# Chapter 14

# Optical Devices

Although there are a few examples of optical actuators, the vast majority of optical devices are sensors. In an optical sensor an optical signal (e.g. a light beam) is modified by the mechanical variable to be measured. Since we want an electrical interface to the sensor, the device must include means of producing and detecting the optical signal electrically. Figure 14.1 shows a block diagram of a typical optical sensor.

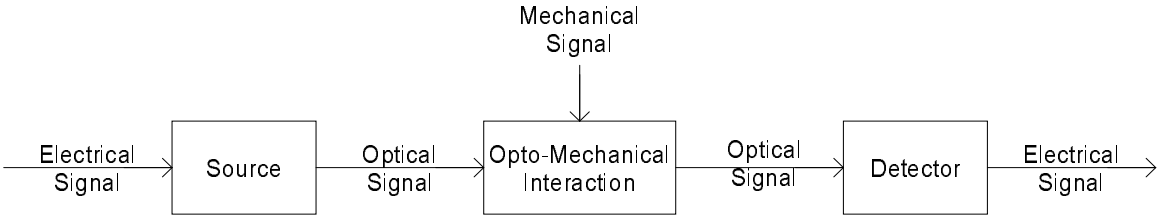



Figure 14.1: Signal flow in a typical optical sensor.

Figure 14.1 has the same basic structure as a non-self-generating electrical sensor, but there are additional conversions of the electrical signal to and from an optical signal. Although this double conversion appears to be an unnecessary complication, the low cost and ease of integration of semiconductor light emitters and detectors often make optical devices less expensive than equivalent non-optical ones. Other advantages of optical sensors include: EMI resistance, compactness, ease of access to moving parts, and the ability to sense without mechanical contact.

To build an optical sensor, we will need to assemble the following components:

- Optoelectronic devices (sources and detectors) to create the initial optical signal and to convert the modulated optical to an electrical signal.
- Optical "conductors" to carry the optical signal to and from its point of interaction with the mechanical signal being measured. These may consist of explicitly guided paths such as optical fibers, or free space propagation through the air.
- A means of modulating a measurable property of the optical signal (e.g. amplitude, phase, polarization) via an interaction with appropriate surface or volume elements of the mechanism. Possible interactions include reflection, refraction, or absorption.

## 14.1 Light

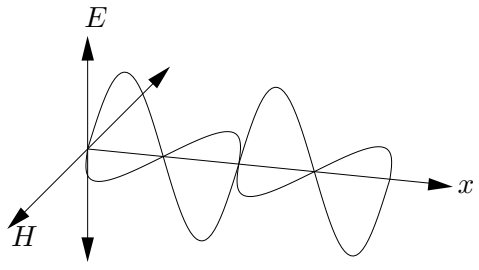
Light is electromagnetic radiation which occupies the region of the spectrum between microwaves and x-rays. More specifically, the wavelength of *visible light* ranges from 380 nm (deep blue) to 780 nm (deep red). This wavelength is sufficiently short that the quantum nature of light must be considered.

### 14.1.1 Waves

One of the most important results of Maxwell's work was the prediction of propagating electromagnetic waves. His equations showed that the velocity of propagation in a medium having permeability $\mu$ and permittivity $\epsilon$ would be $\frac{1}{\sqrt{\mu\epsilon}}$. Since $\frac{1}{\sqrt{\mu_0\epsilon_0}} = c$ (the speed of light in vacuum) the inescapable conclusion was that light was a form of electromagnetic radiation.

Electromagnetic waves are transverse waves with **E** and **H** orthogonal, so for a plane wave of wavelength $\lambda$ propagating in the $x$ direction with the $E$ field aligned with the $y$ axis:

$$\begin{aligned} E_y(x,t) &= E_0 \sin(kx - \omega t) \\ H_z(x,t) &= H_0 \sin(kx - \omega t) \end{aligned}$$




where $k = 2\pi/\lambda$ is the wave number and $\omega = 2\pi c/\lambda$ is the angular frequency.

If we connect the points of equal phase, we define a surface called a *wavefront* of the propagating wave. In free space, the wavefront of a point source is a sphere whose radius is expanding at the speed of light. If we consider the set of wavefronts defined by the crests (maxima) of the waves, a source of wavelength $\lambda$ will emit a series of these propagating spherical wavefronts separated by $\lambda$.

### 14.1.2 Photons

The *wave model* explains most of the observed behavior of light, including reflection, refraction, diffraction, polarization, and interference. However, in order to satisfactorily explain the emission and absorption of light, it is necessary to assume that its energy is carried in discrete *quanta* or *photons* rather than as a continuum. The energy of a photon is given by:

$$E = hf$$

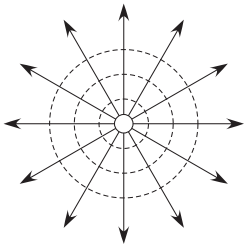
where $h$ is *Planck's constant* ($6.62 \times 10^{-34}$ Js) and $f$ is the frequency (in Hz). It is often convenient to describe photon energies in electron volts, in which case we would use $h = 4.14 \times 10^{-15}$ eVs.

In addition, to account for the photoelectric effect, these quanta must be spatially localized (i.e. particles). However, the *particle model* although consistent with most observed behavior of light, does not account for diffraction and interference. Thus we must accept the notion of a *wave-particle duality* where a quantum of light, or *photon*, may exhibit either particle-like or wave-like behavior.

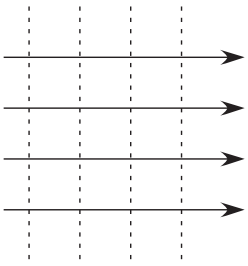
### 14.1.3 Rays

Whether modeling light as particles or waves, we can describe its path in terms of *rays*. We can think of a ray as either the path of a photon or as a line pointing in the direction of the advancing wavefront. In the latter case it will be perpendicular to the wavefront. A light ray originates from a source and terminates when absorbed.

For example, consider a point source of light of wavelength $\lambda$ (shown in cross section at the right). The dashed lines, separated by $\lambda$, represent the spherical wavefronts, and the arrows are the rays.



At a sufficiently large distance from the source, the radius of curvature of the wavefronts is large enough that they may be considered to be planes. The rays representing these *plane waves* will be parallel.



### 14.1.4 Energy Flow: Radiometry and Photometry

Plotting wavefronts or rays shows the *path* of light, but describing the *quantity* is most easily done in terms of power. We can determine the power in a light beam using either the wave or particle model.

The energy density of an electric field of magnitude $E$ is $w_e = \frac{1}{2}\epsilon E^2$. For a magnetic field of magnitude $H$, $w_m = \frac{1}{2}\mu H^2$. The total energy density is $w = w_e + w_m = \frac{1}{2}\epsilon E^2 = \frac{1}{2}\mu H^2$. In an electromagnetic wave in free space $\frac{E}{H} = \sqrt{\frac{\mu_0}{\epsilon_0}}$ so that $w_m = \frac{1}{2}\mu_0 H^2 = \frac{1}{2}\epsilon_0 E^2 = w_e$ and the total energy density is $w = w_e + w_m = 2w_e = 2w_m = \epsilon_0 E^2 = \mu_0 H^2$. Since the wave is propagating with velocity $c$, the power flow per unit area is

$$S = cw = c(w_e + w_m) = c(\frac{1}{2}\epsilon_0 E^2 + \frac{1}{2}\mu_0 H^2)$$

With $c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$ and $\frac{E}{H} = \sqrt{\frac{\mu_0}{\epsilon_0}}$ we can show $S = EH$ and this power flow is in the direction of propagation of the wave. This is compactly stated in vector form by the *Poynting vector*

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \tag{14.1}$$

which describes both the magnitude and direction of the flow of power in an EM wave.

Equation (14.1) gives the instantaneous power in terms of **E** and **H**. For sinusoidally varying $E$ and $H$ the average power density will be

$$\mathbf{S}_{av} = \frac{E_0 H_0}{2}$$

#### 14.1.4.1 Radiant Flux

We can get a similar and equivalent picture using photons: simply count the number (or rate) of photons passing through a surface and multiply by their energy. In either case we can define the *radiant flux* $\phi_e$ through a surface as the energy per unit time (i.e. power) crossing the surface. The subscript $e$ (for electromagnetic) is used to distinguish radiometric variables form the corresponding photometric variables, which we will meet a few paragraphs later.

#### 14.1.4.2 Intensity

Flux describes the total power from a source, but does not describe how it is distributed. A 100 W spotlight produces the same radiant flux as a 100 W bare bulb, but appears much "brighter" since it concentrates that flux into a narrower region of space. We can formalize this by defining the *radiant intensity*

$$I_e = \frac{d\phi_e}{d\Omega}$$

i.e. the flux per unit solid angle, measured from the source. Intensity has units of $\frac{\mathrm{W}}{\mathrm{sr}}$. Note that the radiant intensity of a source in a particular direction is independent of the distance at which it is measured.

Radiant intensity is a property of a source, and implicitly assumes its location is known. For a plane wave, the apparent source is at infinity, so a more meaningful measure of the "strength" of a parallel beam is the power per unit area of the wavefront, as given by Equation 14.1.4. This quantity is sometimes referred to as the *intensity* of the wave, a source of confusion which can be resolved by checking the units. In either case, proper integration of the intensity over a surface gives the total flux through the surface.

#### 14.1.4.3 Emittance

For a true point source, the power density at the source itself would be infinite. A real source will have some non-zero surface area, with the distribution of emitted power defined by the *radiant exitance* or *emittance*

$$M_e = \frac{d\phi_e}{dA}$$

i.e. the flux per unit area leaving an emitting surface in $\frac{\mathrm{W}}{\mathrm{m}^2}$.

#### 14.1.4.4 Irradiance

Intensity describes the distribution of power produced by a source. We also need a way of measuring the density of power incident upon a surface. To do this, we define the *radiant incidence* or *irradiance*

$$E_e = \frac{\phi_e}{dA}$$

in $\frac{\mathrm{W}}{\mathrm{m}^2}$ as the flux per unit area at a receiving surface.

#### 14.1.4.5 An Example

To illustrate some of these concepts, consider a 100 W incandescent light bulb at a distance of 6 ft from a sheet of paper, as shown in Figure 14.2. Assume that 20% of the input power is conducted away as heat through the base of the bulb, and the rest is radiated. I.e. the radiant flux is $\phi_e$=80 W. If we assume the lamp is an ideal point source, then the intensity will be $I_e = \frac{80}{4\pi} = 6.37\frac{\mathrm{W}}{\mathrm{sr}}$. (In fact, the base of the bulb will block some of the light and the shape of the filament will cause some directional non-uniformity, but it's still a fair approximation.)

At a distance of 6 ft the irradiance will be $E_e = \frac{80}{4\pi \cdot 6^2} = 0.177\frac{\mathrm{W}}{\mathrm{ft}^2} = 0.0164\frac{\mathrm{W}}{\mathrm{m}^2}$. Since the sheet of paper is flat, rather than spherical, there will be a slight variation of intensity over

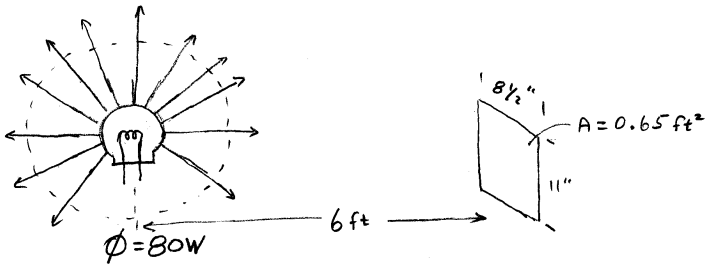



Figure 14.2: A Radiometry Example.

its surface. But within the limits set by our other approximations, we can say that the incident power received by the sheet is $P = I_e A$=1.0 mW.

#### 14.1.4.6 Photometry

Of the 1 mW that is illuminating the paper, only a small fraction of that is useful for reading what's on the paper. Most of the 80 watts radiated from the bulb is in the infrared and invisible to the human eye. To evaluate illumination in terms of human vision, we need to weight its spectrum according to the eye's response. This curve, shown in Figure 14.3, is known as the Spectral Luminous Efficiency Function.

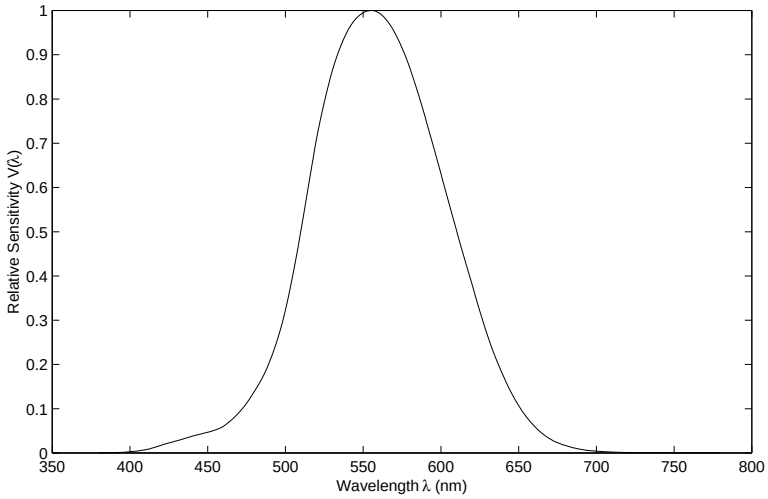



Figure 14.3: The Spectral Luminous Efficiency Function.

The concepts of flux, intensity, and irradiance are equally applicable to the human perception of light as they are to its physical measurement. The science of measuring light for human consumption is called *photometry* and has the same structure and vocabulary as radiometry, but with "radiant" replaced with "luminous" and the $e$ subscript replaced with

$v$ (for *visual*).

The weighted unit of flux (*luminous flux*, $\phi_v$) is called the *lumen* (lm) and corresponds to $\frac{1}{683}$ W at 555 nm. A typical 100 W incandescent bulb has an output of 1750 lumens. Luminous intensity $I_v$ is measured in $\frac{\text{lumens}}{\text{steradian}}$ and has units of *candelas* (cd). Luminous incidence $E_v$ is measured in $\frac{\text{lumens}}{\text{m}^2}$ and has units of *lux* (lx).

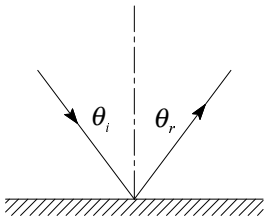
## 14.2 Optics

Optics is concerned with the interaction of light with *optical elements* such as mirrors and lenses. *Physical optics* (or *wave optics*) examines the passage of a wavefront through an optical system, determining its interaction with the optical elements based on their electromagnetic properties. Physical optics can describe reflection and refraction, and must be used to describe diffraction and interference, but is a computationally expensive procedure.

However, in situations where the effects of diffraction and interference are negligible, the simpler model of *geometric optics* is sufficient. In this model, light is presumed to travel in straight lines or *rays* directed radially from a source. The interaction of these rays with the reflecting and refracting surfaces of an optical system determines its behavior.

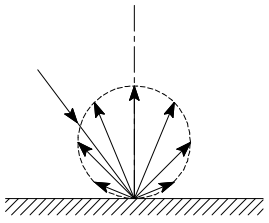
### 14.2.1 Reflection

**Specular Reflection** A light ray incident on a polished metal surface will be almost totally reflected, with the angle between the incident ray and the normal to the surface (the *angle of incidence*, $\theta_i$) equal to the angle between the normal and the reflected ray (the *angle of reflection*, $\theta_r$). The incident ray, the reflected ray, and the normal to the surface are all in the same plane.




**Diffuse Reflection** If the surface is rough, the incident ray will be *scattered* over a range of directions. For a perfectly diffuse or *Lambertian* surface, the radiant intensity varies with the cosine of the angle to the surface normal ($\theta$), i.e.

$$I(\theta) = I_0 \cos(\theta)$$

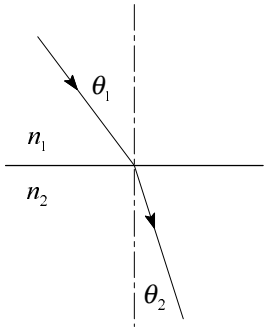


where $I_0$ is the intensity in the direction normal to the surface.

### 14.2.2 Refraction

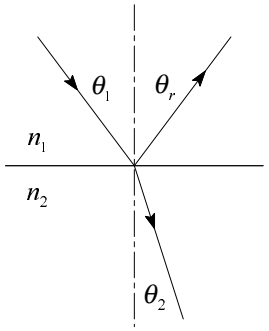
When a light ray encounters a boundary between two different transparent materials, it is partially reflected back into the first material and partially transmitted into the second. The magnitude and direction of the two rays depends on the optical properties of the two materials.

The speed of light in a material depends on the material's *refractive index*, with $c_{material} = c_{vacuum}/n$. When light passes from material of one refractive index to another, its speed changes. If the ray is normal to the boundary surface between the two materials, then the direction is unchanged, but if it is incident on the surface at an angle, it will be *refracted*, with the new direction determined by *Snell's Law*:




$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{14.2}$$

It is important to note that although most of the beam is refracted into the second medium, a portion (typically a few percent) is reflected back into the first, with the angle of reflection equal to the angle of incidence.

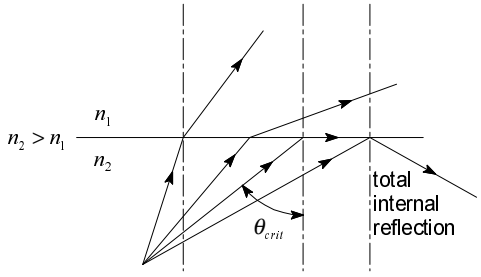



In addition, a ray propagating in any medium other than vacuum is subject to *absorption* or *attenuation*. Both refractive index and the coefficients of reflection and absorption vary with wavelength.

We can write Equation 14.2 as

$$\sin\theta_1 = \frac{n_2}{n_1}\sin\theta_2$$

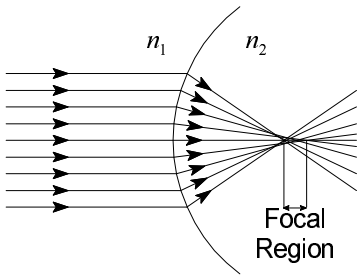
If the ray is traveling from material 2 to material 1 and $n_2 > n_1$ then any angle of incidence with $\theta_2 > \sin^{-1}(\frac{n_1}{n_2}) = \theta_{crit}$ requires $\sin\theta_1 > 1$, which is impossible for real values of $\theta_1$. In this case, the incident ray is reflected back into the original material (with $\theta_r = \theta_2$) with 100% efficiency. This condition is called *total internal refection*.



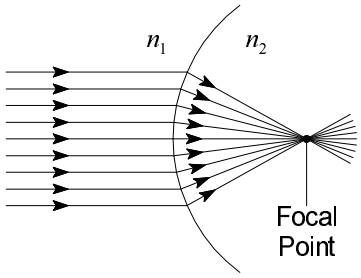
### 14.2.3 Lenses and Imaging

If the boundary surface between two regions of different refractive index is spherical rather than planar, the angle of the surface normal varies with position. If a plane wave, consisting of a "bundle" of parallel rays, crosses this surface, each ray will strike the surface at a different point, and hence be refracted by a different amount.

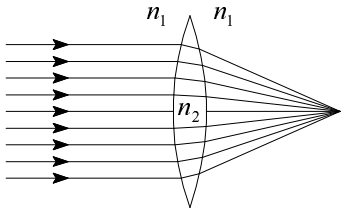
If the surface is convex as seen from the region of lower index, the rays will come to a *focus* in a small region of space before crossing over and diverging.



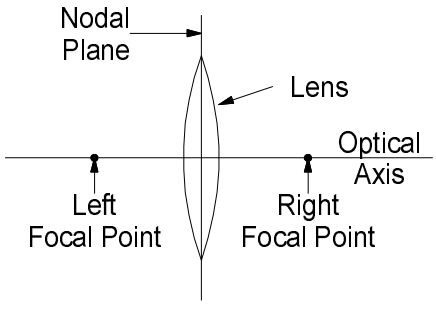

If an appropriate *aspherical* curvature is chosen, the rays will all come together at a single *focal point*. The spreading out of the focal point in systems with spherical surfaces is called *spherical aberration*, a condition which reduces the resolution of optical systems. Nevertheless, spherical surfaces are utilized in most optical systems because of their ease of manufacture.



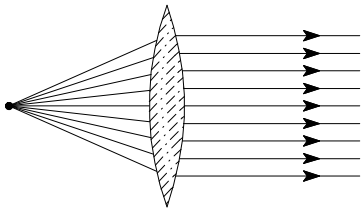

Note that for a single surface the focal point will be inside the medium of higher index (usually glass). If the converging rays pass through a second, glass-to-air surface before coming to a focus, the focus will occur in air.



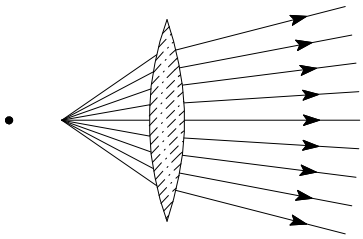

An optical element having two spherical surfaces, or one spherical and one flat surface, is called a *lens*. The line between the two centers of curvature defines the *optical axis* of the lens. An optical system is created by combining several optical elements (e.g. lenses, prisms, mirrors), usually along a common optical axis.



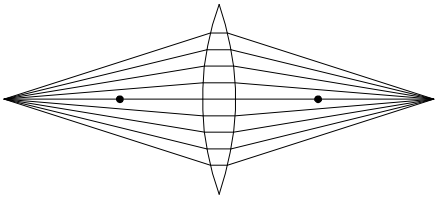
If a point source of light is placed at the focal point of a lens, the rays will follow the same paths (but in the reverse direction) as the parallel light being focused in the previous example. It is traditional to draw optical systems with light traveling from left to right, so the figure shows the focal point to the left of the lens. (In fact, a lens has two focal points, one on each side. These are rather unimaginatively named the left hand and right hand focal points.) The process of converting a point source to a parallel beam is called *collimation*.



If the point source in the previous example is moved closer to the lens, the amount by which the rays are refracted is no longer sufficient to collimate them. Instead, they continue to diverge, but less strongly than when they entered the lens.



Conversely, if the source is moved further from the lens (to the left of the left hand focal point), the rays emerging from the lens will *converge*, coming to a focus somewhere to the right of the right hand focal point.



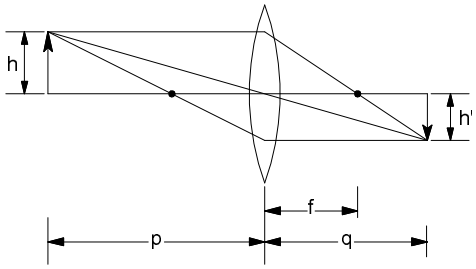
As the source moves still further away, the focus moves closer to the right hand focal point, eventually reaching it when the source reaches infinity.

#### 14.2.3.1 Imaging

If an illuminated *object* is placed to the left of the left hand focal point, the pattern of intensities and wavelengths of light emitted or reflected by the object will form an *image* on the other side. The light from each point on the object will be focused at a corresponding image point. For a flat or shallow object, the image points will lie in an *image plane* The orientation of the image will be reversed with respect to the object.

The relationship between the *object distance* $p$, the *image distance* $q$ and the focal length of the lens $f$ is given by the *lens equation*

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f}$$




The *linear magnification* produced by the lens is

$$m = \frac{\text{image height}}{\text{object height}} = \frac{h'}{h} = -\frac{q}{p}$$

The negative sign is required since the image is inverted.

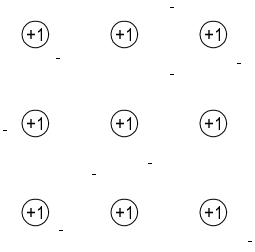
## 14.3 Optoelectronics: Solid State Light Sources and Detectors

There are many different types of light sources: incandescent bulbs, fluorescent bulbs, gas discharge tubes, electroluminescent panels, etc. There are also a variety of ways to convert light to an electrical signal. However, the vast majority of modern optical devices use semiconductor emitters and detectors.

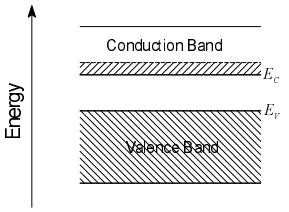
### 14.3.1 Semiconductors

#### 14.3.1.1 Conduction in Metals

In both metals and semiconductors the atoms are arranged in a regular lattice or crystal structure. In a metallic crystal, the nuclei and innermost shells of electrons are bound tightly together at the vertices of the lattice, but the electrons from the outermost shells are *free* to move about the crystal. (In the figure the number in the circle represents the net charge of the *atomic core*, i.e. the nucleus and the inner electron shells.) Electrical conduction in metals takes place via the motion of these free electrons in response to an externally applied electric field.



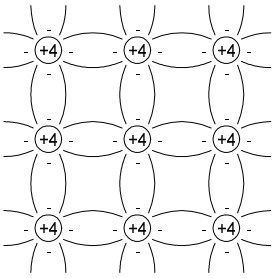
In an isolated atom, each electron can have only certain, discrete *energy levels*. In order to satisfy the exclusion principle, these levels broaden into *energy bands* as the atoms are brought together to form a crystal. The highest energy level in the outermost filled or *valence band* is called $E_v$. The lowest energy level in the next higher empty or partially filled *conduction band* is called $E_c$. Between the top of the valence band and the bottom of the conduction band is a range of forbidden energy states. The width of this forbidden region is called the *band gap* $E_g = E_c - E_v$.



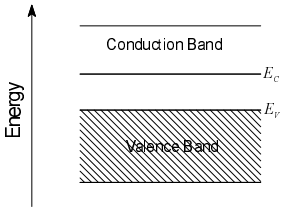

In terms of allowed electron energies, conduction electrons are free to move because there is a range of allowed states available between the highest filled level and the top of the conduction band to accommodate the increased kinetic energy of an electron as it is accelerated by the applied field.

#### 14.3.1.2 Intrinsic Semiconductors

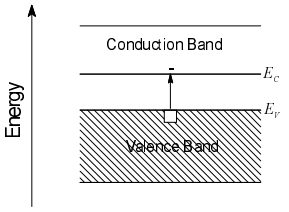
In a crystal of silicon, the atoms form *covalent* rather than *metallic* bonds. The four valence electrons of each atom are shared with four adjacent atoms, resulting in a completely filled valence band.



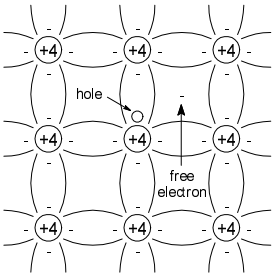

In this case there are no available states immediately above the highest energy electrons in the band, so they cannot accept the additional kinetic energy that would result from motion. Since the electrons are immobile, silicon should be an insulator, at least at absolute zero.



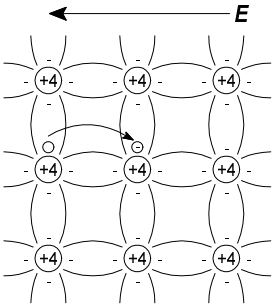

In semiconductors the band gap is small enough that at room temperature sufficient thermally excited electrons are promoted to the conduction band to support conduction. Since the number of electrons with thermal energy greater than $E_g$ increases with temperature, the conductivity of semiconductors is strongly temperature dependent.



When an electron is thermally promoted to the conduction band, it leaves behind an available, unoccupied state in the valence band. This state is called a *hole* and represents a localized net positive charge in the material.




Since this hole is an available state in the valence band, it can accommodate an accelerated valence band electron. In the presence of an electric field, an adjacent electron can move into the hole's position, with the former location of the electron becoming a new hole. Thus a hole may be thought of as a virtual positive charge which can move in response to an electric field (i.e. it is mobile).

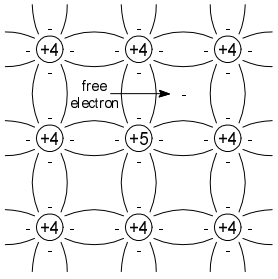


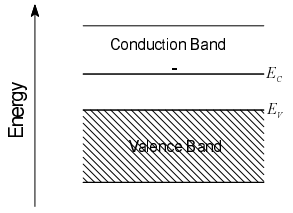

Conduction in a semiconductor involves both positive and negative carriers (holes and electrons). In a pure semiconductor, each free electron is associated with a corresponding hole, so current is carried equally by both.

#### 14.3.1.3 Extrinsic Semiconductors

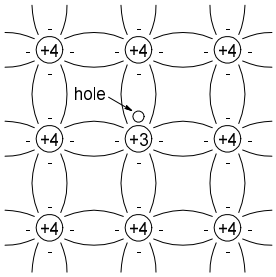
In a pure or *intrinsic* semiconductor, thermal or photo generated electrons and holes are created in pairs (*electron-hole pairs*), so that the density of electrons ($n$) is equal to the density of holes ($p$). In silicon at room temperature $n = p = n_i \approx 1.5{\times}10^{16}\text{m}^{-3}$. Unbalanced holes or electrons can be created by adding small amounts of impurities, a process called *doping*.

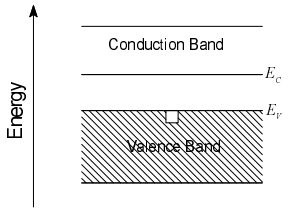
A pentavalent element, such as phosphorus, is called a *donor* impurity. Its atoms can fit onto the lattice in place of a silicon atom, with four of its outer shell electrons participating in the covalent bond. The remaining electron becomes a free electron. Typically the impurity concentration is made much higher than the intrinsic carrier density, so that $n \gg n_i$. Because of the higher density of charge carriers, $n$-type material has a much higher conductivity than the intrinsic semiconductor. In $n$-type material, $n \gg p$ so that electrons are the *majority carriers* of charge. Current in $n$-type material is carried almost entirely by electrons.



Similarly, trivalent *acceptor* impurities (such as boron) create a large majority of positive charge carriers (holes) so that $p \gg n_i$, producing $p$-type semiconductor material. In $p$-type material, holes are the majority carriers.

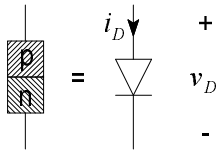



#### 14.3.1.4 The $p$-$n$ Junction

At the *junction* of $p$-type and $n$-type regions, carrier diffusion and the internal electric field produced by the fixed donor and acceptor sites produces a *depletion region* which is free of mobile charge carriers. An external voltage applied such that the $n$ side is positive with respect to the $p$ side (*reverse bias*) will widen the depletion region, reducing conductivity even further. An applied voltage in the opposite direction (*forward bias*) narrows, and eventually eliminates, the depletion region, leading to strong conduction.

As an approximation, the $p$-$n$ junction conducts current in the forward direction and blocks current in the reverse direction, making it useful as a rectifier or *diode*.

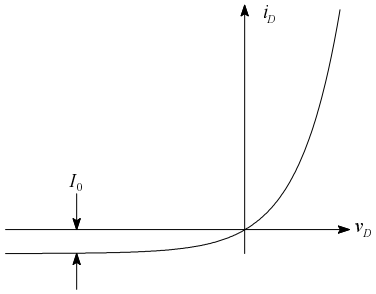



The I-V relationship for a semiconductor diode is given by the *diode equation*

$$i_{DD} = I_0[e^{\frac{q_e v_D}{kT}} - 1]$$

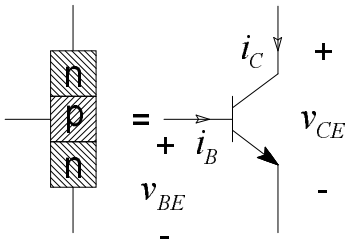
where

$i_{DD}$ = diode dark current
$q_e$ = electron charge ($1.60 \times 10^{-19}$ C)
$v_D$ = diode voltage
$k$ = Boltzman's constant ($1.38 \times 10^{-23}$ J/K)
$T$ = temperature (in K)



#### 14.3.1.5 Transistors

By combining two *pn* junctions in series (actually an *np* and a *pn*) we can produce a device called a *transistor* which can amplify currents. Leaving out the details, we can summarize the behavior of a transistor as follows: If an external voltage source is provided to insure that $v_{CE} > v_{BE}$ then we will have $i_B = \beta i_B$ where $\beta$ is the *current gain* or amplification factor of the transistor.




### 14.3.2 Optoelectronic Detectors

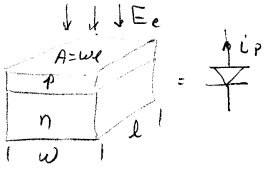
#### 14.3.2.1 Photoconductivity

If a photon of energy greater than $E_g$ is absorbed by a semiconductor, a photoelectron is promoted from the valence to the conduction band, creating a mobile electron-hole pair. The presence of additional charge carriers increases the material's conductivity, a phenomenon known as *photoconductivity*.

#### 14.3.2.2 Photodiodes

If sufficiently energetic photons are absorbed in the depletion region of a *pn*-junction, the electron-hole pairs that are created can be separated by the internal field, giving rise to a *photocurrent* $i_P$.

$$i_P = \frac{\eta E_e A q_e \lambda}{hc} \tag{14.3}$$



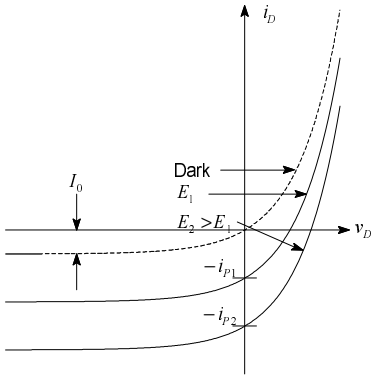

where

$\eta$ = quantum efficiency (electrons/photon)
$E_e$ = irradiance (W/m$^2$)
$A$ = diode junction area (m$^2$)
$h$ = Planck's constant
$\lambda$ = wavelength
$c$ = speed of light
$q_e$ = charge on electron

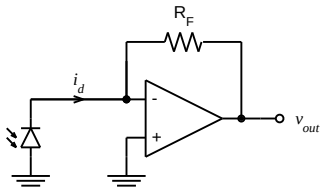
The diode equation describes the behavior of a *dark* semiconductor junction. If light is incident on the junction, the total diode current is the sum of the photocurrent and the dark current

$$i_D = i_{DD} - i_P$$

The minus sign indicates that the photocurrent flows in the opposite direction from the forward bias current.

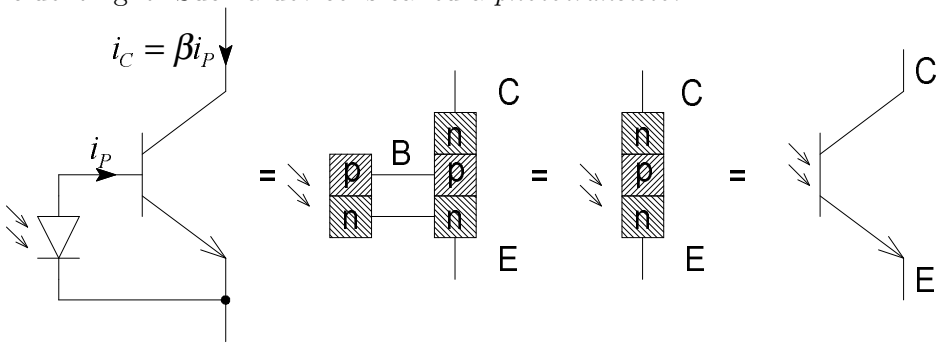



The photocurrent can be converted to a voltage by using a *transresistance amplifier* For this circuit, $v_{out} = -R_F i_d$, and since $v_D$ is forced to zero by the op amp, $i_d = -i_P$ so that $v_{out}$ is directly proportional to the irradiance.

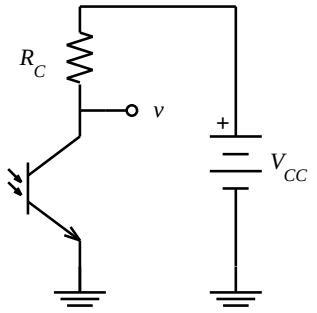



#### 14.3.2.3 Phototransistor

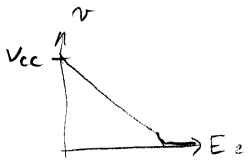
For most sensors we want the sensitive area $A$ to be small to provide fine resolution. From Equation 14.3 we see that this results in a correspondingly small photocurrent. In order to provide a reasonable level of current to the circuit, it will be necessary to amplify the photocurrent. We could do this with an external amplifier as shown above or we can build the amplification into the device. As the sequence below indicates, the effect of combining a photodiode and a transistor can be achieved with a transistor alone, if its base is exposed to the incident light. Such a device is called a *phototransistor*.



However, unlike the photodiode, a phototransistor cannot *produce* current, only *control* it. It requires an external voltage source ($V_{CC}$) to bias its junction, with the resulting current ($i_C$) being modulated by the light falling on the base.




With a resistor in series with the collector bias source, as in the above circuit, the collector voltage decreases as the irradiance increases.

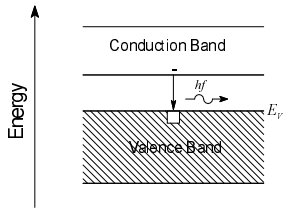



### 14.3.3 Light Emitting Diodes

When current flows through a forward biased ($pn$ junction) diode, electrons from the $n$-type material cross into the $p$ side of the junction, and vice versa. This results in a high concentration of both holes and electrons in the junction area. This condition is favorable for the *recombination* of holes and electrons. During recombination an electron in the conduction band returns to an unoccupied state in the valence band (the hole), in the process giving up an energy of approximately $E_g$.

In silicon the mechanism of recombination results in the energy being released as heat. But in other (so called *direct band gap*) semiconductors the energy is released as a photon, with

$$\lambda \approx \frac{ch}{E_g}$$




A diode exhibiting such photoemissive recombination is called a *light emitting diode* (LED). Since the rate of recombination will be approximately proportional to the carrier density, we have

$$\phi_e \propto i_D$$

over several orders of magnitude.

## 14.4 Sensors Based on Light Beam Interruption

### 14.4.1 The Emitter/Detector Pair

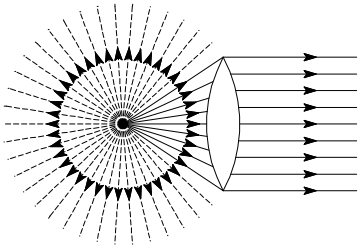
The common component of all sensors of the form shown in Figure 14.1 is the *emitter/detector pair* which performs the conversion from electrical to optical energy and back again.

If we replaced the sheet of paper in Figure 14.2 with a photodiode, we would have an emitter/detector pair. However, it wouldn't be a very good one. A typical photodiode has a junction area of a few $mm^2$, so the power incident on the junction would be less than 100 nW. Most of this power is in the infrared, below the band gap of silicon, and hence unable to generate photoelectrons. The resulting photocurrent would be only a few nA.

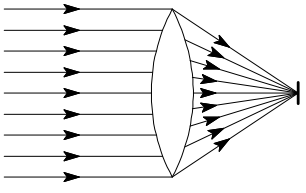
We could increase this by:

1. Increasing the power of the source (but 100 W is already pretty high).
2. Using a source with more of its optical power in the region of sensitivity of the detector (e.g. an LED).
3. Increasing the area of the detector (but larger detectors are slower and more expensive).
4. Reducing the distance between the source and the detector.
5. Focusing more of the flux produced by the source onto the detector.

As an example of 5, if we place a point source at the focal point of a positive lens it will capture a fraction of the total flux and focus it into a *collimated beam* whose diameter is equal to the diameter of the lens. If a reflector is placed directly behind the source, this amount will be doubled.



The same idea can be used to increase the effective area of the detector. If the detector is placed at the focal point of a lens receiving a collimated beam, then all of the light incident on the lens will be focused onto the detector. The detector's effective area becomes the area of the lens. Figure 14.4 shows an emitter/detector pair using a collimated beam.



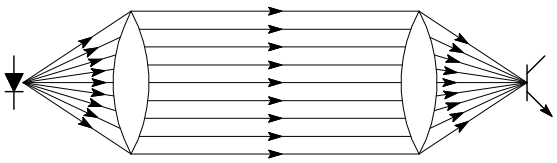
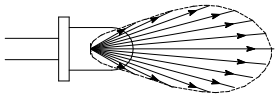
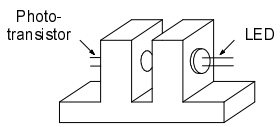


Figure 14.4: A Collimated Beam Emitter/Detector Pair.

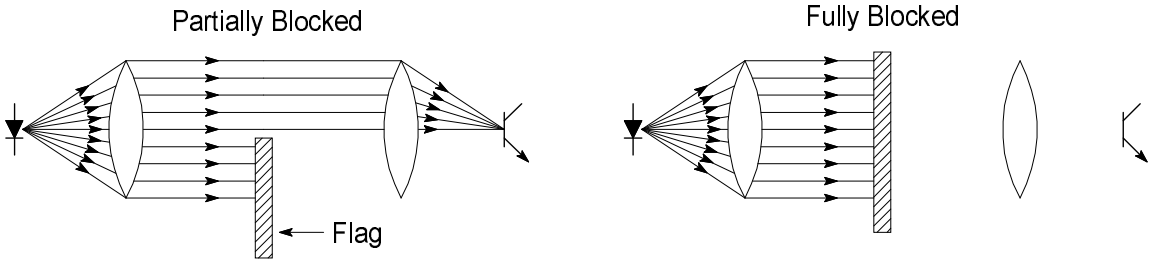


Such a lens is built into the package of most LEDs, photodiodes, and phototransistors by the simple expedient of moulding the plastic case into the appropriate shape for the lens. In addition, LEDs are mounted on a reflector which directs more of the light in a forward direction. However, LEDs are not true point sources, so the resulting beam has an intensity profile with a significant amount of divergence (usually $> 10°$). This is satisfactory for short emitter-detector separations. For larger distances, more elaborate arrangements using discrete lenses or laser sources may be required.

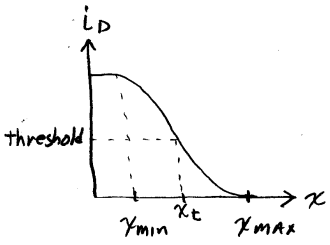



For short emitter to detector distances, an *optical interrupter* is the most convenient form of emitter/detector pair. It consists of an LED and a phototransistor moulded into an opaque plastic housing, with provisions for electrical connection and mechanical mounting, and an open slot through which the beam passes. A portion of the mechanism to be sensed passes into this slot to block the beam.

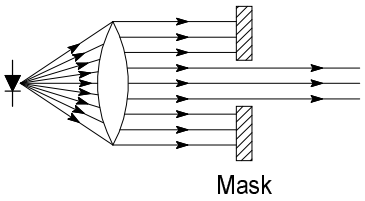
The most common use of a collimated beam emitter/detector involves modulating the received beam power by blocking part of all of the beam by an opaque shutter or "flag" attached to a portion of the mechanism whose position is to be sensed.



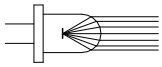

Since the detector current is proportional to the incident power, it will vary between its maximum value and zero as the shutter moves from the fully unblocked to the fully blocked position. The detector current may be interpreted as a continuous variable representing position. More frequently it is compared to a threshold to provide a binary indication of the position of the shutter.



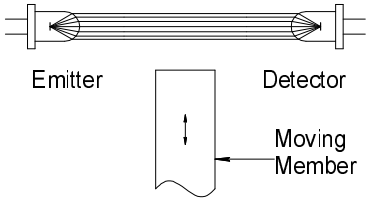
In the later case, the aperture is usually restricted by a mask to produce a narrower beam.



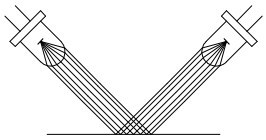

To simplify the following drawings, we will denote the combination of emitter, lens, and mask by a single package with a narrow beam emerging.



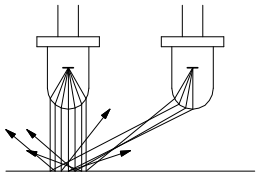
A typical application for an interrupter is as a *limit switch*, to provide an indication when a machine member is leaving its normal range of motion, so that it may be stopped before damage occurs.




The emitter/detector arrangement described above is called *transmittive*. It is also possible to produce a *reflective* emitter/detector structure. In this case the emitted beam is reflected from a portion of the mechanism. If this is a polished metal surface, resulting in a specular reflection, the light reaching the detector will be roughly the same as in the transmittive case.



For a diffuse reflection (e.g. from paper of white paint) the reflected light will be scattered in all directions and the irradiance at the detector will be considerably reduced. On the other hand, the alignment between emitter, detector, and reflecting surface is much less critical, and in fact the emitter and detector can be parallel rather than inclined.

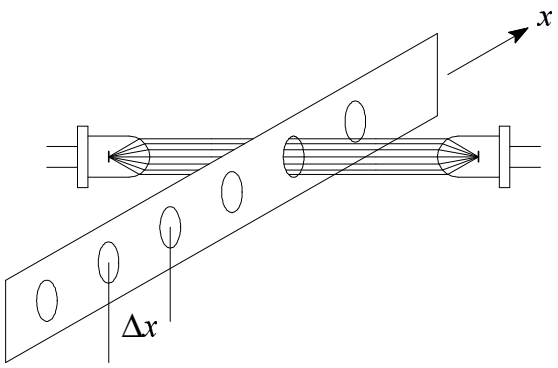


To interrupt the beam, the reflecting surface can be removed (e.g. by a slot) or rendered less reflective (e.g. by a stripe of black paint. Although the reflective structure is more convenient to use (requiring access to only one side of the moving member), it also has lower contrast between on and off states, resulting in increased sensitivity to noise, especially from ambient light.

### 14.4.2 Optical Encoder

Although partial blockage of the beam can be used to provide continuous measurement of a range of positions, difficulty in achieving a uniform beam intensity limits the accuracy of this approach. A more common technique for position sensing is to use an *optical encoder*.

Instead of interrupting the beam with a solid shutter, a strip with a regularly spaced array of holes is placed between the emitter and detector. Each time the strip is moved through a distance of $\Delta x$, a pulse will be produced on the output. By counting the number of pulses and multiplying by $\Delta x$ the total distance moved may be determined. Since this device produces a pulse for each *increment* of motion, it is called an *incremental encoder*.



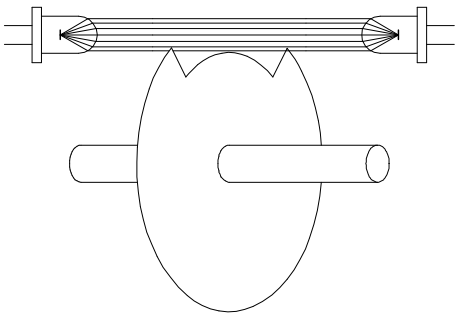

There are several problems with this configuration:

1. The position sensing is relative, not absolute. For absolute position sensing, a separate means of determining the "origin" must be provided (e.g. a limit switch) and provision must be made in the initialization procedure for determining this position.
2. There is no way to determine the direction of motion, so only unidirectional motion may be sensed.
3. As the resolution is increased, spacing between the holes decreases, so they must be made smaller, passing less light. Eventually, not enough light is available for accurate sensing.

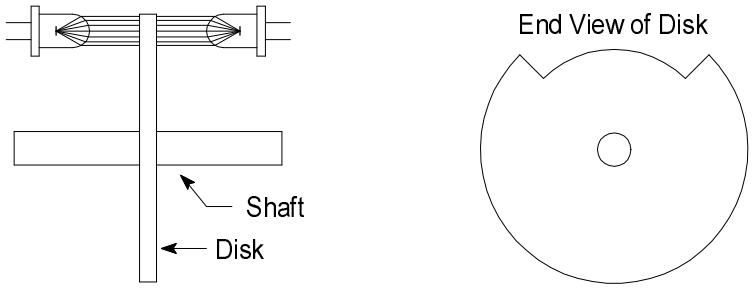
Before addressing these problems, let's see how we can use this same idea to produce a rotational position sensor.

#### 14.4.2.1 Rotational Encoders

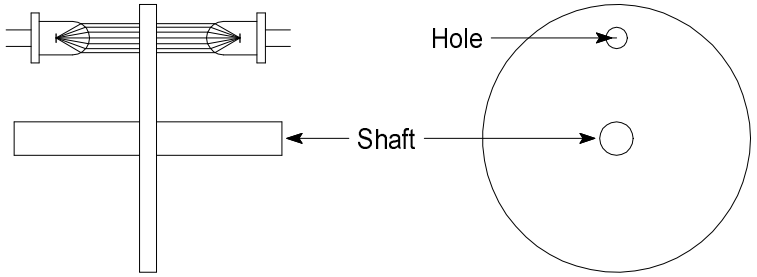
We can start with a rotational limit sensor. In this case, instead of having a flag at each limit of motion, we can simply cut out a notch which denotes the allowed range of motion.



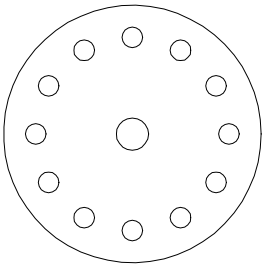
To simplify the drawings, we will show a cross-sectional view of the complete device along with an end-on view of the disk.




In addition to sensing whether the shaft is within a certain range of angles, we can also determine when it reaches a single, specific angle by drilling a single hole at that position.




An obvious extension is to drill a uniformly spaced array of holes, giving a rotational or angular incremental encoder.

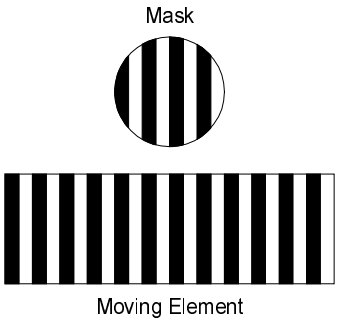


As each successive hole passes in front of the mask, the amount of light passing through to the photodetector varies continuously from zero to the maximum and back to zero. The exact shape of the resulting waveform will depend on the shape of the mask and the holes. We could use this continuous waveform to interpolate between adjacent hole positions, but usually we simply compare it to a threshold to produce a digital waveform, ideally a symmetric square wave. If we are able to count both rising and falling edges, the resolution will be half the hole-to-hole spacing.

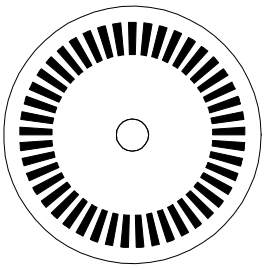
#### 14.4.2.2 Practical Encoders

Now let's address the problems with our prototype encoder, starting with number 3. The first thing to note is that holes need to become smaller only in the direction of motion; their transverse dimension can remain the same. I.e. instead of round holes we can use rectangular slits. This helps somewhat since the area now goes down as $\frac{1}{n}$ rather than $\frac{1}{n^2}$ (where $n$ is the number of holes per unit motion), but it still goes down.

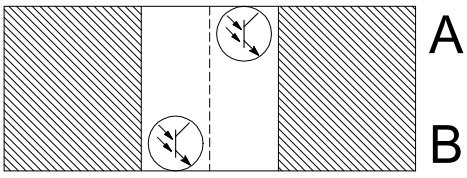
The key is to replace the single slit in the mask with an array of equally spaced slits, i.e. make the stationary mask identical to the moving element, but limited to the aperture of the emitter/detector pair. If the pitch of both gratings is the same and the open and blocked areas are of equal width, the amount of unblocked area will range between 0 and 50%, regardless of the pitch.



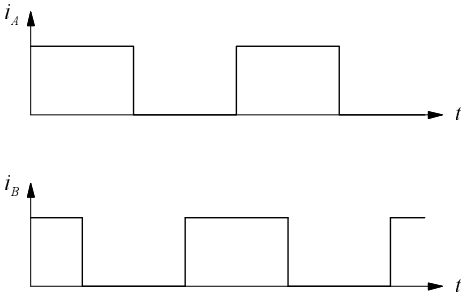

This idea also works for rotational encoders, with slits of uniform angular, rather than linear, spacing.



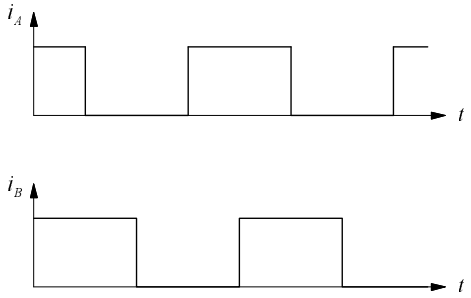
Moving on to problem number 2: we can determine the direction of motion by creating two (waveforms) 90° out of phase. This can be done by having two detectors, shifted by 1/4 of the slot pitch. For clarity, the detector apertures are shown as being circular with diameter equal to half the slot width. In an actual device they would be slotted masks as shown above.




For left to right motion, the waveform produced by detector A *lags* behind that of detector B by 90°.

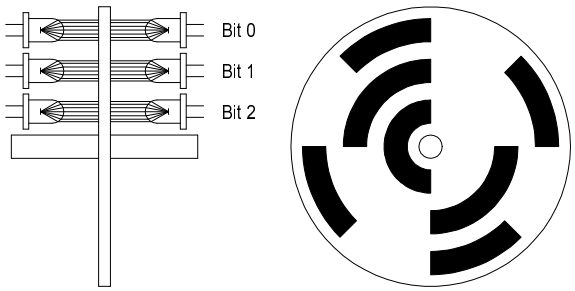


However, when the direction of motion is reversed, output A *leads* output B by 90°.




Note that we can use an optical encoder to measure velocity. The frequency of the output waveform of a rotational encoder is equal to $N$ times the rotational velocity, where $N$ is the number of slots in the disk.

We can solve problem 1 (as well as problem 2) by producing an *absolute encoder*. In this case, instead of counting pulses to determine position, we encode the position directly in the pattern of bits coming from the encoder. This requires a separate track for each binary bit in the word.



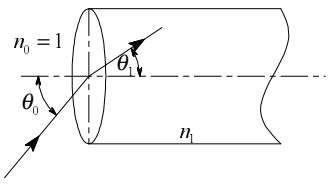

## 14.5 Fiber Optic Sensors

### 14.5.1 Optical Fiber

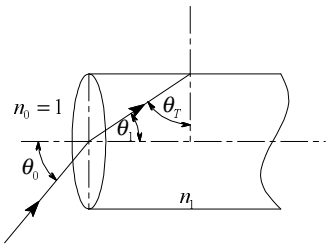
An optical fiber is a long, thin cylinder of transparent glass or plastic with the end faces polished flat. A light ray entering the end of a fiber at an angle of $\theta_0$ with respect to the axis of the fiber will be refracted so that

$$n_0 \sin\theta_0 = n_1 \sin\theta_1 \tag{14.4}$$

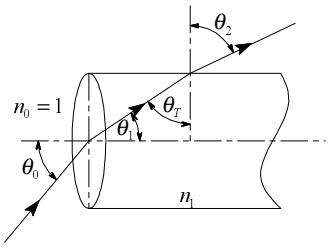
where $n_0$ is the refractive index of the initial medium (usually air, so that $n_0 \approx 1$) and $n_1$ is the refractive index of the fiber. We will assume that $n_1 > n_0$.



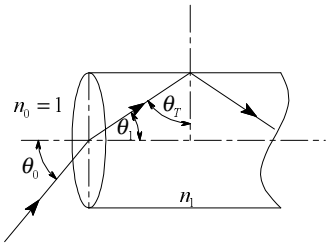
If $\theta_0 \neq 0$, the refracted ray will eventually strike the wall of the cylinder, at an angle of $\theta_T$ with respect to the normal to the surface, where $\theta_T = \frac{\pi}{2} - \theta_1$.



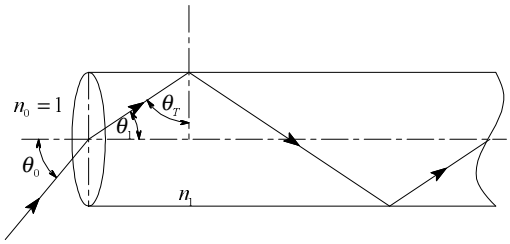

If $\theta_T$ is small, the ray will pass through the wall of the cylinder back into the original medium, having been refracted so that $n_0 \sin\theta_2 = n_1 \sin\theta_T$.



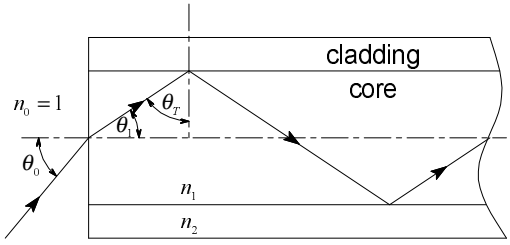

As $\theta_T$ increases, we approach the condition for total internal reflection $\theta_{crit} = \sin^{-1}(\frac{n_0}{n_1})$. For $\theta_T > \theta_{crit}$, the ray will be totally reflected from the upper wall toward the lower wall.




But since the walls are parallel, $\theta_T$ at the lower wall will be the same as $\theta_T$ at the upper wall. Total internal reflection again occurs and the ray continues down the length of the fiber, bouncing back and forth off the walls.




For total internal reflection to be truly total, the interface between the two refracting materials must be perfectly clean. To prevent contamination of this critical surface, practical fibers are formed by *cladding* the *core* cylinder of the fiber with a layer of compatible transparent material with $n_2 < n_1$. In this case the condition for total internal reflection is $\sin\theta_T > \frac{n_2}{n_1}$.




Since $\theta_T = \frac{\pi}{2} - \theta_1$, we require $\cos\theta_1 > \frac{n_2}{n_1}$. But $\cos\theta_1 = \sqrt{1 - \sin{\theta_1}^2}$ and from Equation 14.4 $\sin\theta_1 = \frac{n_0}{n_1}\sin\theta_0$. With a little algebra, and assuming $n_0 = 1$, we can show that we must

have

$$\sin\theta_0 < \sqrt{{n_1}^2 - {n_2}^2}$$

in order for total internal reflection to occur.

This defines the *angle of acceptance* of the fiber

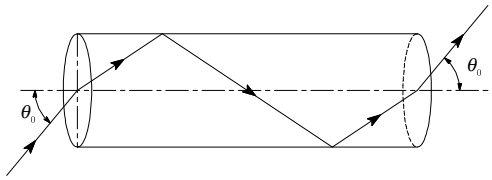
$$\theta_A = \sin^{-1}(\sqrt{{n_1}^2 - {n_2}^2})$$

Rays entering the fiber at angles less than $\theta_A$ will be accepted and propagate down the fiber. For angles greater than $\theta_A$, the rays will leak into the cladding and be lost. The sine of the angle of acceptance is called the *numerical aperture* of the fiber:

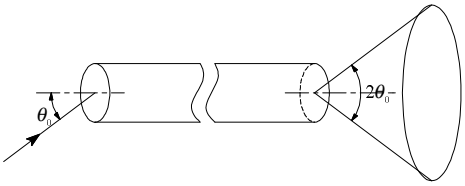
$$NA = \sin\theta_A = \sqrt{{n_1}^2 - {n_2}^2}$$

#### 14.5.1.1 Fiber Output Patterns

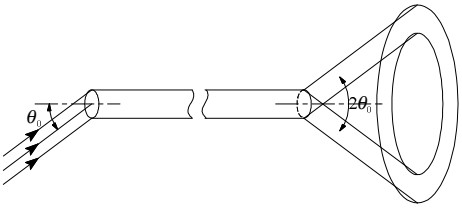
From the previous figures it appears that a single ray entering one end of a fiber will emerge as a single ray at the same angle from the opposite end.



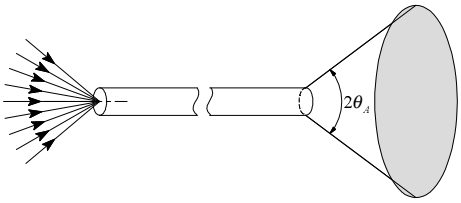

In fact, unless the ray passes through the axis of the fiber, it will emerge within a hollow cone whose total include angle is twice the angle of the input ray.




A collimated beam will produce a hollow cone whose wall thickness is equal to the core diameter of the fiber.




If the input end is illuminated by light arriving from all angles within the angle of acceptance, the output will be a solid cone of light with included angle of $2\theta_A$.




#### 14.5.1.2 Coupling to a Fiber

There are two common ways of efficiently achieving the condition described in the last example above, where the cone of acceptance of the fiber is completely filled. A distributed

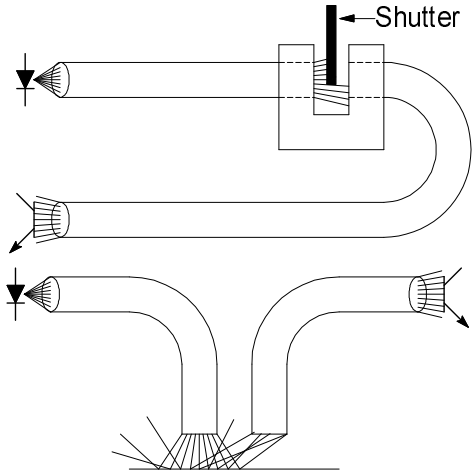
source (such as an LED) can be placed on or very close to the face of the fiber. Alternately, a point source or collimated beam (as from a laser) could be focused on the entry face.

### 14.5.2 Extrinsic Fiber Optic Sensors

Fiber optic sensors may be divided into two groups: *extrinsic* sensors where the fiber serves primarily to conduct light to and from the point of interaction with the mechanism, and *intrinsic* where the propagation of light is modified by mechanically altering the behavior of the fiber.

We can convert the devices from Section 14.4 to extrinsic fiber sensors by replacing the local semiconductor emitters and detectors with remotely mounted emitters and detectors connected by fibers to their previous locations.

We could create fiber coupled transmittive interrupters ...



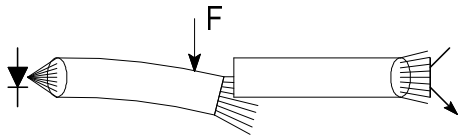

... reflective interrupters ...

... and of course optical encoders of various forms. While this appears to unnecessarily add additional cost and complexity, there are some potential advantages:

- An optical fiber, basically a solid piece of glass, can operate in environments of high radiation, temperature, or G-forces that would incapacitate or destroy semiconductor devices.
- Since the signal is conveyed as light, rather than electric current, fibers can operate in levels of EMI that would corrupt electrical signals.
- The small size of a fiber allows a larger number of sources or detectors to be placed in the same area than would be possible with discrete devices. This would allow, for example, a smaller disk to be used on an absolute encoder for the same number of bits of resolution.

A more interesting situation arises when we incorporate the fiber more directly into the operation of the sensor.

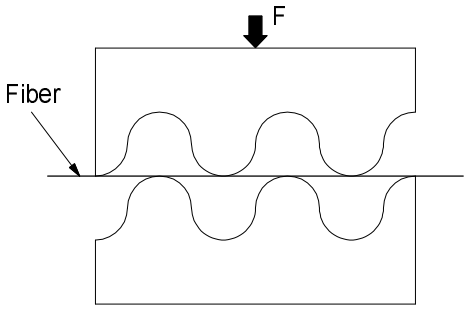
For example, we can create a bending beam load cell utilizing the fiber as the elastic beam. With no force applied, the two fibers are aligned along a common axis. As the amount of force is increased, the left hand fiber deflects, decreasing the amount of light coupled into the right hand fiber.




### 14.5.3 Intrinsic Fiber Optic Sensors

In an intrinsic sensor a single length of fiber carries light from the emitter to the detector. The amount of light reaching the detector is modulated by mechanical action on the fiber itself.

For example, in the microbending sensor, pressure applied to a fiber laid between a pair of grooved plates results in a number of short radius bends in the fiber. At these bends, the angle of incidence of some of the rays in the fiber will exceed the critical angle and leak out of the core into the cladding. The amount of light lost will depend on the amount of bending. This phenomenon can be used to measure displacement, strain, force, or pressure, depending on how the sensor is utilized.




Another application is to detect cracks in structural members. A fiber is glued to the surface along the path to be monitored, perpendicular to the expected direction of crack formation. If a crack forms, the brittle fiber will fracture and the amount of light at the detector will drop.