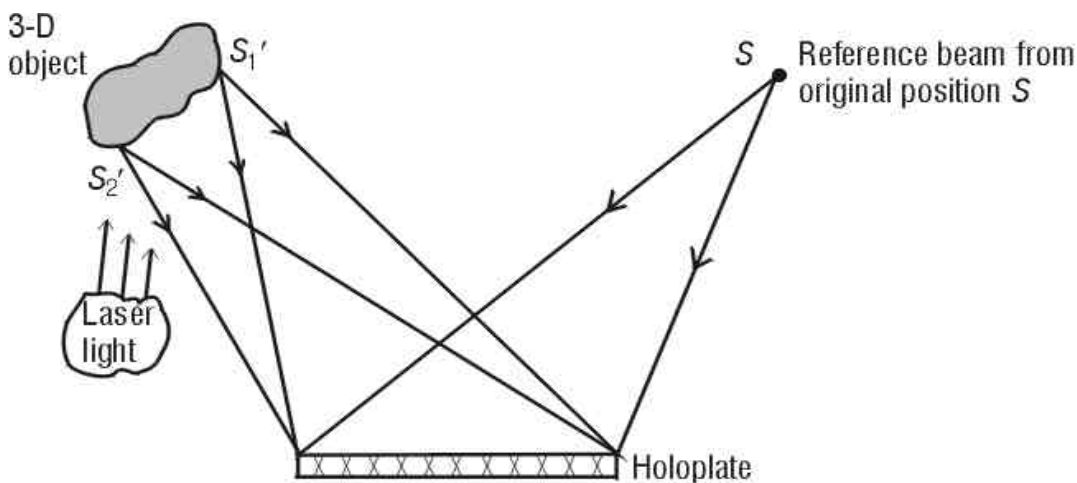


Figure 10-8 A hologram of a three-dimensional object can be considered as a superposition of many individual holograms of points on an object.

A general statement of the model can now be given as follows:

Imagine all hyperboloidal surfaces that represent the interference maxima due to two interfering sources to be partially reflecting surfaces. When a hologram is made, the volume throughout the emulsion records a sum of a multitude of hyperboloidal sets of partial mirrors, each set being created by the interference between the reference beam (S) and light from each point on the object ($S_1', S_2', \dots, S_n'$). When the hologram is viewed by illuminating it with S , each mirror set reflects light and forms a virtual image of each object point, thus recreating the wave front of the original object.

Creating a Real Image

Take the hologram from Figure 10-7 and illuminate it in a “backward direction” by focusing a beam of light back toward S (Figure 10-9). The reflected light from our hyperboloidal mirrors will focus at S' so that, if a projection screen were present there, we would have a *real image* of the original 3-D object. This can be done also with the hologram formed in Figure 10-8. The real image in this case will appear on the screen as a two-dimensional image of our original object. Depending on the location of the screen, different slices of this scene will come into focus.

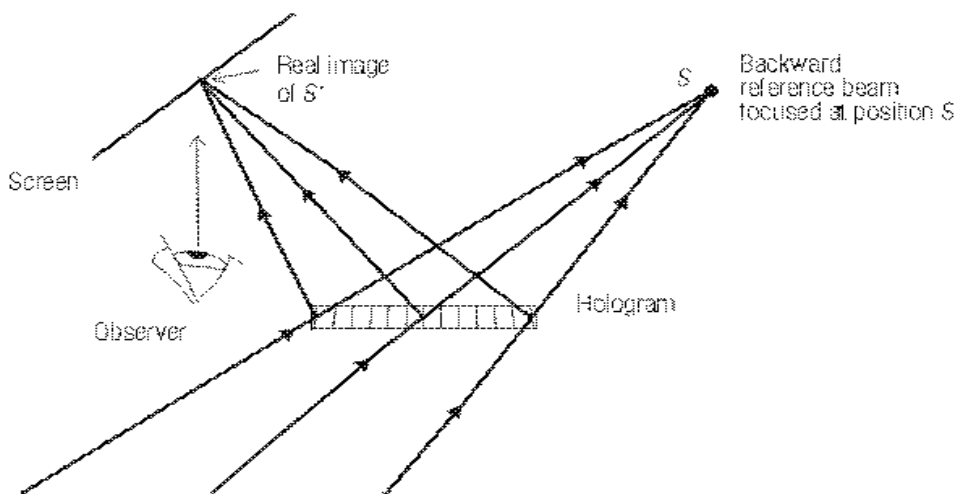


Figure 10-9 If a hologram is illuminated from the “front,” exactly backward toward S , the hologram will reflect the light behind the hologram to form a real image of S' on a screen.

Redundancy

If a transmission hologram is broken into pieces, each piece will give a complete perspective of the original scene. This can be understood from Figure 10-10. Imagine that the holoplate were half or a small fraction of its original size shown earlier in Figure 10-8. Since every elementary volume in the hologram was formed with light from a complete perspective of the scene, each of these elementary volumes will produce a complete perspective. In other words, the size of the holoplate used to make a hologram is independent of the size of the scene. *A large hologram can be considered as the sum of many smaller holograms.*

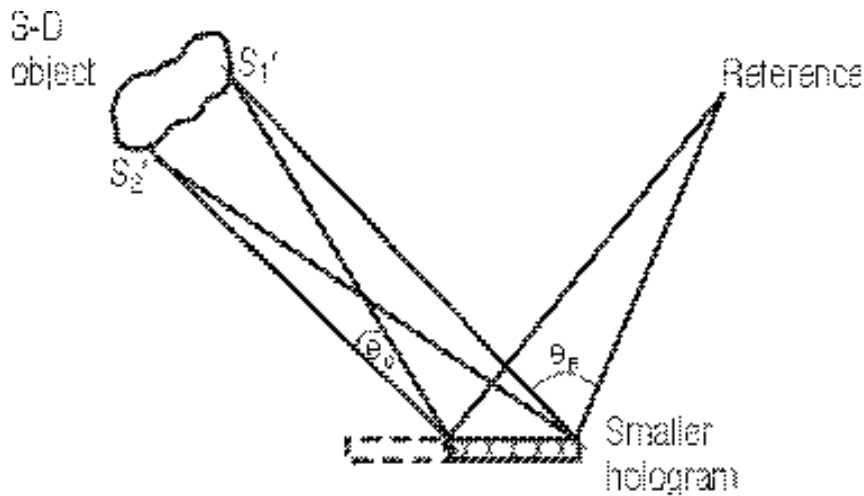


Figure 10-10 Each piece of a hologram can recreate a complete view of the object. A large hologram can be considered as a collection of many smaller holograms.

For the purpose of projecting a real image on a screen with a laser beam, it is desirable to select only a narrow area by using an undiverged beam so that the area covered does not exceed a few millimeters in diameter. In this case, the real image consists of rays at small angles relative to one another. This increases the *depth of field*, allowing us to have a *focused* image over a long distance along the beam paths that form the real image. Many laws of geometric optics operate here, i.e., *aperture*, *depth of field*, and *resolution*.

Dynamic Range

Not only the locations of all points on an object are reproduced in the hologram, but their relative intensities as well. Consider two spots S_1' and S_2' on the object, with S_1' having the same intensity as S and S_2' being less intense than S (Figure 10-11). In the areas where the light waves from S are in phase with those from S_1' , the total amplitude is doubled. Where they are out of phase, the total amplitude is zero. The same cannot occur between waves from S and S_2' . The result is that the interference pattern formed between S and S_1' has a higher contrast than those formed between S and S_2' , and results in having “mirrors” with higher reflectivity. In other words, S_1' simply makes a brighter hologram than S_2' .

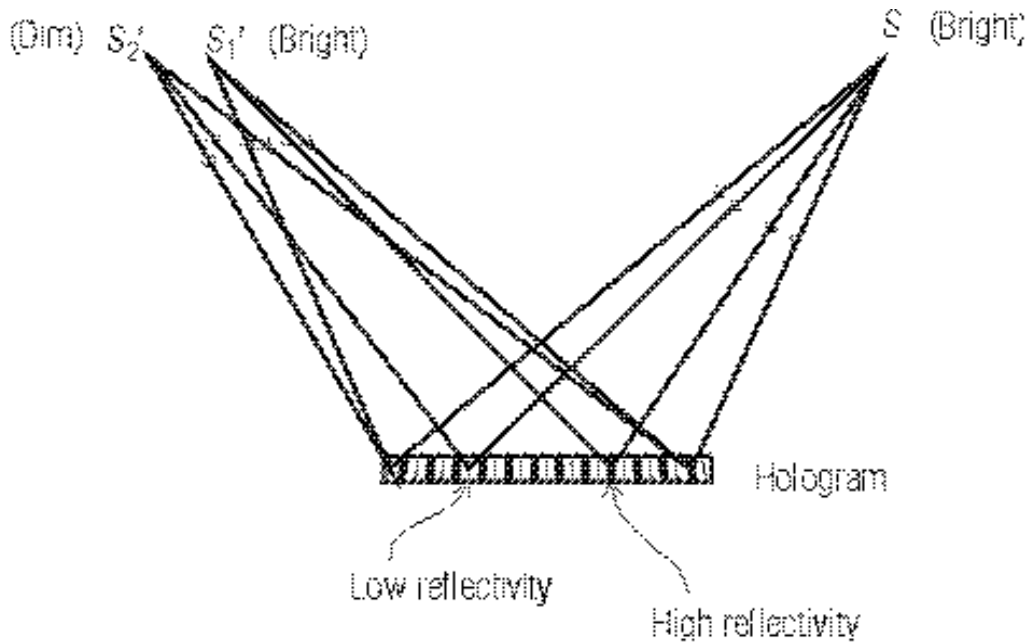


Figure 10-11 A bright spot S_1' creates a brighter hologram with S than does a dim spot S_2' .

Thus, a hologram can recreate images with intensity levels varying from almost complete darkness to glares. The holographic image of a diamond, for example, can be seen to “sparkle.” The hologram is said to have a large *dynamic range*.

By contrast, a photograph (or a TV picture) cannot reproduce glares or sparkles. The brightest part of such an image is the diffused white color of paper (or screen).

“Noise”

As is true in all information-transmission systems (radio, TV, recording a picture, etc.), the output always has *background noise* added to it. Similarly, a copy of a photograph is never as high in quality as the original.

Besides the so-called grain “noise” of photographic emulsion, which is caused by the scattering of light by the silver grains, another source is called *intermodulation noise*. In Figure 10-11, in addition to the interference patterns formed between S and S_1' , and S and S_2' , there is a pattern formed between S_1' and S_2' (not shown). The latter pattern forms a set of hyperboloids throughout the volume of the emulsion but of much lower spatial frequency (fewer lines per millimeter across the surface of the film due to the close proximity between S_1' and S_2'). These extra “mirrors” direct light in arbitrary directions when the hologram is illuminated with S alone, and create a background of “fog.”

When the scene consists of a three-dimensional object, each pair of points on the object creates a set of *unwanted interference patterns*. With a small object located far from the hologram, the range of angles between pairs of points (S' and S'') on the object is small and the noise problem is not severe. In the extreme case where the object is simply a point source S' , there is no intermodulation noise at all.

On the other hand, a large object situated near the hologram can create severe noise problems because every pair of points on it creates a set of unwanted interference patterns. Referring to Figure 10-8, light from S_1' and S_2' will interfere throughout the emulsion. The larger the object and the closer it is to the hologram, the higher the spatial frequency of its interference patterns (i.e., more mirrors created), along with the increase in the severity of the noise problem.

Beam Ratio

To help minimize the effects of intermodulation noise, practical holography requires that the reference beam be of higher intensity than light from any point on the object. In practice, the intensity ratio—as measured (at the location of the holoplate) by using a diffuser in front of a light meter, between the reference and object beams—varies from 1:1 to 10:1 for *transmission holograms*, the type so far under discussion. This allows the “mirrors” to form due to light from the reference beam and points on the object to be generally of higher reflectivity than those formed due to light from any pair of points on the object. Also, the noise can be further minimized by having the smallest angle between the reference beam and any object beam (θ_R in Figure 10-10) be greater than the largest angle formed by a pair of points on the object (θ_O in Figure 10-10). This ensures that the minimum spatial frequency formed by the object and the reference beam is greater than the spatial frequency of the noise. When the hologram is illuminated, the intermodulation noise is diverted to angles smaller than the signal. In this way, even though we cannot eliminate the noise, we can isolate it. In engineering terms, we have spatially isolated the *frequency domains* of the signal from the noise. In doing so, we have increased the signal-to-noise ratio, i.e., increased the quality of the hologram.

Multiple Scenes

One of the most dramatic features of a hologram is that it is capable of recording more than one independent scene in the same volume of space and then displaying them independently by changing the angle between the hologram and the reference beam. This phenomenon was discovered by a father and son team (William Henry and William Lawrence Bragg) through the study of X-ray diffraction by crystals. They were awarded the 1915 Nobel Prize in physics.

To understand what follows, it is necessary to review the subject of *thin-film interference* (Module 1-4, *Basic Physical Optics*) and be able to explain what causes the beautiful colors when diffused white light is reflected by soap bubbles.

Figure 10-12 shows a realistic cross-sectional view of the interference pattern formed in a microscopically small piece of hologram created using the configuration shown in Figure 10-5. (For simplicity, Snell refraction [Module 1-3, *Basic Geometrical Optics*] has been ignored by assuming that the hologram is sandwiched between material with the same index of refraction as the emulsion.) The hyperboloidal surfaces inside the thin emulsion layer approach flat planes and are perpendicular to the film surface, like venetian blinds in the “open” position. Because of the large angle between rays from S and S' (about 90°), the separation between “mirrors” is about λ , much smaller than the emulsion thickness of 10λ . Here, λ represents the wavelength of light *inside the medium*.

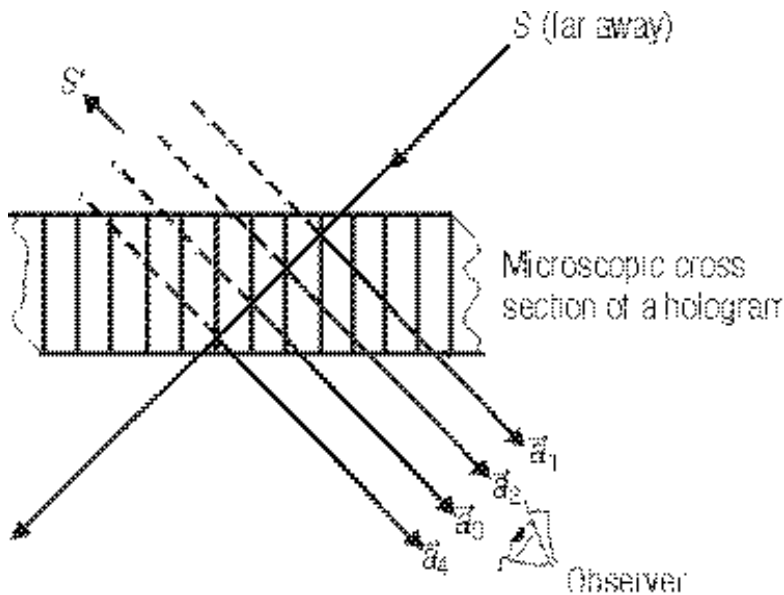


Figure 10-12 A microscopic cross section of a hologram. The thickness of the hologram far exceeds the distance between interference orders, and light from S undergoes multiple reflections.

When the hologram is viewed with S alone, the light undergoes multiple reflection by successive planes when it penetrates the film. However, because of the inherent characteristics of this family of hyperboloid “mirrors” (recall Figure 10-5), each successive reflection will have a precise phase shift of 2π because the optical path is increased by precisely the distance of one wavelength. All the reflected waves are precisely in phase and, therefore, add in amplitude, resulting in a strongest possible wave front representing the object beam. If the angle of incidence of S is deviated from the original reference beam relative to the hologram, all the multiply reflected beams will have phase differences other than 2π and the resultant reflected wave will be much lower in intensity, even zero.

Quantitatively, we can state that, in the case of correct illumination, the absolute values of all the amplitudes— $a_1, a_2, \dots, a_n$, of successive reflections—add directly and the intensity is $I = (|a_1| + |a_2| + \dots + |a_n|)^2$. In the case of misaligned illumination, the phase shift of each successive reflection is different from 2π and $I' = (\bar{a}_1 + \bar{a}_2 + \dots + \bar{a}_n)^2$, so that $I > I'$ always.

In practice, when the illuminating angle is significantly different from the correct angle, no image can be seen.

This phenomenon—*Bragg diffraction*—is of great historical significance. Using a beam of X rays of known wavelength, directing it at a crystal such as rock salt, and studying the angles of maximum diffraction, the distance between the atomic layers was measured. In turn, using crystals with known atomic spacing, X-ray wavelengths were accurately measured. This allowed scientists to study the inner workings of the atom.

In modern photonics, this principle is used to design multilayer thin films on mirror surfaces such as those used in laser cavities, to make optical fibers that selectively transmit and reflect chosen wavelengths, and to fabricate exotic crystal for massive holographic memories.

To create multiple-scene holograms (see the laboratory for making two-channel transmission holograms), one exposes the film with object O_1 , stops the exposure, changes to a second object

O_2 , changes the angle between the reference beam and the film, and exposes a second time. Generally, each exposure time is one-half that of the exposure for a single-scene hologram, assuming no great change in object brightness. The resultant pattern recorded in the processed film is the equivalence of the superposition of two independent sets of hyperboloids, each corresponding to a given scene. During reconstruction, depending on the orientation of the hologram with respect to the reference beam, the wave front of one *or* the other scene can be recreated. This is true for both the real and the virtual images.

Depending on the thickness of the emulsion, the sizes of O_1 and O_2 , and their proximity to the film, different degrees of success can be achieved in recording multiple images with minimum “cross-talk” (image overlap).

It is interesting to compare *time-domain* versus *space-domain* in information systems. A long song takes a long *time* to sing; a long written story requires a thick book (large *volume*) to record.

Example

Suppose a 1.0-mm-thick CD can store 1.0 gigabyte of information in the form of digital data. All these data are stored in the top 1.0-micrometer-thick layer. How much information can this CD store if it can record over its entire volume at the same information density?

Answer: 1.0 terabyte

The CD is 1000 micrometers thick and it uses only the 1.0-micrometer top layer to store the 1.0-gigabyte information. The remaining 999 layers can store an additional 999 gigabytes, leading to a total of 1000 gigabytes or 1 terabyte.

White-Light Reflection Holograms

Gabriel Lippmann received the 1908 Nobel Prize in physics for his discovery of recording color photographs on a “black-and-white” silver halide emulsion. Today, we use a similar “Lippmann emulsion” for recording holograms.

Figure 10-13 shows a photographic camera in which the lens focuses a colorful object on the recording emulsion several micrometers in thickness. The back of the emulsion is in physical contact with a pool of mercury, which serves as a mirror and reflects back all light incident on it.

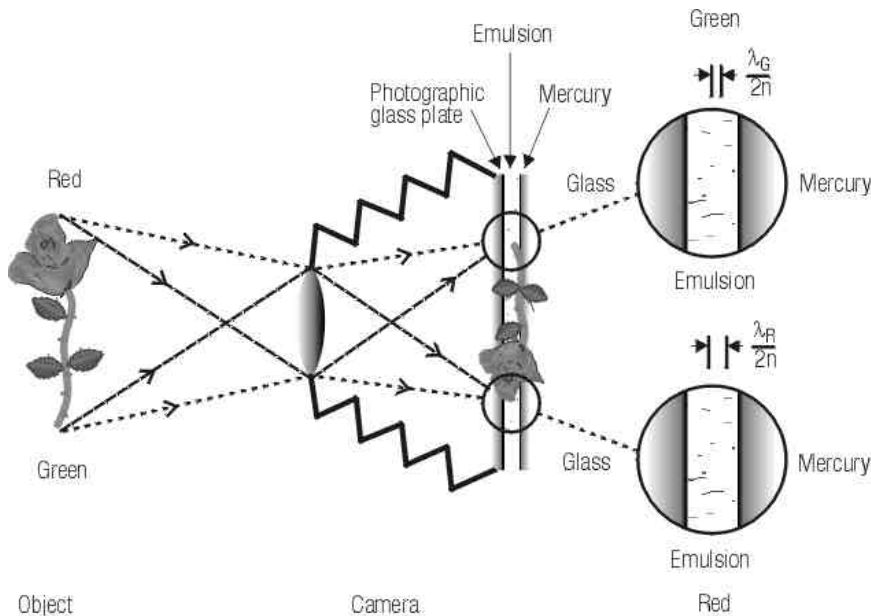


Figure 10-13 Lippmann photographs are recorded in a camera in which mercury is used to reflect the light back through the emulsion to create a standing wave for each color of light.

Suppose that the image of the petal of a red rose falls in a particular region of the emulsion. The red light traverses the emulsion and is reflected back through it. The incident red light interferes with the reflected red light and forms a *standing wave* in which the distance between adjacent maxima (or minima) is $\frac{1}{2}\lambda$. In another area of the emulsion, a green leaf is focused there, forming an image. The green light similarly forms a standing wave pattern throughout the emulsion. Because green light has a shorter wavelength than red light, its interference pattern is of a higher spatial frequency. Another, still higher-frequency, pattern may be formed elsewhere where a patch of the blue sky is focused.

After the photograph is developed, layers of silver remain in all locations where the maxima of all the interference patterns are located. When illuminated with a patch of white light, each area of the photograph reflects—by constructive interference—a color corresponding to the original focused image. The result is a full-color photograph recorded in a colorless emulsion.

With the availability of highly coherent laser light and an improved Lippmann emulsion, we can use the same concept to record three-dimensional scenes in color.

Figure 10-14 shows again the overall interference pattern between point sources of coherent lights S and S' . Note that the region between S and S' is occupied by *standing waves*, where the distance between adjacent maxima (or minima) is $\frac{1}{2}\lambda$. If we place a 10λ -thick emulsion midway between the sources (where MN and PQ are surfaces at the opposite sides of the emulsion) and record the interference pattern, we will have 20 hyperboloidal surfaces throughout the hologram. When illuminated by S alone, a virtual image of S' is created—and vice versa—as discussed earlier.

Recall that the *interference patterns* recorded in a *transmission hologram* consist of slivers of hyperboloidal “mirrors” (Figure 10-6) that are *perpendicular* to the plane of the emulsion, like venetian blinds in the “open” position. In a reflection hologram (Figure 10-14), the “mirrors” are

parallel to the plane of the emulsion. Furthermore, the latter has twenty mirrors—for a 10λ -thick emulsion, as explained above—whereas the other has only four (Figure 10-12) when each is being illuminated by its reference beam.

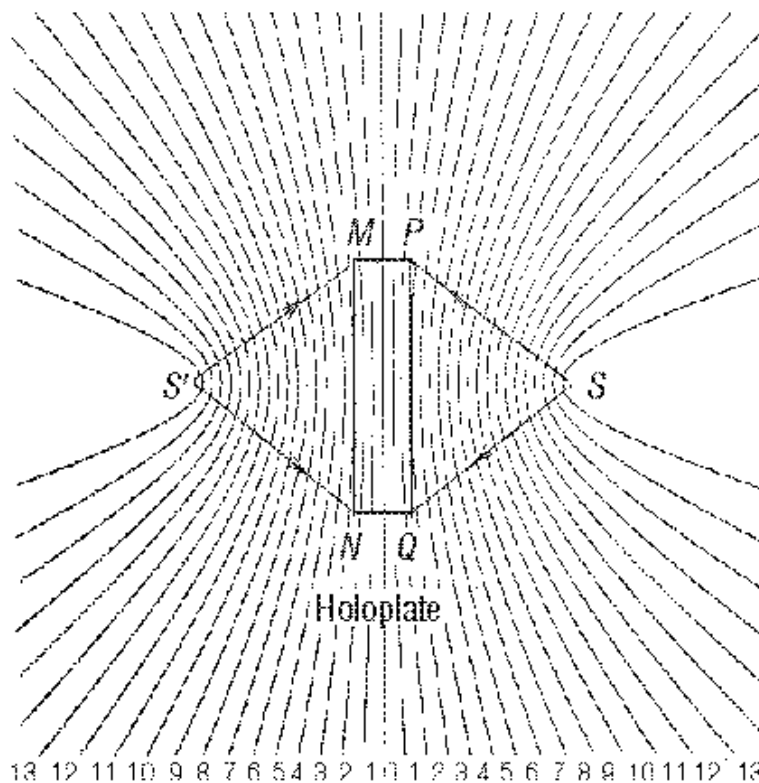


Figure 10-14 Reflection holograms are recorded in the region between the reference and object beams where standing waves are created. MN and PQ are surfaces on opposite sides of the emulsion. (Only a portion of the “mirror” surfaces is shown between planes MN and PQ.)

The result is that the reflection hologram can be viewed with incandescent light located at S, due to the fact that the additional sixteen “mirrors” strongly select the same wavelength of the laser that recorded them.

Example

A reflection hologram is made with red laser light. However, when illuminated with white light, the image appears yellow, or even green. Why?

Answer: The emulsion shrinks when the hologram is developed and dried. This causes the spacing between the hyperbolic surfaces to be decreased. Thus, a shorter wavelength is being reconstructed. Placing the hologram on top of a cup of hot coffee will swell the emulsion, and the color of the image can be tuned back to red momentarily.

Figure 10-15 shows the simplest method of making holograms: A solid object is placed in contact with the holoplate, with the emulsion side facing the object. Light spreading from a

diode laser (located at point S) is allowed to expose the emulsion. Light that passes through the holoplate illuminates the object. Each point on the object scatters light back through the emulsion and interferes with the direct light from S , creating a set of unique hyperboloidal “mirrors” throughout the volume of the hologram.

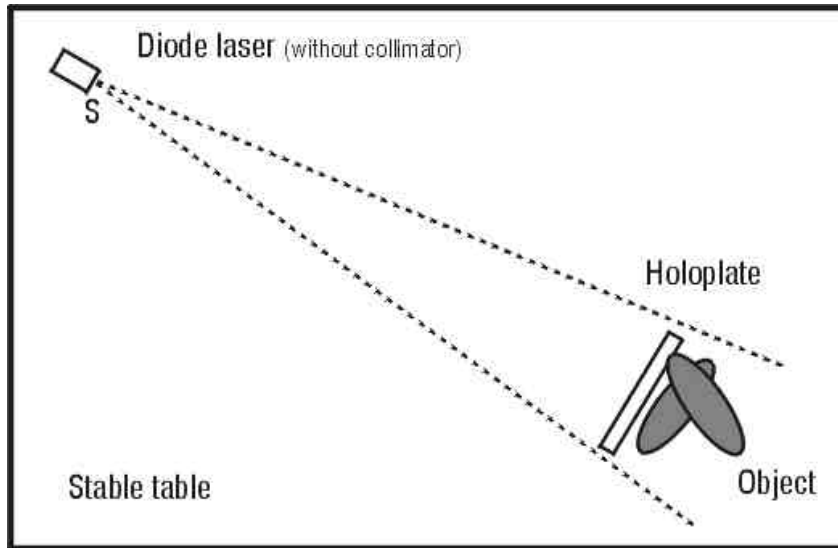


Figure 10-15 Simplest method of making a hologram: The spreading light from a diode laser interferes with the light scattered back by the object and forms standing waves throughout the volume of the recording medium.

When the developed hologram is illuminated by a point source of incandescent light, each set of hyperboloidal “mirrors” reconstructs a virtual image of a point on the object. Together, all the sets of “mirrors” reconstruct the entire three-dimensional image in a single color.

To make full-color holograms, light from three lasers (red, green, and blue) is combined into one single beam and the hologram is recorded in the same way as above. However, a special development process is required to ensure that there is no shrinkage in the emulsion.

At this point, let’s compare the main differences between transmission and reflection holograms: *transmission holograms are made with both reference and object beams on the same side*; the interference patterns recorded are like venetian blinds and are more or less perpendicular to the surface of the hologram. On the other hand, *reflection holograms are made with the reference beam on the opposite side of the object beam*. Its hyperboloidal fringes are more or less parallel to the surface of the hologram.

Since the entire object is on the same side of any hologram, any pair of points on it creates unwanted *intermodulation noise* by forming a transmission type of interference pattern. Therefore, when a reflection hologram is illuminated, the intermodulation noise is transmitted to the side away from the observer, while the image is reflected toward the observer. *Thus, in making reflection holograms, the best beam ratio is one-to-one* because we can ignore the intermodulation noise.

Holographic Interferometry

Consider a “**double exposure**” as represented by Figure 10-16a. Here the object is located at S_1' during the first exposure. It is then moved to S_2' , a distance of the order of a few λ 's, for the second exposure. The resultant hologram is a superposition of two sets of hyperboloids that coincide on the left side of the hologram but fall in between one another on the right side.

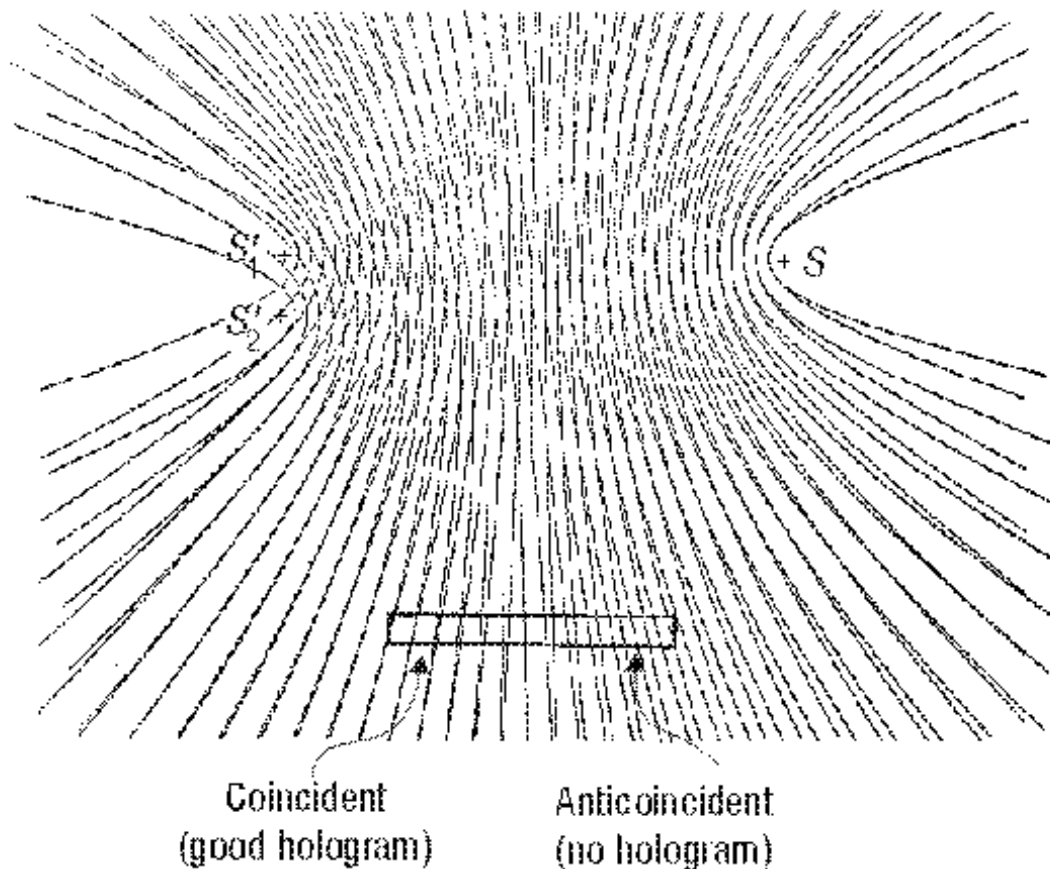


Figure 10-16a A double-exposure hologram of S' in two locations. The two interference patterns are coincident at the left of the hologram but are anticoincident at the right side.

When the virtual image of the hologram is reconstructed by S , a bright point located in the vicinity of the original object is seen from the left side of the hologram. As the viewer moves toward the right side of the hologram, the object S' disappears! The reason is that, at the right side, the two superposed interference patterns cancel each other when the maxima of one set fall in the positions of the minima of the other set.

If the object point had been displaced differently (from S_1' to S_3' , as shown in Figure 10-16b), its interference pattern would be shifted in such a way that the two patterns are anticoincident on the left side of the hologram but coincident on the right side, and the situation then would be reversed. (Note: A comparison of Figures 10-16a and 10-16b will not disclose the small difference between positions S_1' , S_2' in Figure 10-16a and S_1' , S_3' in Figure 10-16b. Nevertheless, the two sets of positions are sufficiently different to interchange the coincident and anticoincident characteristics of the holoplate.)

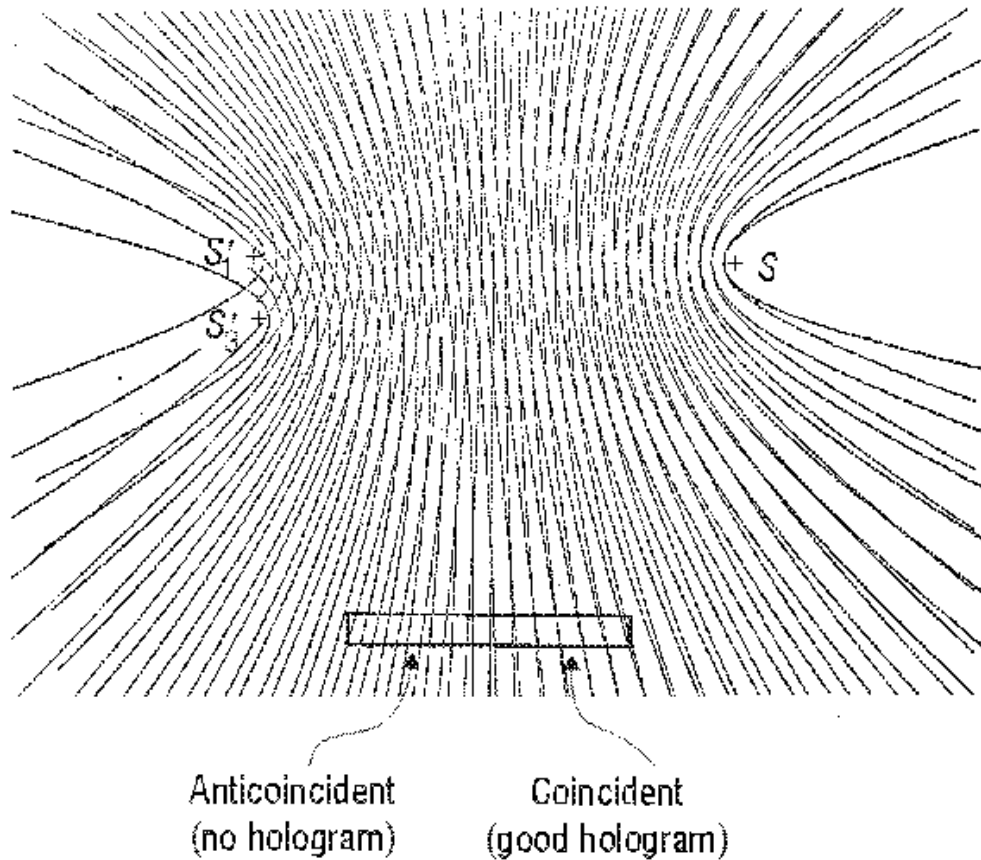


Figure 10-16b A double-exposure hologram of S' in two different locations. The two interference patterns are anticoincident at the left of the hologram but coincident at the right side.

Therefore, when a rigid object is slightly moved (displaced) between the two exposures, a set of straight black lines (fringes) will be superimposed on the image of the object. Each point on any given fringe will displace the same distance. On the other hand, if the object is deformed (such as a “live” plant having grown between exposures), complex curved fringes will be observed.

All points on a given fringe represent locations with the same displacement.

“**Time-averaged**” holographic interferometry is a single-exposure hologram of an object under normal mode vibration. The antinodal areas are moving, so the interference patterns are “smeared out.” The nodal areas have not moved; thus they appear bright. This renders vibration modes directly visible. (This technique has been used in the study of musical instruments.)

“**Real-time**” holographic interferometry involves making a hologram, processing it, relocating it in the position where it was made, and observing the virtual image superposed on the real object. If the object is moved or deformed (such as a growing object), the wave front of the image interferes with light from the real object, and an interference pattern can be observed in real time.

Holographic interferometry encompasses an entire discipline in mechanical engineering and is applied widely to solve myriad technical problems, mostly in the area called *nondestructive testing*.

Coherence Length (temporal coherence)

One of the basic techniques in making good holograms is to ensure that the optical path of the object and the reference beams are equalized well within the *coherence length* of the laser light. Beginning from the output of the laser, trace the total distance traveled by the reference beam to the hologate. Now do the same for the object beam. The difference in the two distances must not exceed the *coherence length*. So, what is this coherence length?

In general, lasers operating without internal etalons emit simultaneously more than one frequency (see Module 1-5, *Lasers*). For example, the low-cost helium-neon lasers emit typically two to three different frequencies simultaneously.

When a hologram is recorded by using a two-frequency laser, one can think of it as the sum of two holograms being made simultaneously, one at each frequency. The higher-frequency interference pattern is caused by light with a shorter wavelength, thus it has more interference orders over any given space. Figure 10-17 is the superposition (summation) of two independent sets of interference patterns, one caused by S_1 and S_1' of one frequency and the other caused by S_2 and S_2' of a slightly higher frequency.

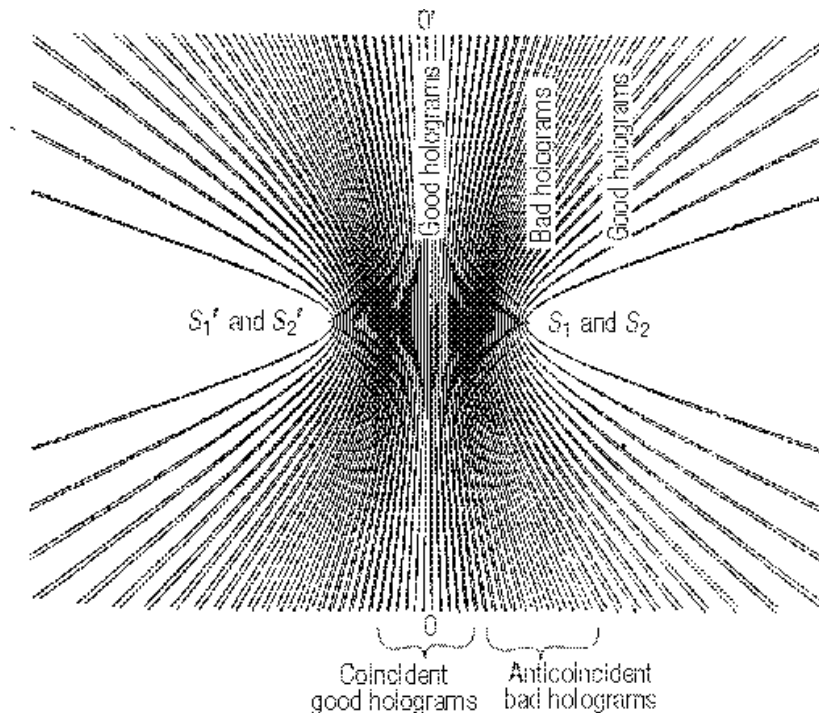


Figure 10-17 A superposition of two independent sets of interference patterns, each caused by light of slightly different wavelength from the other. The smudge-like area is caused by the anticoincidence of the two patterns; no hologram can be recorded there.

Near the zeroth order OO' in Figure 10-17 (the flat plane that is midway between the two interference point sources, the two sets of interference patterns coincide. Near this region, light of *any* color can produce distinct interference patterns. Lippmann photography is based on this fact. If a hologate is located in this area, a good hologram can be made. On the other hand, notice the two dark, smudge-like areas, one on each side of the zeroth order. This is caused by

the fact that the maxima of one pattern are located at or near the minima of the other pattern. If the holoplate is located in this region, the hologram will be very dim or not exist at all.

The *coherence length* of this two-frequency laser can be defined verbally as *the difference in distances travel by light from points S_1, S_2 and S_1', S_2' to the center of either dark area*. For a typical HeNe laser, this distance is about 30 cm.

Notice (Figure 10-17) that, if we increase the difference in the optical paths, the two patterns become coincident again.

For lasers with internal etalons or for highly stabilized diode lasers, the output has only one frequency and the coherence length is many meters. When using such lasers for making holograms, one does not need to equalize optical paths between reference and object beams.

Thin holograms

As is true with all “models,” the physical model for holography must break down at a limit (like the stick-and-ball model for molecules). Our limit is the “thin” hologram.

A hologram is considered “thin” when the separation between successive hyperboloidal surfaces exceeds the thickness of the recording medium. At this point, the model breaks down because there no longer is a set of hyperboloidal surfaces to be recorded. For example, embossed holograms (baseball cards and VISA/MC cards) are made by stamping a set of two-dimensional interference patterns on a sheet of plastic. These holograms cannot reproduce true color because there is no Bragg diffraction.

Then how does one explain the theory of a thin hologram?

A thin hologram is made by recording a two-dimensional interference pattern between a reference beam and light from the object on the surface of a recording medium (Figure 10-18). If the location of the holoplate is sufficiently far from S and S' , the two-dimensional interference pattern formed at the surface of the holoplate consists of *Young's double-slit* interference patterns (Module 1-4, *Basic Physical Optics*). The pattern consists of a set of straight dark and bright parallel lines whose spatial frequency depends on the distance of S from S' (slit separation) and the distance the hologram is from S and S' (screen distance).

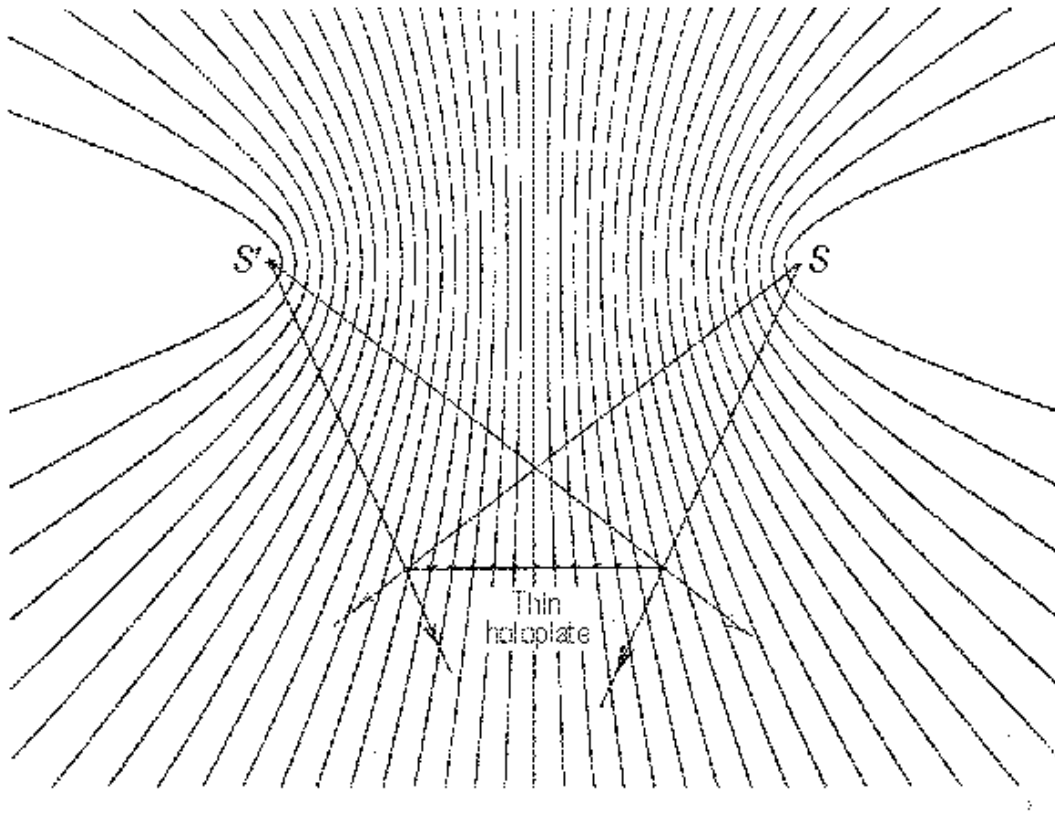


Figure 10-18 A thin hologram is made by recording a two-dimensional interference pattern on the surface of a recording medium. The result is a diffraction grating.

We learned previously (Module 1-4, *Basic Physical Optics*) also that a set of parallel dark and transparent lines is called a *diffraction grating*. When we are making a thin hologram, light from each point (S') on the object interferes with S and forms a two-dimensional diffraction-grating-like pattern on each small area of the hologram. Therefore, *each small area of a thin hologram is a summation of diffraction gratings formed by the interference of light from each point on the object with the reference beam*. The branch of mathematics used for quantitative study of holography is called *Fourier transformation*, and the formation of a hologram is a process called *Fourier synthesis*.

When the hologram is illuminated by S , each grating on each elementary area diffracts light to recreate the wave fronts of an object point (a process of *Fourier analysis*). If white light is used to illuminate this hologram, each different wavelength will be diffracted to a different angle and the image is smeared into a continuous spectrum. Thus, the transmission holograms discussed previously are best viewed with laser light.

In reality, most holograms are “quasi-thick” and “quasi-thin,” requiring a combination of physical explanations.

Laboratory Practices

Introduction

In this section, laboratory exercises will be arranged so that we start with simple procedures using simple equipment. The low cost of equipment makes it possible for many of the following exercises to be performed as “homework.”

All the experiments require the use of a laser and chemicals for processing the holograms. Section I, “Diode laser holography,” uses a class IIIa diode laser without the collimating lens, rendering it useless as a “pointer.” This allows the work to be performed by students at home. When higher-power lasers are used, in section II, “Advanced holography,” supervision by a qualified instructor is required.

Always observe all safety rules concerning lasers and chemicals.

Generally, the developer and bleach solutions are mixed by the instructor. Since the photochemistry of holography is an ongoing research, it will change as improvements are found. Thus, no particular regime is discussed here. Use according to the detailed instructions that accompany each processing kit provided by the manufacturer.

Similarly, the holographic plate or film used will change with time. Use them in accordance with the instructions of the supplier. The exposure time and the appropriate processing scheme are also provided.

Equipment and facilities

Holograms are made in darkened areas free from drafts, vibration, and noise. Because of the relatively low sensitivity of the recording material, sufficient light is allowed so that one can see comfortably after dark adaptation. To achieve this, use a 25-watt green light bulb in a lamp. Place the lamp under the table, cover it with aluminum foil to adjust the light, and direct it toward the floor. Do not allow direct light to shine on the holography system or on the developing station.

If the room has windows, cover them with black plastic sheets. Enough light can leak through to allow minimum vision after dark adaptation. In case of doubt, leave a holographic plate on a table and expose it to the ambient light for ten minutes. Develop it. If it turns dark, there is too much light.

Flowing tap water is desirable but not necessary. A large tray of clean water can be used to rinse the developed hologram. White trays are desirable because they allow continual inspection. An alternative is to use glass trays resting on white paper.

Make sure all fire codes are observed.

I. Diode laser holography

Certain class IIIa diode lasers sold as “pointers” are found to have high-frequency stability and thus long coherence length after an initial warm-up of a few minutes. With the collimating lens removed, the laser light spreads directly from the laser with a highly eccentric elliptical profile. Since the beam does not encounter other optical elements, it is completely “clean.” This allows it to be used to make many types of holograms without additional optical components.

These experiments are to be performed on top of a sturdy lab table or kitchen counter or on the floor. Support a thick 50-cm $\times$ 100-cm wooden board (or optical table) on top of four “lazy balls” (rubber balls that don’t bounce). Put washers under each ball so that they will not roll.

A. Reflection hologram

Equipment requirement: Darkened room with green safe-light, sturdy table or counter, optical table supported by “lazy balls,” mounted diode laser system, object on platform with three-point support, shutter, processing trays with chemicals, and holographic plates.

Figure 10-15—shown earlier—indicates the setup for making a “one-beam reflection hologram,” sometimes called a Lippmann (Nobel Prize in physics, 1908) or Denisyuk hologram.

Procedure

- A. Choose a solid object that looks bright when illuminated with laser light and whose size is not bigger than the hologram to be made. Mount (hot glue) it on a small platform made of wood or sheet metal (15 cm $\times$ 15 cm) with three round-head short screws underneath (to prevent rocking). Mount the laser on a stand about 25 cm high and direct the light down at 45° at the object, with the light spreading horizontally. The distance between the laser and the object is about 40 cm. Now turn on the safe light and turn off the room light.
- B. After the laser has been warmed up for at least five minutes, block the light from reaching the object using a self-standing black cardboard. (We will call this the shutter.)
- C. Lean a holoplate directly on the object, with the sticky side touching it. Wait at least 10 seconds.
- D. Lift the shutter, *but still blocking the light*, for 2 seconds, to allow any vibration to subside. Then lift the shutter away completely to allow the light to pass through the holoplate. The exposure is usually about 5 seconds. (Consult the instructions that accompany the plates.) Then block the light again.
- E. Develop the hologram according to instructions from the manufacturer.

After the hologram is dried, view it with a spot light such as a pen light, projector, or direct sunlight. Optional: Spray paint the sticky side (emulsion side) with a flat (or “antique”) black paint to provide a darker background and greatly improve the visibility of the image.

B. Transmission holograms

1. Without a mirror

Equipment requirement: Same as for the “reflection hologram” in section A. above. In addition, a stand-alone plate holder is needed. Make one exactly the same way as the object platform described above. Instead of the object, install two long (12 cm) screws on top with a separation less than the width of the holoplate to be used. Paint the screws a diffused black color.

Procedure

- A. Set up the system as shown in Figure 10-19. The diode laser is mounted 5 cm above the optical table with the beam spreading horizontally. One side of the beam illuminates the object or objects, and the other side serves as reference beam.

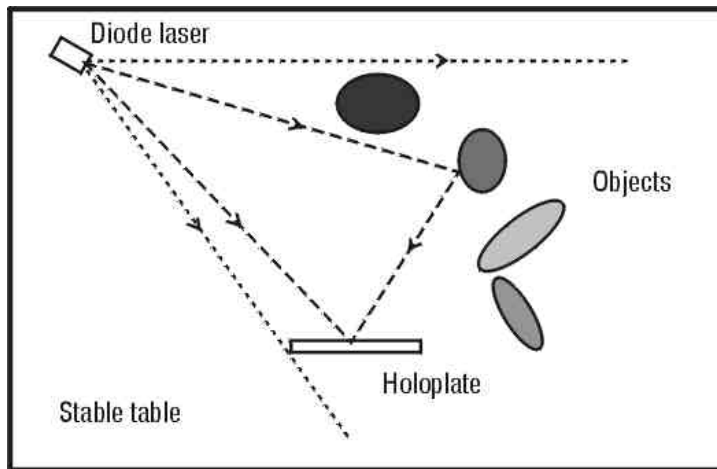


Figure 10-19 The simplest configuration for making a transmission hologram

- B. Block the beam with the shutter, turn off the room light, and, on the stand-alone plate holder, lean a holoplate vertically against the black screws with the sticky side facing the object(s). Wait 10 seconds.
- C. Lift the shutter and expose for about 30 seconds. Note: If there is a draft across your system, the long exposure time of 30 seconds requires you to put a large box over the entire system during the exposure.
- D. Develop and dry as before.
- E. This hologram must be viewed with laser light. To do so, lean the finished hologram back on the black screws the same way as during exposure. Cover or remove the objects and look through the hologram toward the location of the objects. A *virtual image* can be seen as if the object is still there.
- F. To observe the *real image*:
 - Relocate the finished hologram in the position where it was exposed.

- Remove the object and, in its place, position a vertical white screen (cardboard) facing the hologram.
- Darken the room and direct a collimated laser beam *through* the center of the hologram in a direction that is 180° from the original reference beam, i.e., back toward the location of the diode laser used for making the hologram.

All light paths are now reversed and a two-dimensional image is projected onto the screen. Move the laser beam to different locations of the hologram and observe the changing perspectives of the image.

2. Transmission hologram with one mirror

Note that the objects in the foregoing transmission hologram are always illuminated from the side. To get more frontal lighting, we need to add one flat, front-surface mirror.

Equipment requirement: Same as above, plus one front-surface mirror. Mount this mirror in a vertical position by hot-gluing the back side (the glass side) to a block of metal or wood, with the bottom edge raised to the same height as the bases for the object and the plate holder.

Procedure

The procedure is the same as for section B.1 (transmission holograms without a mirror), but using the setup as shown in Figure 10-20. Note that a baffle in the form of a black cardboard is needed to block the laser light from exposing the holoplate from the back side. To project the real image onto a screen, set the finished hologram back onto the plate holder and direct a laser pointer backward through the hologram, onto the mirror so that the beam hits the diode laser. On a screen located at the position of the object, the real image is formed.

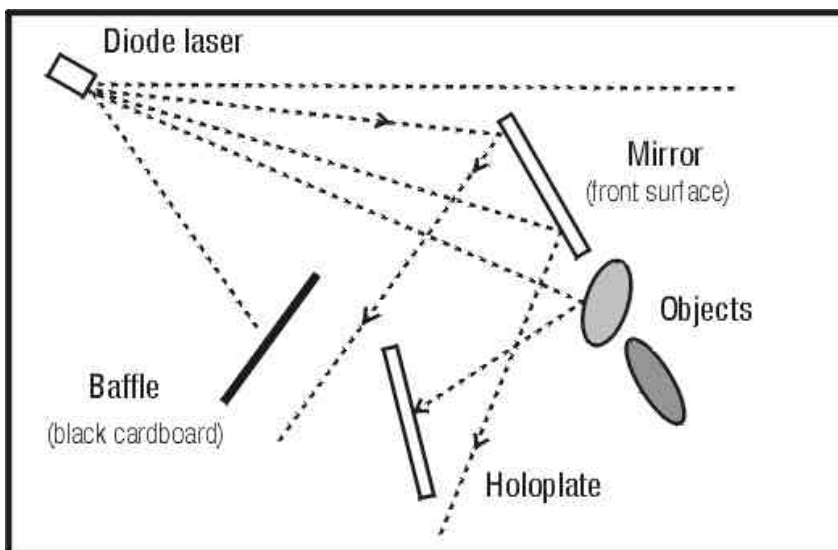


Figure 10-20 With the addition of one front-surface mirror, the hologram can be made facing the objects and the objects are better illuminated.

3. Two-channel transmission hologram

Use the same procedure as for the “transmission hologram with one mirror,” except now make the exposure time one-half as long. Pick up the hologate and rotate it upside down (with emulsion or sticky side still facing the object) and place it back on the plate holder. Replace Object 1 with Object 2. Expose again one-half the full time. Process as before.

In viewing the virtual or real image, notice that, depending on which side is up, a correspondingly different image is observed.

4. Diffraction grating

Set up the experiment according to the layout shown in Figure 10-21. The plane of the mirror should make a 90° angle with the plane of the hologate. Notice that no object is used in this experiment.

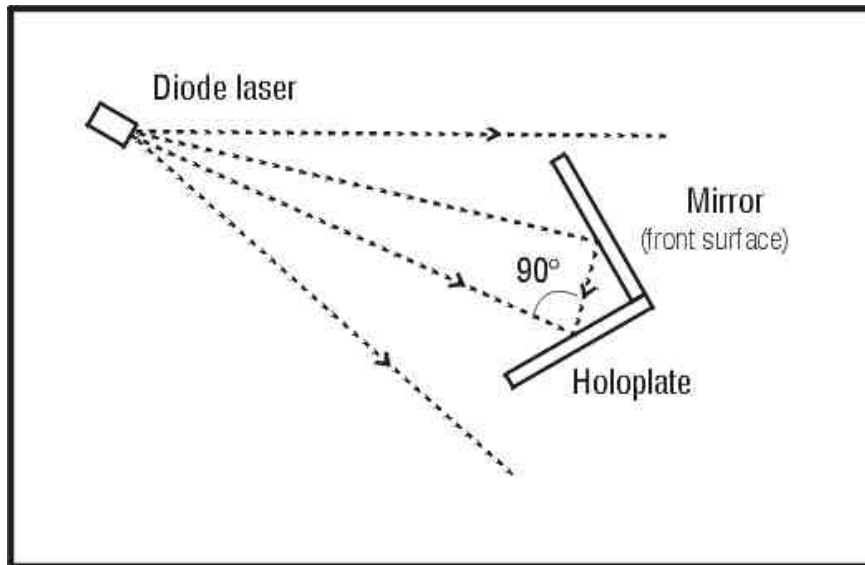


Figure 10-21 Making a hologram of light reflected from a mirror results in a high-dispersion diffraction grating.

Here the “object” is part of the beam from the laser that is redirected by the mirror. This configuration is the same as shown in Figure 10-6, in which both the reference and object beams are point sources.

Use the same procedure as before but with a shorter exposure time (approximately 75% as much) and process similarly.

Project the real image of the finished hologram onto a screen at the location of the mirror. Is there any difference between a real and a virtual image in this case?

If you see more than one spot, the extra ones are caused by internal reflections inside the hologate. To avoid these reflections, an anti-halation backing of the hologate is required. This is done by painting on the glass side of the hologate, before exposure, with a washable blue or black water-color paint. This absorbs the light as it arrives at the back surface during exposure.

This hologram is in effect a high-dispersion diffraction grating. Shine a beam of white light from a slide projector through the hologram and observe the spectrum on a screen. Use other sources such as mercury and sodium lights from the street during the night.

5. Additional project

Design a system using this holographic grating to observe the Fraunhofer lines from the sun. It will show which elements exist in the sun!

II. Advanced holography

The preceding experiments are fundamental. They use simple equipment to demonstrate the major principles of holography. However, they have many limitations. For example:

- The intensity ratio between the reference and object beam cannot be controlled. Thus the quality of the hologram cannot be optimized.
- The object is always illuminated with a single point source from a fixed direction, thereby casting shadows that cannot be controlled.
- The location of the image is always behind the plate.
- A laser is needed to observe the image formed by transmission holograms.

To make holograms with none of the above restrictions, the following **additional equipment** is needed:

- A higher-power laser that operates in a single transverse mode. It's even better if it operates in a single axial mode. A 10- to 30-milliwatt HeNe laser is recommended.
Observe all laser safety rules !!
- A power meter with a sensitivity range of 1 microwatt to 1 watt
- A large (4 feet $\times$ 8 feet is typical) isolation table with adjustable front-surface mirrors (4); lens with the largest aperture and shortest positive focal length (1); variable beam splitters (2); spatial filters with 10X objective and 25-micron pinholes (3); 4 $\times$ 5-inch ground glass plate (1); plate holder (1); and hardware for supporting all the above components
- *Alternative, low-cost system:* Build a sandbox (3 feet $\times$ 4 feet $\times$ 10 inches) with thick plywood and fill it with washed silica sand (the kind sold for sandblasting). Substitute spatial filters with small, double-concave lenses and concave mirrors. Mount (glue) all optical elements on top of 2.5 inch diameter PVC pipes (which allows x-y-z position adjustments when stuck into the sand). Use film instead of plates and sandwich it between large glass plates with strong paper clamps. (Just stick the sandwich into the sand.) Support the sandbox by a dozen "lazy balls" and put the system on a rolling table. The cost of this system is an order of magnitude less than the commercial system and is more mobile and versatile for the purpose of learning.

A. Split-beam transmission hologram

The following symbols are used in the diagrams for the remaining experiments:

	Front surface mirror		Negative Lens
	Beam splitter		Ground glass
	Positive Lens		Pinhole
	Laser		Hologram, or film

Note: A lens followed by a pinhole is a spatial filter.

Figure 10-22 is the layout for making a highest-quality transmission hologram. The procedure is general and is applicable to all of the remaining experiments.

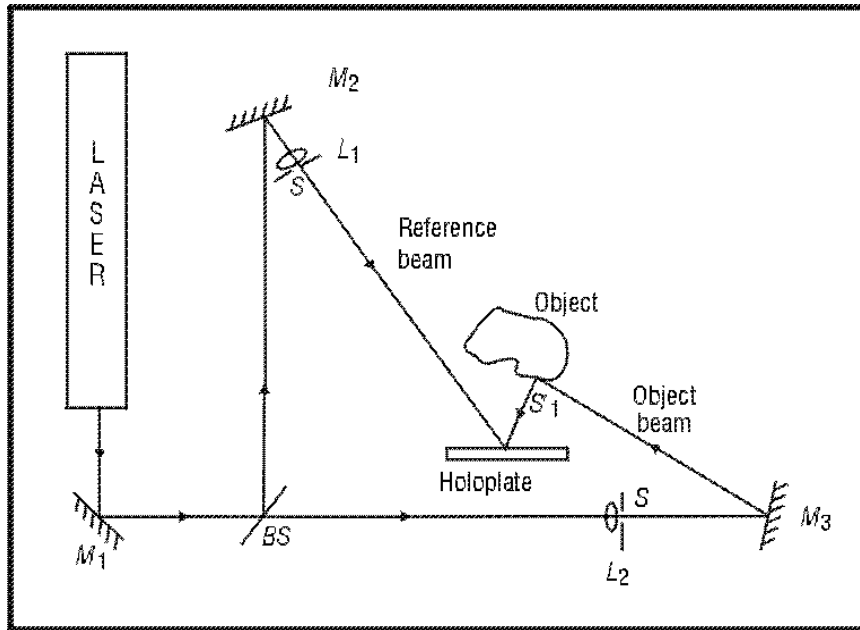


Figure 10-22 A general configuration for making the highest quality transmission holograms

- Lay out the components as shown in Figure 10-22 except for the two spatial filters L_1 , S and L_2 , S (or meniscus double-concave lenses).
- Equalize the beam paths by using a cord. Start from the beam splitter BS , and measure the total distance from BS to M_2 and on to H . Now measure the distance from BS to M_3 , on to the center of the object O and on to H . Equalize the paths within 2 or 3 cm by moving mirror M_2 .
- Expand the two beams by positioning the spatial filters (lenses) as shown. The reference beam should cover the holoplate (here substituted for by a white card during alignment). The object beam should light up the object.

- D. Baffling: From the position of the holoplate, carefully observe light scattered from anywhere other than from the object and the reference beam. Block all the unwanted light using black cardboard.
- E. Measure the intensity ratio of the two beams incident on the holoplate with the power meter by positioning the meter at the location of the holoplate. Block the light from the object beam and read the intensity of the reference; then block the reference beam and read the intensity of the light beam from the object. Adjust the intensity ratio of reference beam to object beam reading to about 4:1 to 8:1 by two methods: (1) use a variable beam splitter; (2) move the beam-expanding lenses appropriately along the beam paths.
- F. Determine the exposure time.
- G. Now expose and develop the hologram.

If “soft” lighting of the object is desired, a ground glass can be placed between the object and mirror M_3 . Carefully baffle any light from the ground glass directly to H .

B. Dual-object beam transmission hologram

Figure 10-23 shows a layout for illuminating the object with two beams. This allows artistic lighting through the addition of a beam splitter, spatial filter, and mirror. Soft lighting with ground glass(es) is also possible.

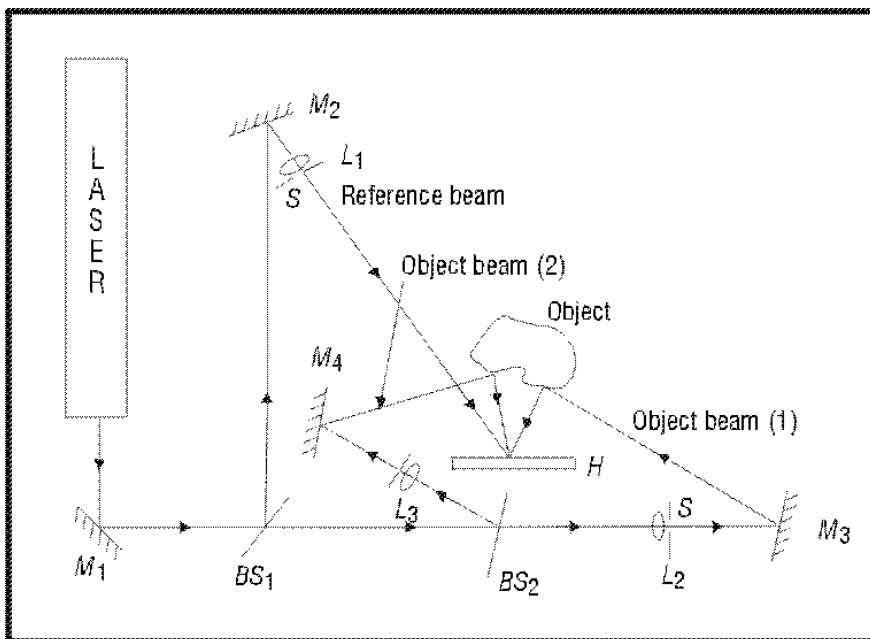


Figure 10-23 The object is illuminated from two independent beams, allowing more artistic lighting.

The beam path of the added beam (object beam 2) must be equalized, starting from the beam splitter BS_1 . Then proceed with instructions given in II.A. above.

C. Focused-image reflection hologram

Figure 10-24 shows a layout for making a reflection hologram in which the image appears in the plane of the hologram. Here the large lens with a short focal length is used to image the object onto the plane of the hologram. The object and the hologate are located at a distance equal to $2f$ on the opposite sides of the lens. This allows the image to be the same size as the object.

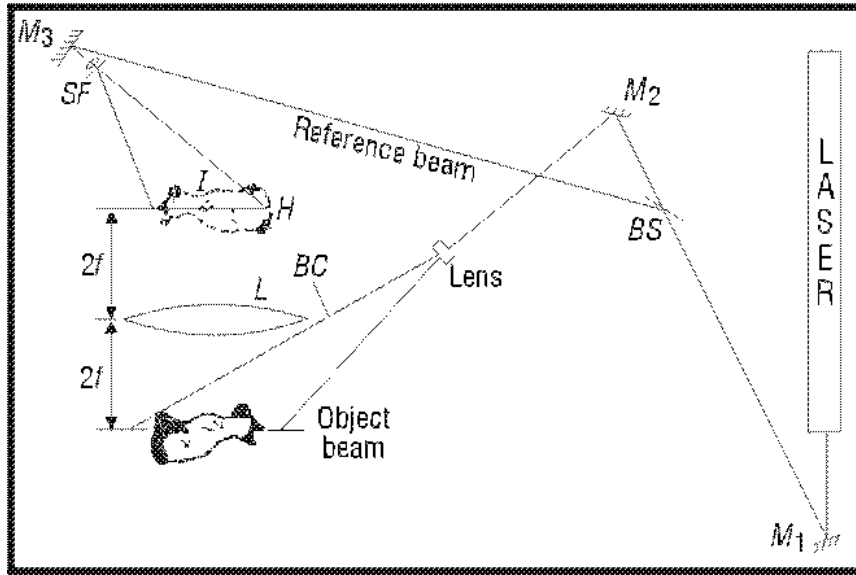


Figure 10-24 A configuration for making a focused image reflection hologram

During alignment, use a white card to find the real image of the object and place the plate holder there. As is true for making reflection holograms in general, a good beam ratio is 1:1.

Notice that in this experiment we are making a photograph and a hologram at the same time!

Problem Exercises/Questions

1. What is the most basic difference in the recording of reflection holograms and transmission holograms?
2. When a transmission hologram is broken into many pieces, each piece can reconstruct the image of an entire view of the object. Why?
3. When a reflection hologram is broken into many pieces, each piece cannot reconstruct the image of an entire view of the object. Why?
4. It is generally not possible to use commonly available photographic film to record holograms. Design a very particular hologram that can be made using these films.
5. What type of hologram, even if it is thinner than a wavelength of light, can record and reproduce a completely three-dimensional image? Explain.

6. Can holograms be made using one laser for the reference beam and another laser for the object beam? Why?
7. If you have to make a hologram using low-power diode lasers near a location that has trains going by all the time, what type of hologram should you try to make to optimize your chance of success?
8. Why can we use the sun to reconstruct the image of a reflection hologram but not of a transmission hologram?
9. Dennis Gabor discovered holography in 1947 before the laser was invented. How do you think he made the first hologram?
10. Explain how a reflection hologram selects the color from the light used to illuminate it.
11. You have made a reflection hologram using red laser light. Why is the image generally not red, but yellowish or greenish?
12. How is it possible to make a hologram with more than one scene that can be viewed one at a time through the full aperture of the hologram?
13. Why do we bleach the holograms after development?
14. Explain how a double-exposure hologram of an object that changes microscopically between exposures reveals the nature of the change.
15. When making holograms, we try to equalize the optical paths traveled by the reference and object beams. Why? Under what condition is this procedure not necessary?

Bibliography

Rossing, Thomas D., and Christopher J. Chiaverina. *Light Science, Physics and the Visual Arts*. New York: Springer-Verlag, Inc., 1999.

Saxby, Graham, *Practical Holography*, 3rd edition. Prentice-Hall, Inc., 1996.

Jeong, Tung H. *Laser Holography—Experiments You Can Do*. Lake Forest, Illinois, Thomas Alva Edison Foundation, Integraf, 1987.

Vest, Charles M. *Holographic Interferometry*. New York: John Wiley & Sons, 1979.

Bjelkhagen, Hans I. *Silver-Halide Recording Materials for Holography and Their Processing*, 2nd edition. New York: Springer-Verlag, 1996.

Pedrotti, Leno S., and Frank L. Pedrotti, *Introduction to Optics*, 2nd edition. Prentice-Hall, Inc., 1993.