CHAPTER SIX

OPTICAL DETECTORS

6. 1 Responsivity, noise equivalent power, quantum efficiency, detectivity, and rise time.

6.2 Light Detection

6.3 Detector Characteristics

6.3 Detector Characteristics

6.5 Types of Detectors

6.6 Calibration

6.7 Power Supplies for Optical Detectors

CHAPTER SIX

OPTICAL DETECTORS

The detection of optical radiation is usually accomplished by converting the optical energy into an electrical signal. Optical detectors include photon detectors, in which one photon of light energy releases one electron that is detected in the electronic circuitry, and thermal detectors, in which the optical energy is converted into heat, which then generates an electrical signal. Often the detection of optical energy must be performed in the presence of noise sources, which interfere with the detection process. The detector circuitry usually employs a bias voltage and a load resistor in series with the detector. The incident light changes the characteristics of the detector and causes the current flowing in the circuit to change. The output signal is the change in voltage drop across the load resistor. Many detector circuits are designed for specific applications.

6. 1 When you complete this chapter, you will be able to:

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 circuitry used with photomultiplier detectors.

5. Draw and explain photodiode circuits for use in the photoconductive and photovoltaic modes of operation.

6. Fabricate a circuit for operation of a photodiode and use it for detection of light in both photoconductive and photovoltaic modes of operation.

6.2 Light Detection

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 hence may be used to monitor the presence of the light and to measure its intensity.

Detectors suitable for monitoring optical power or energy are commonly employed along with lasers and other light sources. For an application such as laser communication, a detector is necessary as the receiver. For applications involving interferometry, detectors are used to measure the position and motion of the fringes in the interference pattern. In applications involving laser material processing, a detector monitors the

laser output to ensure reproducible conditions. In very many applications of light, one desires a detector to determine the output of the laser or other light source. Thus, good optical detectors for measuring optical power and energy are essential.

All optical detectors respond to the power in the optical beam, which is proportional to the square of the electric field. They are thus called "square-law detectors." Microwave detectors, in contrast, can measure the electric field intensity directly. But all the detectors that we consider here exhibit square-law response. This is also true of other common optical detectors such as the human eye and photographic film.

The detection and measurement of optical and infrared radiation is a well-established area of technology. In recent years, this technology has been applied specifically to laser applications, and detectors particularly suitable for use with lasers have been developed. Commercial developments have also kept pace. Detectors specially designed and packaged for use with lasers are marketed by numerous manufacturers. Some detectors are packaged in the format of a power or energy meter. These devices include a complete system for measuring the output of a specific class of lasers, and include a detector, housing, amplification if necessary, and a readout device.

In this chapter, we will describe some of the detectors that are available. 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. We shall also define some of the common terminology.

There are two broad classes of optical detectors: **photon detectors** and **thermal detectors**. Photon detectors rely on the action of quanta of light energy to interact with electrons in the detector material and to generate free electrons. To produce such effects, the quantum of light must have sufficient energy to free an electron. The wavelength response of photon detectors shows a long-wavelength cutoff. When the wavelength is longer than the cutoff wavelength, the photon energy is too small to liberate an electron and the response of the detector drops to zero.

Thermal detectors respond to the heat energy delivered by the light. The response of these detectors involves some temperature-dependent effect, like a change of electrical resistance. Because thermal detectors rely on only the amount of heat energy delivered, their response is independent of wavelength.

The output of photon detectors and thermal detectors as a function of wavelength is shown schematically in Figure 1. This shows how the output of thermal detectors is independent of wavelength. It also shows the typical spectral dependence of the response of photon detectors, which increases with increasing wavelength until the cutoff wavelength is reached. At that point it drops rapidly to zero.

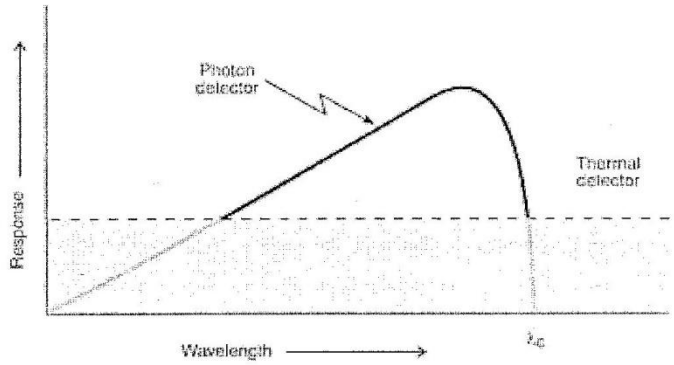


Fig. 1

Schematic drawing of the output of photon detectors and thermal detectors as a function of wavelength. The position of the long-wavelength cutoff, λ_c, for photon detectors is indicated.

Photon detectors may be further subdivided into the following groups:

• **Photoconductive.** The electrical conductivity of the material changes as a function of the intensity of the incident light. Photoconductive detectors

are semiconductor materials. They have an external electrical bias voltage.

• **Photovoltaic**. These detectors contain a p-n semiconductor junction and are often called photodiodes. A voltage is self generated as radiant energy strikes the device. The photovoltaic detector may operate without external bias voltage. A good example is the solar cell used on spacecraft and satellites to convert the sun's light into useful electrical power.

• **Photoemissive.** These detectors use the photoelectric effect, in which incident photons free electrons from the surface of the detector material. These devices include vacuum photodiodes, bipolar phototubes, and photomultiplier tubes.

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 describe these effects in more detail later in the discussion of the types of detectors.

6.3 Detector Characteristics

The performance of optical detectors is commonly described by a number of different figures of merit, which are widely used throughout the field of optical detection. They were developed originally to describe the capabilities of a detector in responding to a small signal in the presence of noise. As such, they are not always pertinent to the detection of laser light. Often in laser applications—for example, in 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 other source that may be present. But in other applications, like laser communications, infrared thermal imaging systems, and detection of backscattered light in laser Doppler anemometry, the signals are small, and noise considerations are important. It is also worthwhile to define these figures of merit because the manufacturers of detectors usually describe the performance of their detectors in these terms.

The first term that is commonly used is *responsivity*. This is defined as 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. This depends on the particular type of detector and how it is used. Figure 1 can be considered to be a representation of how the *responsivity* varies with wavelength for photon detectors and thermal detectors. We note that the *responsivity* gives no information about noise characteristics.

The *responsivity* is an important characteristic that is usually specified by the manufacturer, at least as a nominal value. Knowledge of the *responsivity* allows the user to determine how much detector signal will be available in a specific application. One may also characterize the spectral *responsivity*, which is the *responsivity* as a function of wavelength.

A second figure of merit, one that depends on noise characteristics, is the noise equivalent power **(NEP)**. *This is defined as the radiant power that produces a signal voltage (current) equal to the noise voltage (current) of*

the detector. Since the noise is dependent on the bandwidth of the measurement, that bandwidth must be specified.

The equation defining NEP is

$$NEP = I\,A\,V_N / V_S\,(\Delta f)^{1/2} \quad \text{.....(1)}$$

where I 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 watts per (hertz to the one-half power), commonly 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. The NEP of a detector is dependent on the area of the detector. To provide a figure of merit under standard conditions, a term called ***detectivity*** is defined. Detectivity is represented by the symbol D^*, which is pronounced as D-star. It is defined by

$$D^* = A^{1/2} / NEP \quad \text{.....(2)}$$

Since many detectors have NEP proportional to the square root of their area, D* is independent of the area of the detector and provides a measure of the intrinsic quality of the detector material itself, independent of the area with which the detector happens to be made. When a value of D* for a photodetector 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 Δ f. The dependence of D* on the wavelength λ , the frequency f at which the measurement is made, and the bandwidth Δf are specified in the notation $D^*(\lambda, f, \Delta f)$. The reference bandwidth is frequently taken as 1 hertz. The units of $D^*(\lambda, f, \Delta f)$ are centimeters

(square root hertz) per 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 the modulation frequency and the bandwidth on the noise characteristics.

Another commonly encountered figure of merit for photodetectors 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 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 p-n 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. Thus 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 radiant energy for producing electrical current in a circuit. It may be related to the responsivity by the equation:

$$Q = 100 \times R_d \times hv \qquad = 100 \times R_d \times (1.2395/\lambda) \qquad \ldots\ldots\ldots\ldots (3)$$

where Q is the quantum efficiency (in %) and R_d is the responsivity (in amperes per watt) of the detector at wavelength λ (in micrometers), and hv is the photon energy. The right portion of the equation makes calculation easy for a specific responsivity at a given wavelength.

Another important detector characteristic is the speed of the detector response to changes in light intensity. If a constant source of light energy

is instantaneously turned on and irradiates a photodetector, it will take a finite time for current to appear at the output of the device and for the current to reach a steady value. If the same source is turned off instantaneously, it will take a finite time for the current to decay back to its initial zero value. The term response time generally refers to the time it takes the detector current to rise to a value equal to 63.2% of the steady-state value reached after a relatively long period of time. (This value is numerically equal to $\mathbf{1-1/}e$, where e is the base of the natural logarithms.) The recovery time is the time photocurrent takes to fall to 36.8% of the steady-state value when the light is turned off instantaneously.

Because photodetectors often are used for detection of fast pulses, a more 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 points 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 note that the fall time may be different numerically from the rise time.

Of course light sources are not turned on or off instantaneously. For accurate measurements of rise time and fall time, the source used for the measurement should have a rise time much less that the rise time of the detector that is being tested. Generally one will accept a source whose rise time is less than 10% of the rise time of the detector being tested.

Other factors that affect measured rise times are the limitations introduced by the electrical cables and by the display device, for example,

the oscilloscope or recorder. These devices can make the measured rise time appear longer than the value that arises from the detector alone.

The response time of a photodetector arises from the transit time of photogenerated charge carriers within the detector material and from the inherent capacitance and resistance associated with the device. It is also affected by the value of the load resistance that is used with the detector. 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 or display device as low as possible. This will help to keep the RC 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.

Another important characteristic of detectors is their ***linearity***. Photodetectors are characterized by a response that is linear with incident intensity 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 handle without becoming saturated. Saturation is a condition in which there is no further increase in detector response as the input light is increased. Linearity may be quantified in terms of the maximum

percentage deviation from a straight line over a range of input light levels. For example, the maximum deviation from a straight line could be 5% over the range of input light from 10^{-12} W/cm^2 to 10^{-4} W cm^2. One would state that the linearity is 5% over eight orders of magnitude in the input.

The manufacturer often specifies a maximum allowable continuous light level. Light levels in excess of this maximum may cause saturation, hysteresis effects, and irreversible damage to the detector. 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.

6.4 Noise Considerations

The topic of noise in optical detectors is a complex subject. 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 distinguished as external and internal. External noise involves those disturbances that appear in the detection system because of actions 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 may be considered as an ever-present limit to the smallest signal that may be handled 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. The output from a random noise generator might look like what is shown in Figure 2, which is a plot of instantaneous voltage as a function of time. Because of the random nature of the noise, the voltage fluctuates about an average value V_{av}. How does one describe these variations? A simple average is meaningless because the average is zero. Rather, one uses an average of the squares of the deviations around V_{av}, with the average taken over a period of time T much longer than the period of the fluctuations.

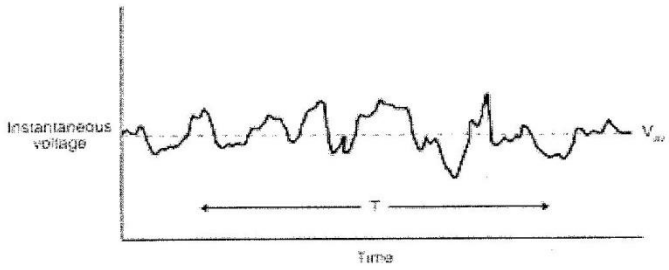


Fig. 2
A record of random noise voltage

Mathematically this is expressed as:

$$\overline{V^2} = \overline{(V(t) - V_{av})^2} = \frac{1}{T}\int_0^T (V(t) - V_{av})^2 dt \qquad \text{.....(4)}$$

where $v(t)$ is the value of the voltage at time t and $\overline{V^2}$ is termed the mean square voltage fluctuation, and the bar over a quantity indicates an average value.

The right side of the equation contains an integral sign with limits 0 and T. This means that one adds all values $(v(t) - V_{av})^2$ for each small increment of time Dt in the interval from time 0 to time T. It is unlikely that one would calculate the noise voltage directly from this equation, because the process would be laborious. The equation does define the basic concept.

If two or more noise sources are present, their total effect is found by adding their mean square voltages. Since mean square voltages are proportional to power, this is equivalent to saying that noise powers are additive, but noise voltages or currents are not. As an example, if two independent noise sources are present in a circuit, with root-mean-square (rms) noise voltages of 30 and 40 microvolts, the total rms noise voltage is $(30^2 + 40^2)^{1/2}$ or 50 microvolts.

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 the four types most often encountered in a system for visible and infrared detection:

- ***Johnson noise***
- ***Shot noise***
- ***1/f noise***
- ***Photon noise***

Johnson noise is a type of noise 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.

The mean square value of the voltage associated with Johnson noise is:

$$\overline{V^2} = 4KTR\Delta f \quad \text{.....(5)}$$

where K is Boltzmann's constant (1.38×10^{-23} joule/degree Kelvin), T is the absolute temperature, R is the circuit resistance (ohms), and Δf is the bandwidth of the amplification (Hz). Since the load resistance is usually greater than the internal resistance of the photodetector, the Johnson noise may be dominated by the load resistor. *This equation is also frequently written as $V^2 = 4KTRB$, with B denoting the bandwidth.*

The equation indicates methods to reduce the magnitude of the 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.

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. It originally was likened to the the noise of a hail of shot striking a target; hence the name

shot noise was applied. In semiconductors, the major source of noise is due to random variations in the rate at which charge carriers are generated and recombine. This noise, called generation-recombination or *gr* noise, is the semiconductor counterpart of shot noise. The mean-square current fluctuation for shot noise in a semiconductor photodetector is:

$$\overline{i^2} = 2eI_{DC}\Delta f \quad \text{.....(6)}$$

where e is the electronic charge (1.6×10^{-19} coulomb), I_{DC} is the DC component of dark leakage current (or any current) in amperes, and Δf is the bandwidth of the amplification (Hz). Note that here we have specified a mean square noise current, instead of a voltage.

The 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. 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. With respect to photodiodes, it is sometimes called boxcar noise, because it can suddenly appear and then disappear in small boxes of noise observed over a period of time.

The mechanisms that produce 1/f noise are poorly understood and there is no simple mathematical expression to define 1/f noise. The noise power is inversely proportional to f, the modulation frequency. This dependence of the noise power leads to the name for this type of noise. To reduce 1/f noise, a photodetector should be operated at a reasonably high frequency, often taken as 1000 Hz. This is a high enough value to reduce the contribution of 1/f noise to a small amount. Since to reduce Johnson noise

and shot noise, the amplification bandwidth should be small (perhaps 1 Hz), measurements of the spectral detectivity are often expressed as D*(λ ,1000,1).

Even if all the previously discussed sources of noise could be eliminated, there would still be some noise present in the output of a photodetector because of the random arrival rate of photons from the source of radiant energy that is being measured and from the background. This contribution to the noise is called photon noise, and 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 may be possible to reduce the field of view of the detector so as to view only the source of interest. In other cases it may be possible to cool the background. Both these measures may be used to reduce the background noise contribution to photon noise.

Any of the types of noise described here, or a combination of the noise powers from a combination of the sources, will set an upper limit to the detectivity of a photodetector system.

6.5 Types of Detectors

6.5.1 Photon Detectors

As mentioned before, photon detectors rely on liberating free electrons and require the photon to have sufficient energy to exceed some threshold, i.e., the wavelength must be shorter than the cutoff wavelength. We will consider three types of ***photoeffect*** that are often used for detectors. These are the **photovoltaic** effect, the **photoemissive** effect, and the **photoconductive** effect.

The photovoltaic effect and the operation of photodiodes both rely on the presence of a p-n junction in a semiconductor. When such a junction is in the dark, an electric field is present internally in the junction region 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.

When light falls on the junction, it 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 a change in voltage that may be measured externally. This process is the origin of the so-called photovoltaic effect. The photovoltaic effect is the generation of a voltage when light strikes a semiconductor p-n junction. We note that the open-circuit voltage generated in the photovoltaic effect may be detected directly and that no bias voltage or ballast resistor is required.

It is also possible to use a p-n junction to detect light if one does apply a bias voltage in the reverse direction. By reverse direction, we mean the direction of low current flow, that is, with the positive voltage applied to the n-type material. A p-n junction detector with bias voltage is termed a photodiode. The current-voltage characteristics of a photodiode are shown in Figure 3. The curve labeled "dark" represents conditions in the absence of light; it displays the familiar rectification characteristics of a p-n semiconductor diode. The other curves show the current-voltage characteristics when the device is illuminated at different light levels. The characteristics of a photovoltaic detector, with zero applied voltage, are represented by the intersections of the different curves with the vertical axis. The photodiode detector is operated in the lower left quadrant of this figure. 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 resistor.

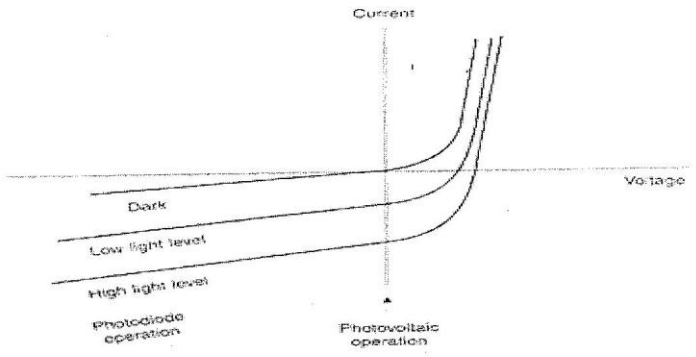


Fig. 3
Current voltage characteristics for photovoltaic detectors and photodiodes. Regions for photodiode and photovoltaic operation are indicated.

To increase the frequency response of photodiodes, a type called the PIN photodiode has been developed. This device has a layer of nearly intrinsic 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. A sufficiently large reverse bias voltage is applied so that the depletion layer, from which free carriers are swept out, spreads to occupy the entire volume of intrinsic material. This volume then has a high and nearly constant electric field. It is called the depletion region because all mobile charges have been removed. 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 such PIN photodiodes can be very high, of the order of 10^{10} Hz. This is higher than the frequency response of p-n junctions without the intrinsic region.

A variety of photodiode structures is available. No single photodiode structure can best meet all system requirements. Therefore, a number of different types has been developed. These include the planar diffused photodiode, shown if Figure 4a, and the Schottky photodiode, shown in Figure 4b. The planar diffused photodiode is formed 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 R_L.

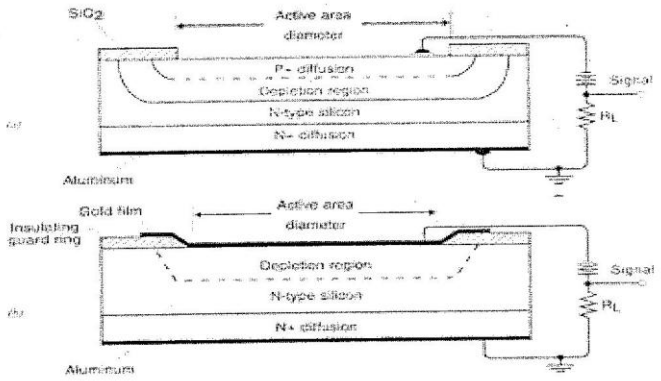


Fig. 4
Photodiode structures. (a) Planar diffused photodiode.
(b) Schottky photodiode. The load resistor is denoted R_L.

A number of different semiconductor materials is in common use as optical detectors. They include silicon in the visible and 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 of typical commercial silicon photodiodes is shown in Figure 5. 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, He-Ne, AlGaAs, and Nd:YAG. As a practical matter, silicon photodiodes have become the detector of choice for many laser applications. They represent well-developed technology and are widely available. They represent the most widely used type of laser detectors for lasers operating in the visible and near-infrared portions of the spectrum.

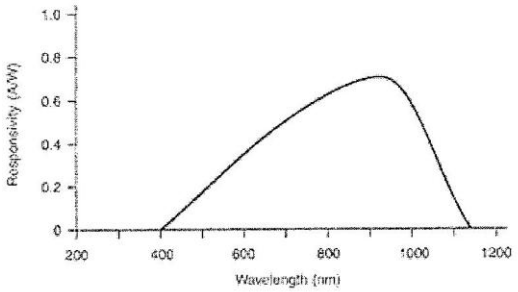


Fig. 5

Responsivity as a function of wavelength
for typical silicon photodiodes

One photon produces one electron-hole pair in the material, so long as the photon energy is high enough. Absorption of one photon then gives a constant response, independent of wavelength (provided that the wavelength lies within the range of spectral sensitivity of the detector). One photon of ultraviolet light and one photon of infrared light each produces the same result, even though they have much different energy. For constant photon arrival rate, as wavelength increases, the incident

power decreases, but the response remains the same. Therefore, the value of D* increases, reaching a maximum at the cutoff wavelength, which is equal to the Planck's constant times the velocity of light divided by the band gap of the material. At longer wavelengths, the detectivity decreases rapidly because the photons do not have enough energy to excite an electron into the conduction band. 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 p-n 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.

Silicon and other photodiodes have been configured as power meters, which are calibrated so that the detector output may be presented on a display as the laser power. These devices may be used directly to measure the power of a continuous laser or of a repetitively pulsed laser operating at a reasonably high pulse-repetition rate. Some commercial units can measure powers down to the nanowatt regime. Because the photodiode response varies with wavelength, the manufacturer usually supplies a calibration graph to allow conversion to wavelengths other than that for

which the photodiode was calibrated. Alternatively, some units are supplied with filters to compensate for the wavelength variation of the detector response, so that the response of the entire unit will be wavelength-independent, at least over some interval. Such packaged power meters provide a convenient and useful monitor of laser output power.

A photoemissive detector employs a cathode coated with a material that emits electrons when light of wavelength shorter than a certain value falls on the surface. The electrons emitted from the surface may be accelerated by a voltage to an anode where they give rise to a current in an external circuit. These detectors are available commercially from a number of manufacturers. They represent an important class of detectors for many applications.

An important variation of the photoemissive detector is the photomultiplier. This is a device with 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 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 7 shows the principles of

operation of the tube. With careful design, photoelectrons emitted from the cathode will 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 are required to 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.

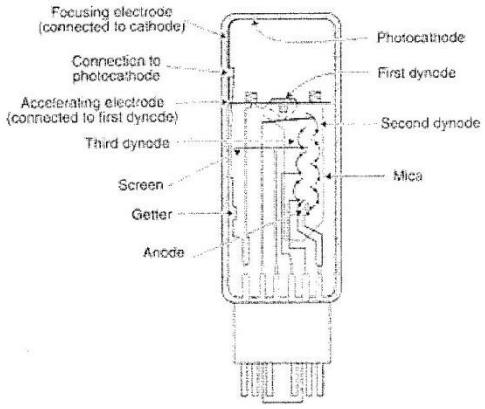


Fig.6

Diagram of typical photomultiplier tube structure

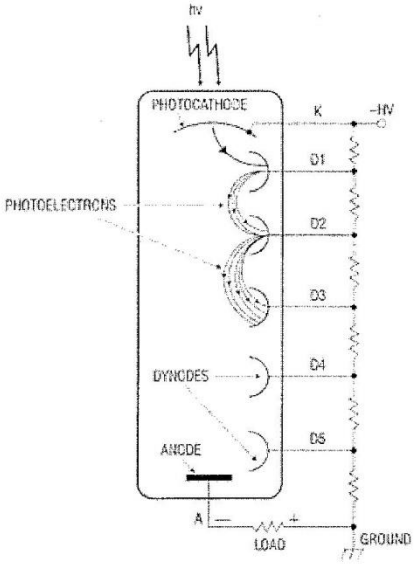


Fig. 7

Principles of photomultiplier operation.
The dynodes are denoted D1, D2, and so on.

The current gain of a photomultiplier is defined as 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. Figure 8 shows a plot of 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. Photomultiplier tubes can in fact detect the arrival of a single photon at the cathode. Applications of

photomultiplier tubes include scintillation counting, air-pollution monitoring, photon counting, star tracking, photometry, and radiometry.

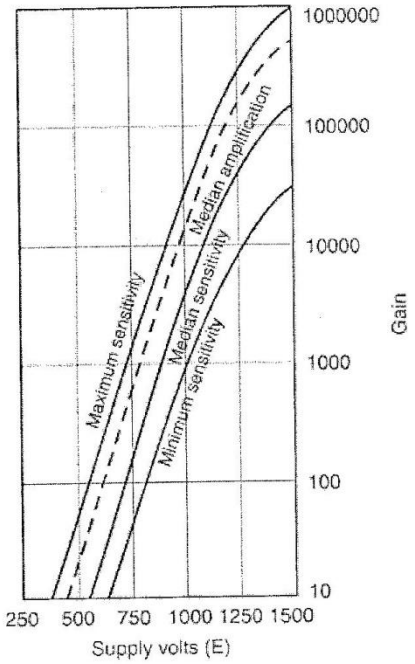


Fig. 8

Photomultiplier gain as a function of applied voltage

A third class of photodetector uses the phenomenon of 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 cause excitation, either by raising electrons across the forbidden band gap or by activating impurities present within the band gap. 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 spectrum, at wavelengths where photoemissive detectors are not available and the wavelengths are beyond the cutoffs of the best photodiodes (silicon and germanium). Many different materials are used as infrared photoconductive detectors. Typical values of spectral detectivity for some common devices operating in the infrared have already been shown in Figure 6. 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 cryogenic temperatures, which may involve some inconvenience in practical applications.

In its most simple form, a photoconductive detector is a crystal of semiconductor material that has low conductance in the dark and an increased value of conductance when it is illuminated. In a series circuit with a battery and a load resistor, the detector element has its conductance increased by light. The sensing of the presence of the light is accomplished via the increased voltage drop across the load resistor. But it is possible to use photodiodes in a photoconductive mode as well as in the photovoltaic mode that we have already described.

Figure 9 shows the voltage-current characteristics in a different way. Note that the curve has been rotated 180° from Figure 3. The photoconductive mode of operation appears on the right side of the figure. Negative values of reverse voltage increase from left to right, starting at the origin. The forward-biased (photovoltaic) device is a

voltage generator. If it operates with a low-resistance load, the operation is along the near-vertical line and the current output is fairly linear with input radiation. As the load resistance increases, the output becomes nonlinear. If one observes the open-circuit voltage (load line horizontal), the open-circuit voltage is found to be proportional to the logarithm of the input light intensity.

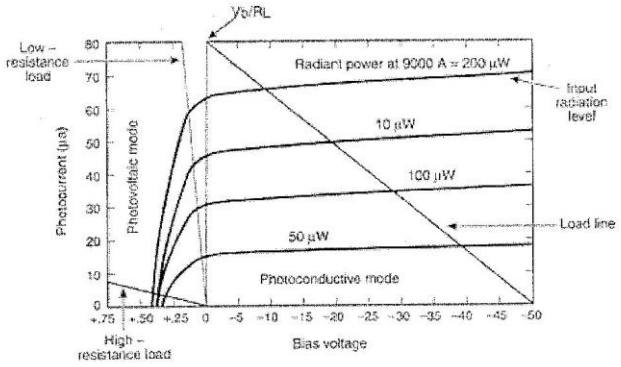


Fig. 9

Volt-ampere characteristics of photodiodes

In the reverse-biased or photoconductive mode, linear operation is maintained so long as the photodiode is not saturated and the bias voltage is higher than the product of the load resistance and the current.

For a reverse-biased device, the photodiode exhibits higher responsivity, faster response time, and greater linearity than when operated in the forward-biased mode. One drawback is the presence of a small dark current. In the forward-biased mode, the dark current may be eliminated. This makes photovoltaic devices desirable for low-level measurements in which the dark current would interfere. But the responsivity and speed

decrease and the response becomes nonlinear for large values of load resistance.

6.5.2 Thermal detectors

Now we turn to the second broad class of photodetectors, thermal detectors. *Thermal detectors respond to the total energy absorbed, regardless of wavelength. They have no long-wavelength cutoff in their response, as photon detectors do. The value for D* for a thermal detector is independent of wavelength. Thermal detectors generally do not offer as rapid response as photon detectors,* and for laser work are not often used in the wavelength region in which photon detectors are most effective ($\leq$ 1.55 μ m). They are often used at longer wavelengths.

Pyroelectric detectors represent one popular form of thermal detector. These detectors respond to the change in electric polarization that occurs in certain classes of crystalline materials as their temperature changes. The change in polarization, called the pyroelectric effect, may be measured as an open-circuit voltage or as a short-circuit current. The temporal response is fast enough to respond to very short laser pulses. This behavior is in contrast to that of many other thermal detectors, which tend to be slower than photon detectors. Pyroelectric detectors are often used in conjunction with CO_2 lasers.

The calorimeter represents another type of thermal detector. Calorimetric measurements yield a simple determination of the total energy in a laser pulse, but usually do not respond rapidly enough to follow the pulse shape. Calorimeters designed for laser measurements usually use a blackbody absorber with low thermal mass 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 laser pulse. The temperature-measuring devices include **thermocouples**, **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.

Many different types of calorimeters have been developed for measuring the total energy in a laser pulse or for integrating the output from a continuous laser. Since the total energy in a laser 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.

One form of 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 acts as a very efficient absorber. Thermistor beads or thermocouples are placed intimately 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 unbalance 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. In some designs, two identical cones are used to form a conjugate pair in the bridge circuit. This approach allows cancellation of drifts in the ambient temperature.

A calorimeter using a carbon cone or similar design is a simple and useful device for measurement of laser pulse energy. In the range of energy below **1 J** or so, an accuracy of a few percent or better should be attainable. The main sources of error in conical calorimeters for pulsed energy measurements are loss of some of the energy by reflection, loss of heat by cooling of the entire system before the heat is distributed uniformly, and imperfect calibration. Calibration using an electrical-current pulse applied to the calorimeter element can make the last source of error small. With careful technique, the other sources of error can be held to a few percent.

Calorimeters using absorbing cones or disks with thermocouples to sense the temperature rise have been developed for laser pulses with energy up to hundreds of joules. When the laser energy becomes high, destructive effects, such as vaporization of the absorbing surface, may limit the usefulness of calorimeters. Since calorimeters require surface absorption of the laser energy, there are limits to the energy that a calorimeter can withstand without damage.

If the response of the calorimeter is fast, it can be used for measurement of power in a continuous laser beam. The temperature of the absorber will reach an equilibrium value dependent on the input power. Such units are available commercially as laser power meters, with different models capable of covering the range from fractions of a milliwatt to ten kilowatts.

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. Also,

power meters based on these thermal detectors tend to cover a higher range of laser power than do the models based on photodiodes.

Some values of D* for thermal detectors are shown in Figure 10. 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.

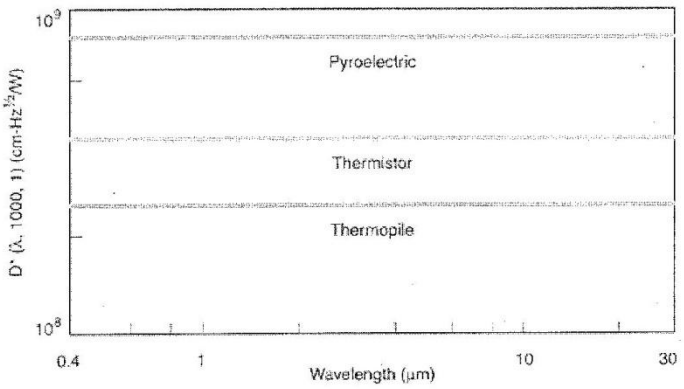


Fig. 10

Detectivity (D*) as a function of wavelength for several typical thermal detectors. The temperature of operation is 295 K.

6.6 Calibration

The response of any photodetector in current (or voltage) per unit input of power is often taken as the nominal value specified by the manufacturer. For precise work, the detector may have to be calibrated by the user. But accurate absolute measurements of power or energy are difficult. A good calibration requires painstaking work.

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. Gelatin or glass neutral-density filters and semiconductor wafers are subject to damage by high-power laser beams. Liquid filters containing a suitable absorber (for example, an aqueous solution of copper sulfate for ruby lasers) are not very susceptible to permanent damage.

The calibration of filters is a difficult task, because the filters also 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. Filters may be calibrated by measuring both the incident power and the transmitted power, but the measurement must be done at low enough irradiance so that the filter (and the detector) does not become saturated.

One useful 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 11. The goniometric distribution of the reflected light is independent of the azimuthal angle and depends on the angle q from the normal to the surface in the following simple manner:

$$P\omega\, d\omega = P_{tot} \cos\theta\, d\omega/\pi \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where $P\omega$ is the power reflected into solid angle $d\omega$ at angle θ from the normal, and P_{tot} is the total power. This relation is called Lambert's cosine law, and a surface that follows this law is called a Lambertian surface. Many practical surfaces follow this relation approximately. The power that reaches the detector after reflection from such a surface is

$$P_{detector} = P_{tot} \cos\theta \, A_d/\pi D^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

where 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 reflector to the detector. This approximation is valid when D is much larger than the detector dimensions and the transverse dimension of the laser beam. With a Lambertian reflector, the power incident on the photosurface may be adjusted simply in a known way by changing the distance D. 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 total beam power.

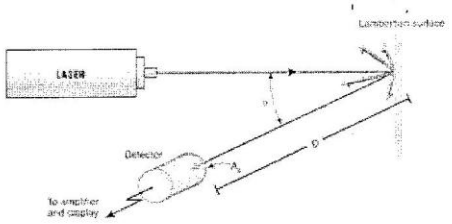


Fig. 11

Arrangement for measuring laser power output

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 many applications, one uses 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 is delivered to a calibrated calorimeter. The total energy arriving at the calorimeter is determined. 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. Finally, the calorimeter is removed and the beam is used for the desired application, while the detector acts as a pulse-to-pulse monitor.

Electrical calibration of power meters has also become common. The absorbing element is heated by an electrical resistance heater. The electrical power dissipation is determined from electrical measurements. The measured response of the instrument to the known electrical input provides the calibration. It is assumed that the deposition of a given amount of energy in the absorber provides the same response, independent of whether the energy was radiant or electrical.

The difficulty of accurate measurement of radiant power on an absolute basis is well known. Different workers attempting the same measurement often obtain substantially different results. This fact emphasizes the need for care in the calibration of optical detectors.

6.7 Power Supplies for Optical Detectors

The basic power supply for a photodetector consists of a bias voltage applied to the detector and a load resistor in series with it. The basic circuit for a photoconductive detector is shown in Figure 12. As the irradiance on the detector element changes, its conductance changes because of the free carriers generated within it. A change in the conductance increases the total current in the circuit and decreases the voltage drop across the detector. The load resistor is necessary to obtain an output signal. If the load resistor were zero, all the bias voltage would appear across the detector and there would be no signal voltage available. In the circuit shown, an increase in light intensity increases the voltage drop across the resistor, yielding a signal that may easily be monitored. If the light intensity is modulated in a periodic fashion, an AC signal will be detected.

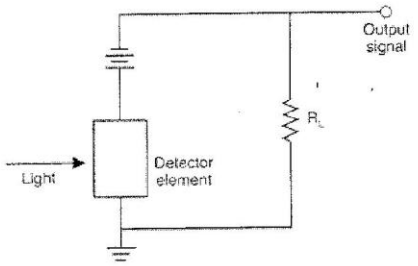


Fig. 12

Basic circuit for operation of a photoconductive detector. The load resistor is R_L.

The magnitude of the available signal increases as the value of the load resistor increases. But this increase in available signal must be balanced

against possible increase in Johnson noise and possible increase in rise time, because of the increased RC time constant of the circuit. The designer must trade these effects against each other to obtain the best result for the particular application.

A photovoltaic detector requires no bias voltage; it is a voltage generator itself. The basic circuit for a photovoltaic detector is shown in Figure 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. The resulting IR drop across the resistor again is available as a signal to be monitored.

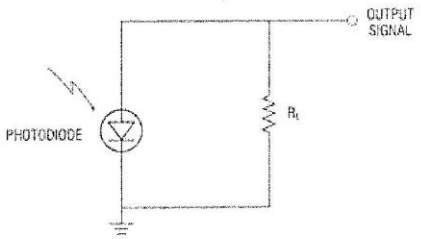


Fig. 13

Basic circuit for operation of a photovoltaic detector.
The symbol for a photodiode is indicated. The load resistor is R_L.

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. The value of the detector shunt resistance drops exponentially as the light intensity increases. The output

voltage then increases logarithmically with light intensity. Disadvantages of this circuit are the nonlinear nature of the response and the fact that the signal depends on the shunt resistance of the detector, which may have a spread in values from different production batches of detectors.

To counter these disadvantages, a photovoltaic photodiode is often used in a circuit such as shown in Figure 14. In this case, the load resistance has a value much less than the shunt resistance of the photodiode. The operation thus corresponds to the line marked low-resistance load in Figure 11. Again the photocurrent flows through the load resistor and produces the observed signal. The photocurrent is fed to the virtual ground of an operational amplifier. This provides amplification to counter the decreased voltage drop resulting from the low value of the load resistor.

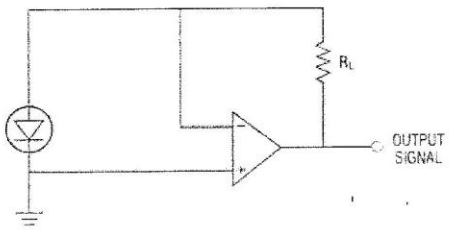


Fig. 14

Circuit for photovoltaic operation with low load resistance. The load resistor is R_L and the feedback resistor is R_F.

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 eliminated.

We have mentioned previously that photodiodes may be operated in a photoconductive mode. A circuit that provides this mode is shown in Figure 15. In this mode, the photocurrent produces a voltage across the load resistor which is in parallel with the shunt resistance of the detector. In this mode, the shunt resistance is nearly constant. The diode is reverse biased, so that the operation is in the third quadrant of Figure 3. One may use large values of load resistance, to obtain large values of signal, and still obtain linear variation of the output with light intensity.

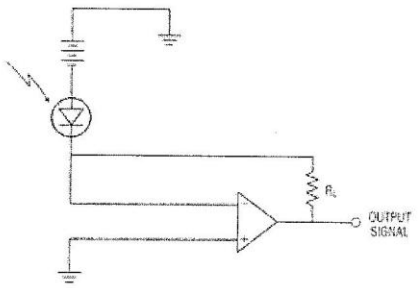


Fig. 15

Circuit for operation of a photodiode in the photoconductive mode. The load resistor is R_L and the feedback resistor is R_F.

This circuit is capable of 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.

Many different types of circuits have been designed for photodetector operation for particular applications. We will present two specialized

circuits for specific applications. These will give an idea of the range of circuits that may be employed for detection of radiant energy.

Figure 16 shows a circuit designed for detection of far-infrared radiation, in the 10-μ m region. The detector is a mercury cadmium telluride detector operated in the photoconductive mode. This detector is a cryogenic detector, operated at a temperature of 77 K. It is generally packaged in a Dewar assembly and cooled with liquid nitrogen. The input radiation is chopped with a rotating mechanical chopper wheel, to provide an AC signal. This circuit has been designed for low-noise operation.

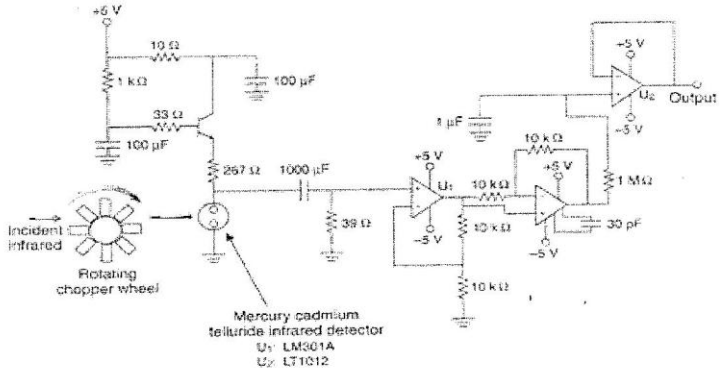


Fig. 16

Circuit for low-noise infrared detector

Figure 17 shows a circuit designed as a receiver for high-frequency fiber-optic communication systems. The circuit uses a back-biased silicon photodiode. It provides an output with transistor-transistor logic (TTL). In this system, the light emerging from the optical fiber is incident on the

detector. The resulting signal is amplified. If this circuit is employed in a relay station in a telecommunications link, the output may be used to drive a laser diode source that retransmits the information into the next segment of the fiber-optic cable.

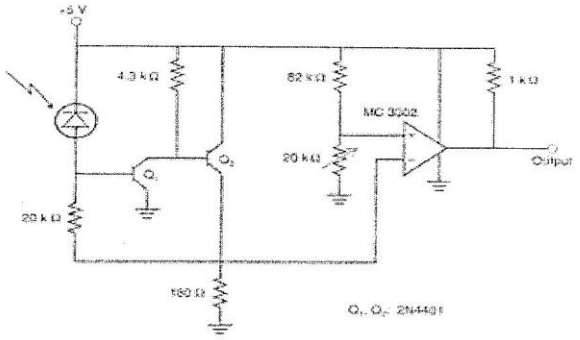


Fig. 17

Fiber-optic receiver circuit

Now let us turn to the circuitry employed for photomultiplier tubes. In the circuits discussed above, the voltages were low, five volts to a few tens of volts. The photomultiplier tube requires higher voltages, in the kilovolt range.

Because the gain of photomultiplier tubes is a strong function of the applied voltage, a small change in power-supply voltage will result in a large change in gain. Thus, one must use a well-regulated, stable power supply for photomultiplier applications.

The power supply for a photomultiplier consists of a voltage-divider circuit, as illustrated in Figure 18 for a 10-stage photomultiplier. The total voltage around 875 V is applied from the anode to cathode. A string of resistors of equal value is connected in parallel with the dynodes. This

ensures that the voltage applied from one dynode to the next is equal. This arrangement is called a voltage-divider network. It is the arrangement usually employed with photomultipliers, instead of applying separate voltage sources to each dynode. The circuit shown is a half-wave rectifier circuit. The connections are to the pins on the socket into which the photomultiplier tube is inserted.

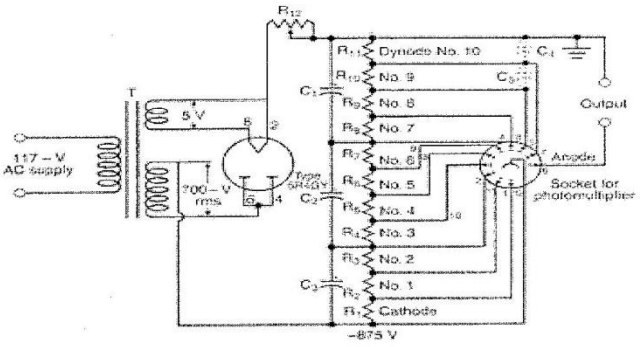


Fig. 18

Typical voltage-divider circuit for a 10-stage photomultiplier tube

An alternative arrangement is shown in Figure 19 for a 12-stage photomultiplier. This circuit uses a regulated DC power supply as the voltage source. The voltage-divider network is similar to that shown in the last figure. Although again the connections would be made directly to the socket pins, the connections are shown to the internal structure of the tube in order to make clear the functionality. The use of the adjustable resistor R_1 is not strictly necessary. Its purpose is to minimize dispersion of the transit time of the electrons, but for many applications it could be eliminated from the circuit.

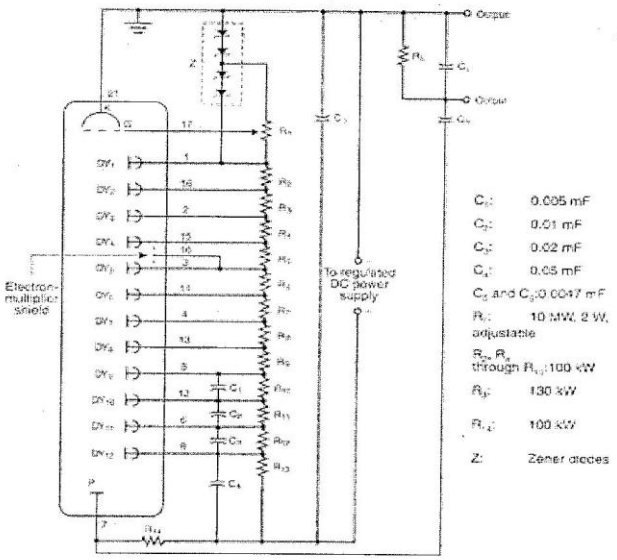


Fig. 19

Alternate voltage-divider circuit for
a 12-stage photomultiplier

These two examples illustrate the general principles of power supplies for photomultiplier tubes.

For many of the diverse applications for which photodetectors are employed, the photodetector circuits are custom designed for the needs of the specific application. One exception, for which standardized packages are available, involves photodiodes for use in the visible and near infrared. Manufacturers offer integrated photodiode/preamplifier circuits. The photodiode is integrated on the same chip with an operational amplifier. Both photovoltaic-mode and photoconductive-mode chips are available. Thus the circuitry is already integrated with the detector and the user needs only to supply the specified voltage inputs.

Monolithic amplifier units specifically designed for use with photodetectors are also available. The amplifier converts the signal from a photodiode directly into a voltage that can be used to drive an oscilloscope, recorder, or other voltage-sensing equipment.

Most of this discussion about circuits has involved circuitry for photon detectors. We briefly describe some circuits for thermal detectors. Figure 20 shows the basic circuit for a thermal detector that employs a temperature-based change in resistance, such as a thermistor or bolometer. In this very simple circuit, the signal arises from the change in voltage drop across the load resistor when the resistance of the detector element is changed by heating by the incident radiant energy. If the load resistance is much smaller than the resistance of the detector element, the percentage change of the signal will be maximized.

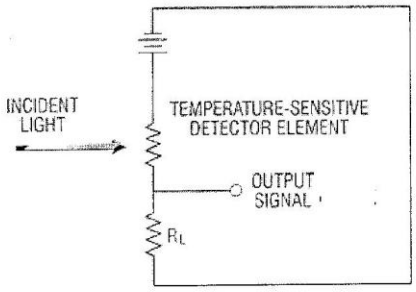


Fig. 20

Basic circuit for a thermal detector using a detector element that changes resistance with temperature. The load resistor is R_L.

Figure 21 shows a circuit for use with a pyroelectric detector and a pulsed source of light energy. The pyroelectric detector is a capacitor formed by depositing metal electrodes on a piece of pyroelectric material, such as lithium tantalate. The detector response arises from the change of electric

polarization of the material with temperature. As radiant energy is absorbed and the temperature rises, the change in electric polarization leads to a displacement current in the material, which in turn causes a compensating current flow in the external circuit. The pyroelectric element in this mode acts directly as a current generator. The output signal is derived from the voltage drop across the load resistor.

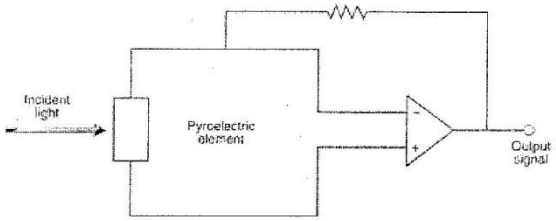


Fig. 21

Circuit for use with a pyroelectric detector and a pulsed light source. The load resistor is R_L.

If the light source is continuous, a chopper must be inserted in the beam.

Chapter Five

Optical Modulation

An *optical modulator* is used to provide directional information for tracking and to suppress unwanted signals from backgrounds. The optical modulator can assume many forms, but basically each can be described as a pattern of alternately clear and opaque areas carried on a suitably transparent substrate. It is common practice to call the optical modulator a *reticle* or a *chopper*; occasionally it is referred to as an *episcotister*, a designation usually restricted to scholarly literature. Reticle patterns for infrared systems range from very simple patterns for converting a dc to an ac system, through patterns used to discriminate against unwanted backgrounds, to patterns that code the radiant flux with information about the direction of a target. In this chapter we shall examine some of these patterns, the characteristics of the signals they generate.

It is unfortunate that the general subject of optical modulation is so heavily obscured by the curtain of military classification. A search of the unclassified literature reveals that there are pitifully few papers dealing with the design, analysis, and comparative performance of reticles. It is all too apparent that most authors simply take the easy way out and automatically classify any paper dealing with this subject. The absurdity of this approach is clearly evident from a study of the patent literature, in which the basic principles of reticle design and performance have been described in detail during the last three or four decades. For example an application filed in 1934 resulted in a patent being issued to Zahl[3] in 1946 on the art of locating objects by their heat radiation. Even today, the teachings of this patent provide a remarkably lucid description of the problems involved in infrared search systems and the means of solving them. Much of the material in this chapter is taken from the patent literature in order to permit a detailed coverage of modern reticles and to forestall the inevitable cries of those who mistakenly think that the entire subject is classified.

5.1 OPTICAL FILTERING FOR BACKGROUND DISCRIMINATION

Whenever the spectral distribution of the flux from a target and its background are different, an optical filter is an inexpensive means of providing some rejection of the unwanted signals from the background. Although it is not a modulator in the strictest sense of the word, a filter is used with a reticle in almost every infrared system. Its primary purpose is to define the spectral bandpass of the system, but at the same time it can also effectively supplement the background rejection capabilities of the reticle. Subject, of course, to other restrictions, we attempt to choose a filter having a high transmittance for the flux from the target and a low transmittance for the flux from the background.

To understand the principles involved, consider the factors that influence the choice of a spectral bandpass for a system that is to detect jet aircraft in the presence of reflected sunlight and thermal radiation from the surface of the earth. The apparent spectral radiance of typical terrain is shown in Figure 3.15. This curve has a double peak, one at short wavelengths, which is due to reflected sunlight, and one at long wavelengths, which is due to thermal emission. Because the rather broad minimum between these two peaks occurs at about 3.5 μ, one is led to conclude that a system operating in this spectral region should have minimum interference from sunlit terrestrial backgrounds. The spectral distribution of the flux from a turbojet has its maximum in the 3.5 to 4 μ region so that the ratio of target-to-background flux has a broad max-

imum around 4 μ. The atmospheric transmittance curve in Figure 4.1 shows that this maximum lies within the 3.2 to 4.8 μ window—a particularly convenient act of nature, to provide a window so well placed for a very common detection task. The theory of optimum spectral filtering has been described by Eldering[4].

When the temperatures of the target and of its backgound are nearly the same, spectral filtering is of little help in discriminating one from the other and detection is dependent on there being an adequate radiation contrast.

The effect of contrast can be readily appreciated from the following experiment. A vehicle, such as a truck, is parked in an open field so that it can be viewed by an appropriate infrared system. When observations extend over a period of 24 hours, marked variations are found in the contrast between the vehicle and its background. In the afternoon the vehicle has been thoroughly heated by the sun and is warmer than the background, so that the contrast is positive. Because of its large thermal capacity, the vehicle cools more slowly than does the background during the early evening hours and the contrast is even greater than that observed during the afternoon. As the night progresses, the vehicle cools more rapidly and eventually becomes colder than the background, causing the contrast to pass through zero and become negative. After sunrise the background warms more rapidly than does the vehicle and the negative contrast is enhanced. By midmorning the heating of the vehicle by the sun is sufficient to again cause a period of zero contrast before the value finally becomes positive. The point to remember is that *for many targets and background combinations, there are two intervals in any 24-hour period during which the target cannot be detected because there is insufficient radiation contrast between it and its background.* Such effects, often called *washout*, have been observed with vehicles, structures, and roads[5]. No amount of spectral filtering or optical modulation can eliminate the possibility of such a washout; one must either wait for it to pass or employ some other means of detection.

5.2. THE USE OF RETICLES FOR BACKGROUND SUPPRESSION

The increase in target-to-background ratio from spectral filtering is rarely sufficient to render system operation independent of background conditions. For a system operating in the 2 to 2.5 μ atmospheric window, for instance, the irradiance on the detector due to sunlight reflected from clouds in the background may be 10^4 to 10^5 times that due to a distant turbojet target. Fortunately, reticles can provide this magnitude of background suppression.

The use of a reticle to increase the detectability of a particular target in the presence of extraneous background detail is called *space filtering*. Most targets of interest have the common characteristic that they are much smaller in angular extent than are objects in the background; a turbojet in front of a sunlit cloud, a ship against the sea, and a vehicle against terrain are typical examples. For such detection tasks, space filtering is used to enhance the signal from objects of small angular extent and to suppress signals from objects subtending large angles.

A simple example of space filtering by a rotating reticle is shown in Figure 5.1. The reticle pattern consists of a series of fan-shaped segments, alternately transparent and opaque. The reticle is placed at the image plane of the optics, and its center is coincident with the optical axis. The target and a (background) cloud illuminated by sunlight would normally be imaged on the reticle. They are shown to the side in Figure 6.1 in order that the action of the reticle can be seen more readily. As the reticle rotates at high speed about the center of the pattern, imagine that it moves slowly to the right so that it passes across the images of the target and the cloud. As the reticle passes across the target image, the image is chopped, that is, it is alternately transmitted and blocked by the

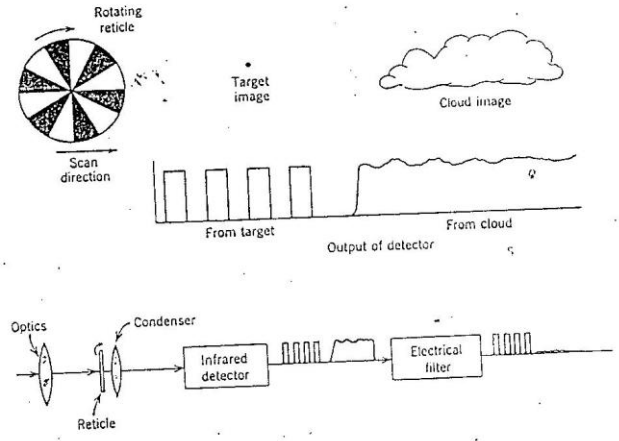


Figure 5.1 Space filtering by a rotating reticle.

elements of the pattern. Since the openings in the reticle are of approximately the same size as the image of the target, the electrical signal from the detector is a series of pulses at the chopping frequency f_c.

$$f_c = nf_r. \qquad 5\text{-}1$$

where n is the number of pairs of clear and opaque segments in the reticle and f_r is its rotational frequency expressed in revolutions per second.[1] As the reticle passes across the relatively large cloud image, the image covers several segments of the reticle pattern at any given instant of time. As a consequence, the irradiance on the detector is increased but there is very little chopping action on the cloud image. With the target and cloud both imaged on the reticle, the output of the detector consists of a large dc signal with a small ripple from the cloud and a pulsed signal from the target. When these are amplified and passed through an electrical filter with its passband centered at the chopping frequency, only the ac signals remain and the effect of the cloud is suppressed.

simplified example. Since most cloud edges are irregularly shaped, they will undergo some chopping action. This is responsible for the ripple on the otherwise constant signal from the cloud in Figure 5.1. As the range to the target increases, the chopped signal from the target eventually becomes less than that from the cloud edges and the system becomes *background limited*. Most of the early infrared systems working in the 2 to 2.5 μ atmospheric window were background limited in the presence of sunlit clouds. Even the most advanced reticle techniques were not efficient enough to remove this limitation except at short target ranges. As new detectors became available and it was possible to use the 3.2 to 4.8 μ window in order to escape more of the reflected solar radiation, sunlit clouds ceased to be a problem for most infrared systems.

5.3 THE USE OF RETICLES TO PROVIDE DIRECTIONAL INFORMATION

Many infrared systems are designed to detect and track targets of interest. In this category are seekers for guided missiles, star trackers for navigational purposes, and fire control systems[11]. In such systems the reticle is used to modulate the incident flux with information that can be used to determine the direction of the target. Since backgrounds are likely to be a problem with such systems, the reticle must provide good background rejection together with the directional information. Reticles developing amplitude (AM), frequency (FM), and pulse (PM) modulation are found in contemporary infrared equipment[11].

Rotating Reticles

The two-sector reticle shown in Figure 5.2 is one of the simplest ways to provide directional information[13–16]. This reticle, which has one transparent and one opaque sector, rotates about the optical axis of the system. At (*a*) the image of a target is shown slightly to the left of the center of the reticle. The detector output is shown to the right; it is a train of pulses at the chopping frequency, which, in this case, is identical with the rotational frequency of the reticle. At (*b*) the target is shown with the same radial displacement, but the angular displacement has been increased. The output of the detector is similar to that obtained before except that the pulses have been displaced along the time axis. The relative phase shifts δ_1 and δ_2 of the pulses are seen to be proportional to the angular displacement of the target. It is necessary to provide a phase reference so that the phase shifts can be measured. One of the simplest phase-reference generators is a small magnet fastened to the

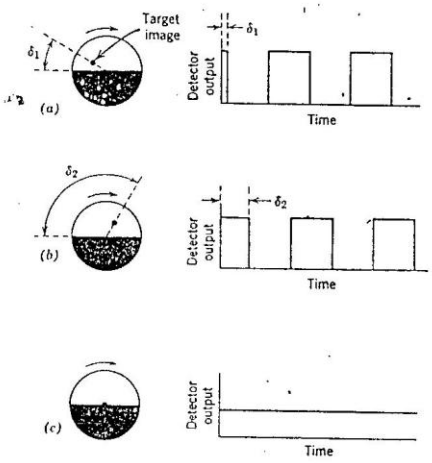


Figure 5.2 Simple two-segment reticle for generating target directional information (adapted from Carbonara et al.[15]).

periphery of a reticle with one or more pickup or *pip coils* mounted on a fixed frame near the path of the magnet [16]. Each time that the magnet passes the coil, a sharp pulse or *pip* is generated that can be used as a stable reference from which phase can be measured.

When the system of Figure 5.2 is pointed directly at the target, so that there is no pointing error, the target image lies at the center of the reticle and there is no chopping action, as shown at (*c*). As a result there is no output signal generated with zero pointing error—an unfortunate condition, since it is indistinguishable from that found when there is no target in the field of view. The reticle generates a chopping frequency analagous to the *carrier frequency* in a communications system. The carrier is phase modulated with information concerning the direction to the target. At zero pointing error the carrier disappears, taking with it the phase information. One way to prevent this is to use a double modulation scheme, such as that shown in Figure 5.3, so that a carrier is present whenever a target is in the field of view [16, 17]. A second reticle, with a

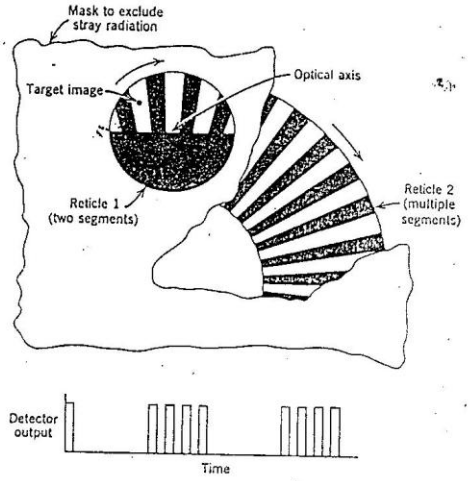


Figure 5.3 Double modulation arrangement to prevent loss of carrier with zero pointing error (adapted from Robert and Deslandes [16] and Chitayet [17]).

large number of segments, is placed immediately behind the two-segment reticle. It produces a high-frequency carrier regardless of the position of the target within the field of view. The two-segment reticle pulse modulates the carrier, and the phase of the pulses contains the information on target direction. The carrier and the modulation frequencies thus generated can be separated after amplification by passing them through filters tuned to the respective frequencies. Since the amplitude of the carrier signal is proportional to the irradiance from the target, it can be used for radiometric purposes, for automatic gain control, or for operating a device to indicate the presence of a target. Since the opening in the two-segment reticle is relatively large, its background rejection capability is poor. The addition of the second reticle gives a combination having quite good background rejection.

A design combining the merits of the reticles shown in Figures 5.1 and 5.2 was developed by Biberman and Estey[18]. Using the simple fan-bladed chopper shown in Figure 5.1, they made a systematic study of the signals it produced when scanning various types of sky backgrounds. They found strong signals from radiance gradients at the rotational frequency of the reticle and at its first few harmonics. The signals generated by clouds were similar but somewhat richer in harmonics. In no case did they find background signals beyond the eighth harmonic. By contrast, the signals from distant small targets, such as aircraft, were found to have strong harmonics to well beyond the twentieth.

On the basis of these observations, Bibermen and Estey designed the "rising sun" reticle, shown in Figure 5.4, so that it would generate a carrier frequency at least eight times higher than the reticle rotation frequency. Their reticle consists of two semicircular sectors. One contains alternately transparent and opaque fan-shaped segments for target sensing and background suppression. The other sector is semitransparent

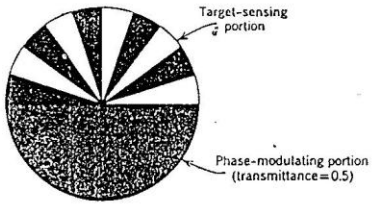


Figure 5.4 Reticle with good rejection of sky backgrounds (adapted from Biberman and Estey[18]).

and has a transmittance of 0.5. It provides the phase modulation that indicates target direction. The carrier frequency f_c generated by this reticle is

$$f_c = Knf_r, \qquad 5\text{-}2$$

where n is the number of pairs of clear and opaque segments in the target-sensing portion of the reticle, f_r the reticle rotational frequency, and K the reciprocal of the fraction of the total area of the reticle occupied by the target-sensing portion. In Figure 5.4 the value of K is 2 and n is 5 so that the carrier frequency is 10 times the reticle rotational frequency.

Biberman and Estey were the first to recognise that in a reticle of this type the transmittance of the phasing sector should be equal to 0.5. Since the average transmittance of the target-sensing portion is also 0.5, the reticle is balanced; that is, the transmittance of the two sectors is identical for images of large area. Figure 5.5 shows the detector outputs

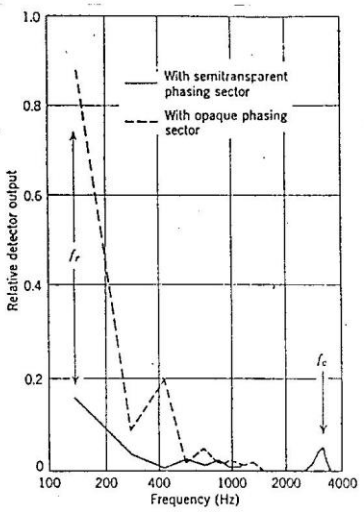


Figure 5.5 Comparison of the frequency spectrum of signals generated by reticles having semitransparent and opaque phasing sectors (adapted from Biberman and Estey[18]).

caused by a sky background and by a distant aircraft, with two different reticles. In one reticle the phasing section was opaque; in the second reticle it had a transmittance of 0.5. It is evident from Figure 5.5 that the maximum amplitude of the background signal from the reticle with the opaque phasing sector is five or six times that from the reticle with the semitransparent phasing sector. On the other hand, the target signal (at the carrier frequency) has about the same amplitude with either reticle. Thus the change from an opaque to a semitransparent phasing sector increases the ratio of target-to-background signals by a factor of 5 or 6. An amplifier having a narrow bandpass centered about the carrier frequency will further reject the background signals. Note, however, that this design, like that shown in Figure 5.2, generates no carrier for zero pointing error.

The "rising sun" reticle generates an amplitude modulation as the target image moves out from its center that is due to the changing relationship between the size of the blur circle and the size of the openings in the reticle. Thus the amplitude of the modulation indicates the radial coordinate of the target position and the phase of the modulation indicates the angular coordinate.

When the reticle pattern consists of segments bounded by straight lines, there is a tendency for the reticle to generate a larger signal when chopping a line image that when chopping a point image. This is a particularly undesirable characteristic if the horizon should appear in the background. Davis[20] has described an improved pattern for the target-sensing sector in the Biberman and Estey reticle. As shown in Figure 5.6, it

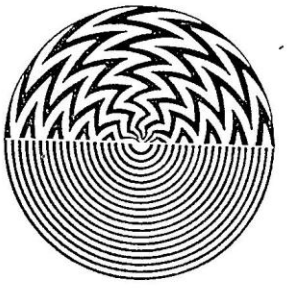


Figure 5.6 Reticle pattern having improved rejection of straight-line background elements.

consists of a series of zigzag elements bounded by portions of curved lines spiraling outward from the center of the reticle. The reticle is balanced since the average transmittance of either portion is 0.5. Davis claims that this pattern increases the rejection of spurious signals from cloud and horizon backgrounds. However, his performance curve does not support this conclusion since it is identical to that given by Biberman and Estey to support the claims for their reticle.[1] Aroyan and Cushner[21] have described the usefulness of involute patterns for the rejection of signals from straight-line sources.

Another reticle pattern that generates modulation indicative of both the angular and radial coordinates of target position[22] is shown in Figure 5.7. In the target-sensing portion the pattern is formed from curvilinear strips that originate along a diameter of the reticle and have equal arcuate

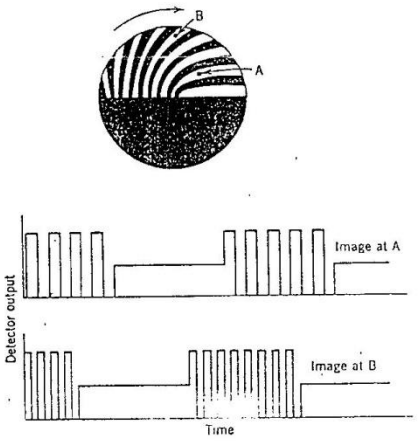


Figure 5.7 A reticle for generating both frequency and phase modulation (adapted from Lovell[22]).

radial angular

widths at equal distances from the center of the reticle. The carrier frequency generated by this pattern is a function of the radial position of the image. In Figure 5.7 two images are shown on the reticle. For the image at A, the carrier frequency is 10 cycles per reticle revolution; for the image at B the frequency is 16 cycles per reticle revolution. The output of the detector is shown in the lower part of the figure for both image positions. As before, the phasing sector pulse modulates the carrier so that the phase of the pulses indicates the angular position of the target. In addition, the carrier is frequency modulated to indicate the radial position of the target.

The basic fan-bladed pattern can be modified as shown in Figure 5.8 so as to generate a frequency-modulated carrier[23, 24]. The angular width of the individual elements of the pattern varies as a sinusoidal function of the azimuth angle around the reticle. Other reticles that generate frequency modulation

Two reticles, placed one behind the other, can be used to scan a circular field of view[29, 30]. Usually one reticle carries a single spiral slit and is rotated in front of a fixed reticle carrying one or more straight slits.

Figure 5.9 shows a reticle that generates pulse width and phase modulation. The pattern consists of triangular-shaped elements with the bases of the triangles alternately toward or away from the center of the reticle. An image on the optical axis is chopped into a series of pulses. If the image moves vertically from the optical axis, the pulse width varies in proportion to the distance from the axis. If the image moves horizontally, it causes a change in the phase of the pulses. Unlike the previous reticles, the modulation generated by this reticle is proportional to the cartesian rather than to the polar coordinates of target position. The equally spaced slots around the periphery of the reticle are used to generate a phase reference. A photodiode viewing a small source, such as a pilot lamp, through these slots is an effective phase-reference generator. Merlén[31] has described

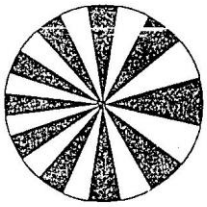


Figure 5.8 A reticle for generating a frequency-modulated carrier.

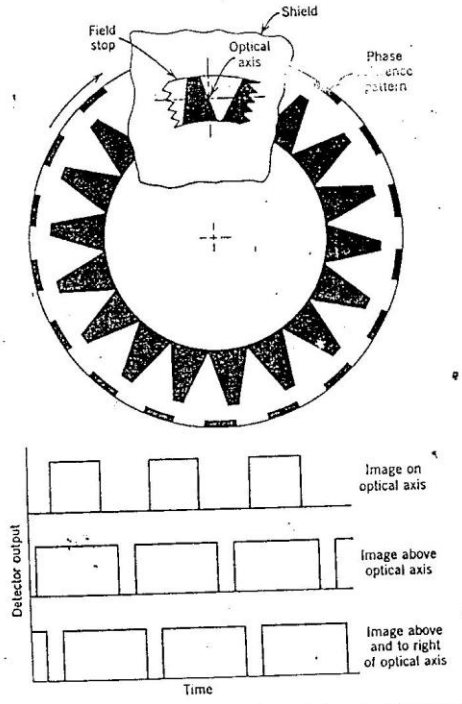


Figure 5.9 A reticle for generating both pulse-width and phase modulation (adapted from Merlen[31]).

a similar reticle in which the triangular elements contain much smaller fan-shaped segments. The number of segments in alternate triangular elements is in the ratio of 2 to 1. The output of the detector consists of bursts of alternately high- and low-frequency pulses. The smaller chopping segments improve the background rejection capabilities. A shield in front of the reticle carries an aperture that forms the field stop for the optics. Merlen claims improved background rejection by the use of sawtooth edges on the field stop.

A reticle that divides the field of view into a number of concentric zones and generates a different carrier frequency for each zone is shown in Figure 5.10, along with its rather unusual detector array [32, 33]. The detectors are made in the form of 10 rings concentric about the center of a flat disk. A narrow space between rings ensures the electrical isolation of each detector. The reticle, which is just in front of the detectors, has an opaque phasing portion and a target-sensing portion consisting of 10 concentric rings. The number of pairs of segments in each ring are in the ratio of 20 to 22 to 24, and so on, with the smallest number being in the innermost ring. The radius of each ring is chosen so that all rings have the same area.

If the reticle rotation rate is chosen so that the chopping frequency of the inner ring is, for instance, 1000 Hz, the second ring will generate 1100 Hz, the third 1200 Hz, and so on. Thus the chopping frequency is a function of the radial position of the image. All detectors are connected in parallel to a single amplifier. After amplification, a bank of 10 bandpass filters, each tuned to one of the frequencies generated by the reticle, is used to separate the signals. This arrangement can be used to track several targets simultaneously because no confusion exists unless more than one target occupies the same ring of the reticle.

The concentric multielement detector is probably not practical because of the difficulty of manufacturing it because the concentric detectors contribute nothing to the modulation process, they can be replaced by a single detector. Such a single detector would have to be as large as the reticle. Because it is desirable to keep the detector small in order to minimize its noise, it would be necessary to use an optical condenser along

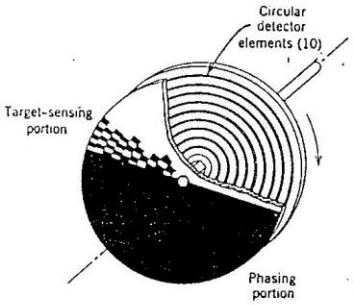


Figure 5.10 A ten-frequency reticle and multielement detector combination (adapted from Shapiro [32]).

with the smaller detector.

Stationary Reticles

Thus far the discussion has been limited to *rotating reticle systems*. It is just as feasible to rotate the image optically with respect to a fixed reticle. Such *stationary reticle systems* offer additional flexibility in the types of modulation that can be produced and they offer the important feature that there is no loss of carrier for zero pointing error.

Rotation of the image is called *nutation*;[1] it can be accomplished in several ways[6, 13, 15, 37–40, 47–50]. A common method is shown in Figure 5.11. The lens is mounted so that it can be rotated about an axis normal to and passing through the center of the reticle. The lens, however, is displaced laterally a distance d so that the optical and rotational axes are parallel but not coincident. When the lens is rotated, the image follows a circular path called the *nutation circle*. When the pointing error is zero, the nutation circle is concentric with the rotational axis and with the center of the reticle. The linear diameter of the nutation circle is just twice the distance d by which the optics are displaced. The angular subtense of the nutation circle is

$$\delta_n = 2000\frac{d}{f}, \qquad 5-3$$

where δ_n is expressed in milliradians, d is the displacement of the lens, and f is the equivalent focal length measured in the same units as d. In a typical system using a lens having a focal length of 4 in. (10 cm), a nutation circle subtending 35 mrad (2 deg) would require that the lens be

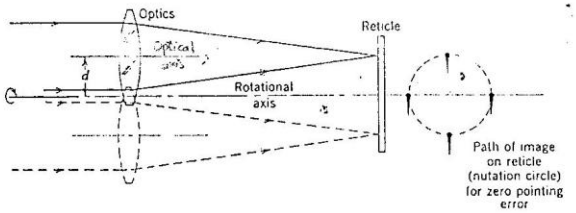


Figure 5.11 Rotation of decentered optics to produce nutation.

displaced by 0.070 in. (0.175 cm). The designer may put the required offset either in the metal mounting or in the lens. In the first case, the edge of the lens is ground concentric with its optical axis and the metal mount is machined eccentrically to provide the offset. Alternatively, the edge of the lens is ground so as to be concentric with the rotational axis, thus eliminating the need for an eccentric mount. Since the rotational speed of the lens, the *nutation frequency*, may be quite high, it is usually necessary to dynamically balance the mounted optical assembly. Although Figure 6.11 shows a lens, the principle is equally applicable to reflective optics[47].

Nutating systems are often referred to as *rotating field systems*. In Figure 5.11 the image is shown at several points around the nutation circle; we see that its spatial orientation does not change. Instead it undergoes a simple translation around the center of the nutation circle. Thus the designation of a rotating field is a misnomer if this term is taken literally.

When a pointing error exists, the nutation circle is no longer concentric with the center of the reticle. Turck [6] has used this effect in the tracking arrangement shown in Figure 5.12. A simple reticle with fan-blade segments is used for illustration; the reader will recognize that many of the patterns previously discussed can also be used. At (*a*) there is no pointing error. The image travels around a nutation circle that is concentric with the center of the reticle.

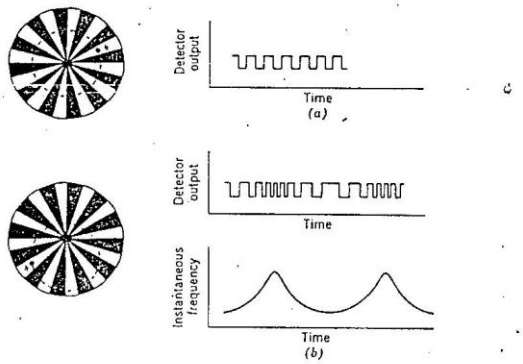


Figure 5.12 Nutating system generating frequency modulation (adapted from Turck [6]).

5.4 TRACKING SYSTEMS WITHOUT RETICLES

By using multiple-element detectors it is possible to build tracking systems that do not use reticles. The basic principles involved have been known since the 1920's, or earlier [46]. One such system, using a four-element detector [47–49], is shown in Figure 5.13. Rotating optics are used with a detector consisting of four rectangular elements arranged in a cross-shaped pattern. In the absence of a pointing error, the center of the nutation circle coincides with that of the detector array. The image crosses the four elements at equal time intervals, and the detector output is a series of pulses occurring at a constant rate. With a pointing error, the nutation circle is not concentric with the center of the array and the time intervals between crossings of successive detector elements are no longer constant. By comparing these time intervals with reference signals derived from the rotating optics, the retangular components of the target

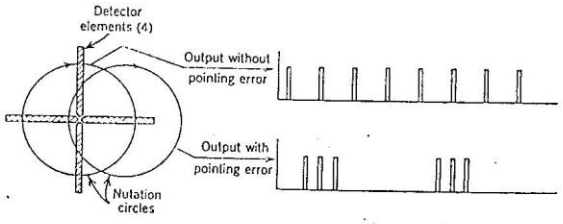


Figure 5.13 A pulse-modulation system without a reticle (adapted from Buntenbach [47]).

coordinates can be determined. Thus the combination produces a pulse position modulation in which variations in pulse position indicate the direction of the target. Similar performance can be obtained from a two-element array in which the elements are arranged in an L-shaped pattern (equivalent to removing one vertical and one horizontal element in Figure 5.13)[47, 50–52].

The background rejection characteristics of such pulse systems are excellent. Since the motion of the target image is across the short dimension of the detector elements, this dimension should meet the same criterion used in determining the width of reticle openings; that is, the width of the element should be from one to three times the diameter of the blur circle of the optics. Each pair of detector elements (those lying in the same straight line) can be oppositely biased so as to cancel signals from large straight edges, such as the horizon. Unfortunately, good rejection requires that the edges be accurately parallel to the detector elements, a condition difficult to maintain.

Tiny mirrors mounted on the tines of electrically driven tuning forks have been used to provide chopping or image motion over a reticle[53].

5.5 COMMENTS ON RETICLE DESIGN

At no other point in the design of an infrared system is there likely to be as large a gulf between the analytical and the hardware people as there is in the area of reticle design and comparative performance. At first glance the subject appears to be ideally suited to the mathematical approach. As one proceeds, more complicated mathematical techniques are required, and more than one promising design has become lost in an elegant set of equations having no solution.

If one is interested only in target detection and not in its direction, the analytical approaches described by Aroyan[7] are excellent. Supplementing these with probability distributions of background radiances and their spatial distributions can lead quite directly to very efficient filtering. In a tracking system the designer is vitally interested in the *error response* of the system, that is, the relationship between pointing error and the error signals available at the output of the signal processor. The particular circuitry chosen for the signal processor and the tracking servo places strong demands on the choice of a reticle and the modulation it generates. It is extremely unlikely that a reticle designed by purely mathematical methods will meet all of these diverse requirements. Such designs do, however, offer a good starting point from which the design can be empirically modified to one having an error response curve of the desired shape and slope while still retaining good background rejection, good resolution against multiple targets, and the desired field of view. The exact procedures for refining a reticle design are usually regarded as proprietary by the organizations involved, and therefore each designer must develop his own solution to the problem. Graphical analysis, simulation techniques[54], and tests of the proposed reticle in the final system have proved to be useful approaches.

As shown in Section 5.3, typical tracking systems may employ a rotating reticle, a stationary reticle and nutating optics, or a combination of these to generate amplitude, frequency, or pulse modulation. Unfortunately there is very little information available in the open literature on the comparative performance of these combinations[23, 24, 55, 56]. For moderate signal levels, an FM system usually has a higher signal-to-noise ratio than an AM system. When the signal level drops so low that the limiter no longer functions, the performance of an FM system becomes inferior to that of an AM system. Therefore, if obtaining the maximum possible tracking range is the principal system requirement, it can probably best be met by the use of AM.

rotating-reticle systems in the presence of a high-radiance background has not been mentioned previously. Even though the optical condenser focuses the image of the entrance aperture on the detector, there is still an out-of-focus image of the reticle on the detector. Since it is likely that the sensitivity of the detector will vary across its surface, the out-of-focus image of the rotating reticle may generate spurious signals when the system views a high-radiance background. This difficulty is not encountered with a stationary reticle system since there is no relative motion between reticle and detector.

If several targets are in the field of view, most reticle systems will track the target that has the highest radiant intensity. If the targets have about the same radiant intensity, most reticle systems will track the effective radiation centroid. This can lead to serious difficulties if, for instance, the system is to be used to guide a missile against multi-engine aircraft. Proper design of the reticle pattern or of the shape of a superimposed radial transmission gradient can force the system to select one target rather than continuing to track the centroid [57]. An infrared guided missile fired toward the forward aspect of a turbojet target may consistently miss because its seeker does not see the hot tailpipe but, instead, tracks the centroid of the exhaust plume. Pulse modulation systems offer a possible solution because they can readily track the edge of a radiating source.

TABLE 5.1 A COMPARISON OF SEVERAL TYPES OF RETICLES FOR TRACKING SYSTEMS

Type	Type of Modulation	Advantages	Disadvantages
Rotating reticle, fixed optics	AM	Simple mechanical construction	Loss of carrier in absence of pointing error
		Adequate discrimination against low-radiance backgrounds (when combined with optical filtering)	Relatively poor discrimination against high-radiance backgrounds
	FM	Same as for AM	Same as for AM Wide electrical bandwidth
Stationary reticle, nutating optics	AM	Excellent discrimination against high-radiance backgrounds	Moderately wide electrical bandwidth
		No loss of carrier with zero pointing error	Shift in apparent position of target for a large image size
		Relatively simple mechanical construction	
	FM or FM/AM	Same as for AM	Reversal of FM signal when image crosses center of reticle Abrupt shift in apparent position of target for a large image size Wide electrical bandwidth
Rotating reticle, nutating optics	FM	Good discrimination against high-radiance backgrounds	Highly complex mechanical construction
		No loss of carrier with zero pointing error	Wide electrical bandwidth

5.6 FABRICATION OF RETICLES

In many systems the reticle, or a mask placed immediately in front of it, acts as a field stop. The diameter of the field stop is

$$d = 0.0174 f\beta, \tag{5.4}$$

where d and f are measured in the same units and β is the instantaneous field of view measured in degrees. This expression gives an error of less

be well polished, their surfaces should be flat, parallel, and free of scratches and pits, and they should be edged to the desired reticle diameter. An evaporated metallic coating, usually of aluminum, is applied to one surface of the blank. Next a thin layer of photoresist is applied over the metallic coating. The prepared blank is placed in contact with the master negative and exposed through it to the ultraviolet from a mercury arc. Extreme care is required to ensure that the blank is held concentric with the pattern on the master negative in order to meet the tolerance limits on concentricity. Photoresist is a photosensitive material that is polymerized upon exposure to ultraviolet so that is is not affected by the acids used for etching. After the exposure the blank is washed to remove the photoresist from those portions of the pattern that were protected from the ultraviolet by the master negative. Immersion in an etchant removes the metal film from these areas. The final result is a reticle pattern etched in a metal film with no material in the clear spaces other than the substrate. With careful attention to detail in every step of the process, pattern openings as narrow as 0.00008 in. (2 μ) can be reproduced. The edges of such openings show no irregularities, even when viewed under a 900-power microscope.

It is not practical to apply an antireflection coating to the substrate exposed in open spaces of the reticle pattern, but the second surface of the substrate can be coated. Reflection losses can be entirely eliminated by etching the reticle pattern into a thin metal foil and then cementing it to a

Chapter Three Sources of Infrared Radiation

3-1 Introduction

: This chapter emphasizes practical details concerning radiation from sources that are of interest to an infrared system designer. Considered first is the design of blackbody-type sources for use in calibration and the peculiar problems arising from the lack of a traceable national standard of blackbody radiation. Sources useful in the laboratory, and others such as turbojets, rockets, vehicles, and personnel, which are often the targets for infrared systems.

3-2 Blackbody-Type Sources

: Blackbody-type sources are widely used for the absolute calibration of infrared equipment. Before discussing the various theoretical and practical factors to be considered in their design and use, a plea should be made for proper terminology. A blackbody represents a theoretical concept; that is, it is an ideal thermal radiator to which all others can be compared. by its very definition, we cannot hope to build a blackbody. Any one who must design blackbodies for highly accurate calibration purposes will wish to supplement this material with a study of some of the additional

references to be mentioned later. on the assumption that the walls are diffuse reflectors, Gouffé finds that the effective emissivity of a cavity is

$$\epsilon' = \frac{\epsilon (1+K)}{\epsilon (1 - A/S') + A/S},$$

where

ϵ' = effective emissivity of the cavity

ϵ = emissivity of the cavity walls

A = area of the opening through which radiation leaves the cavity, cm^2

S' = total surface area of the cavity, including that of the opening, cm^2

$K = (1 - \epsilon)(A/S' - A/S_0')$

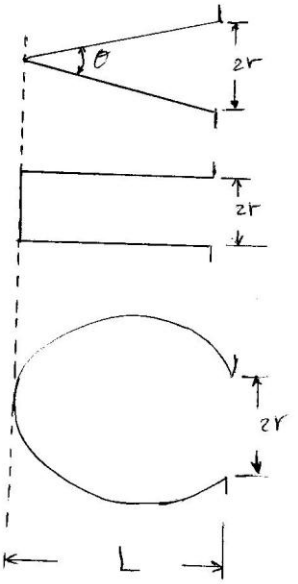
S_0' = surface area of a sphere whose diameter is equal to the depth of the cavity (measured from the plane of the opening to the deepest point of the cavity).

it is convenient to write as

$$\epsilon' = \epsilon_0 (1+K)$$

The numerical value of K is small; for a spherical cavity K is equal to zero and the effective emissivity is equal to ϵ_0.

Three typical cavity configurations are shown in figure below. They are more easily characterized by the depth of the cavity L and the diameter of the opening 2r than by the areas in eq $\epsilon' = \frac{\epsilon(1+k)}{\epsilon(1-A/S') + A/S'}$. The values of L and 2r are identical in each of the configurations shown.



Typical cavity configurations for blackbody-type sources.

Note that for a sphere, L is not equal to the diameter since it is measured normal to the plane of the opening to the deepest point of the cavity. The cavity effect is clearly evident; the effective emissivity of a cavity always exceeds that of its surfaces. As the emissivity of the surface decreases, the cavity effect becomes increasingly evident.

Another important result from the analysis of Goulfé is that for a given value of L/r, the cavity with the largest surface area has the highest effective emissivity.

3.3 General Purpose Sources of Infrared

Several sources are available for use in system checkout and alignment, spectrometers, communications devices, and Solar Simulators. In general the characteristics of these sources are not as well known as those of a blackbody. but their intended applications do not require such detailed knowledge. In addition, such sources are often relatively inexpensive readily portable, and simple to use.

(a) The Nernst Glower: A Nernst Glower is often found in infrared spectrometers that are used to measure the transmittance, reflectance, or absorptance of various materials it consists of a relatively fragile cylinder made by sintering a mixture of zirconium, yttrium, thorium and certain oxides. when cold, it does not conduct, but when it is heated to 400°C by a flame or self contained tungsten filament, it becomes conductive and can be further heated by passing an electrical current through it.

For an average glower that is one about 3 cm long and 0.15 cm in diameter, an input of 0.5 amp at 20 V is required after the initial heating. under these conditions the effective temperature of the glower is about 2100°K. Because of the large negative temperature coefficient of resistance, a current - limiting ballast is needed. The emissivity of the glower varies somewhat with wavelength and has an average value of about 0.6 from 2 to 15 μ.

(b) The Globar : Another source often used with infrared spectrometers is the Globar. it is a rod of silicon carbide, typically 5 to 10 cm long and 0.5 cm in diameter. It is heated to an operating temperature of about 1500°K by an input of 3 to 5 amp. at 50 Volt. the Globar needs no separate heater since the heating current is passed directly through the silicon carbide rod. The emissivity varies somewhat with wavelength and has an average value of about 0.8 from 2 to 15 μ.

© The Carbone Arc: A low-intensity carbon arc has been used as a spectrometer source when a greater radiance than that of the Globar or Nernst Glower was needed. A source temperature of about 3900°K is reached. A five fold decrease in emissivity occurs as the wavelength increase from 2 to 6 μ.

The high-intensity carbon arc, which operates at 5800 to 6000 °K is used in solar simulators. The arc current is Three to four times greater than that of the low-intensity arc and the operating life of the electrodes is proportionately less.

(d) The Tungsten Lamp: Tungsten lamps are used as sources, but only for the near infrared since their glass envelopes do not transmit radiant energy beyond 4 μ. Filament temperature as high as 3300°K can be obtained The average emissivity of a tungsten filament at 2800 °K is about 0.23 from 2 to 3 μ.

Tungsten lamps are surprisingly inefficient sources of visible light, Ten Percent of the input Power to a

typical 100 W household lamp is radiated beyond the bulb as visible light, 70 Per cent is radiated in the near infrared, and 20 Per Cent is absorbed by the gas in the lamp and by its glass envelope. The glass envelope can readily a temperature of 150°C. As a result, equipment operating in the intermediate and the far infrared may receive strong signals from tungsten lamps.

(e) The Xenon Arc Lamp: The xenon arc lamp has been used in near-infrared communication system. It is particular advantage is the ease with which the output can be modulated by varying the current supplied to the lamp. Most of energy from the xenon arc is radiated in the visible and ultraviolet, but there is a useful output in the near infrared, extending to a wavelength of about 1.5 μ.

(f) The Laser: The laser, an acronym for "light amplification of stimulated emission of radiation", represents an entirely new family of quantum electronic devices. Laser Provide coherent Sources of extremely high radiance in the Portion of the spectrum extending from the ultraviolet to microwaves.

Probably the first application of the laser in the infrared portion of the spectrum will be for communication systems that can exploit the laser's coherence, high radiance, and the ease with which it can be modulated

(9) The Sun : For convenience in calculations it is often assumed that the Sun radiates as a 5900 K Blackbody. however, show that for accurate calculation no single effective temperature can be assumed for the sun since the value appears to decrease with increasing wavelength. The irradiance at the surface of the earth is about two-thirds of this value (0.140 w/cm^2) or 0.09 w/cm^2. Since many infrared systems are designed to detect targets that produce an irradiance of 10^{-10} w/cm^2 or less, an inadvertent look at the sun may seriously overload or even permanently damage these systems.

3.4 Targets : Those objects that infrared system are designed to detect, the radiating characteristics of some targets are classified by the military, reasonably accurate estimates can often be made by applying the radiation laws.

① The turbojet engine
② The turbofan engine
③ The Boeing 707 jet Transport
④ After burning
⑤ The Ramjet
⑥ The Rocket engine
⑦ Aerodynamic heating
⑧ Personal
⑨ Surface Vehicles
⑩ Stars and Planets

Chapter Four Optics

4.1 Introduction: optical design is a highly complex subject, and the services of an experienced optical designer are essential to the project team responsible for the development of a new infrared system.

4.2 Refraction and Reflection: The velocity of light which is the velocity in empty space, that is, in a perfect vacuum; its velocity in any other medium is less. For a particular material and wavelength, the ratio of the velocity of light in a vacuum to that in the material is called its index of refraction. A graph of the index of refraction plotted as a function of wavelength is called a dispersion curve. Ordinary glass has an index of refraction of 1.5, thus the velocity of light in glass is two-thirds of what it is in a vacuum. optical materials useful in the infrared have indices ranging from about 1.3 to 4.

The index of air is about 1.00029, and the correction to vacuum conditions is very small and need only be applied to ultraprecise wavelength determinations. When a portion of the wave front passes through an optical system, its curvature is changed. Thus we could describe an optical system by the change it causes in the curvature of the wavefront. Since such changes are not easily visualized, the idea of rays has been introduced.

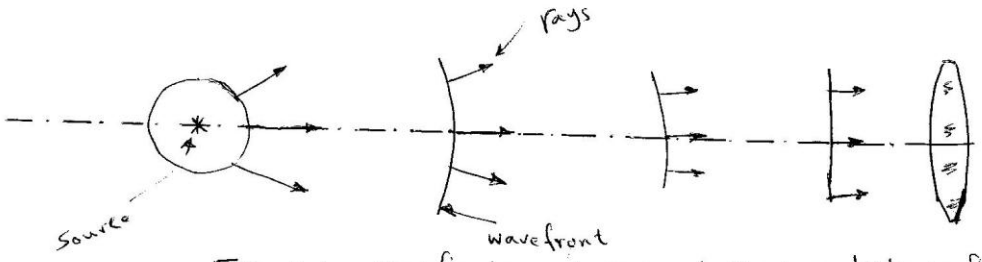


Fig 4.1 Wavefronts and rays at various distance from a source.

A rays is simply a normal to the wavefront and points in the direction in which the wave is moving when the radiant flux is considered as a stream of photons, the ray represents the path followed by a photon.

One of the Key assumptions in describing the Performance of an optical system is the use of a distant source or a source at infinity. These terms describe a condition in which the wavefornt entering the optics can be considered as a plane rather than as a portion of a sphere. Alternatively they mean that all of the rays entering the optics are parallel to one another. This effect is shown in figure (4-1) where several wave fronts and rays are moving toward a lens. If we look only at the portion of the wave front that the lens accepts, it is evident that the curvature of this portion decreases with distance from the source and that the rays tend to become parallel to one another. obviously, only when the source is at infinity is this portion of wave front truly plane and the rays parallel to one another. The collimator is a simple optical means of producing a source at infinity; it's widely used for testing infrared equipment

When aray meets the interface between two materials having different indices of refraction, it is split into two rays, as shown in figure (4-2), one ray is reflected and the other is refracted into the second medium.

Senll's law or the law of refraction describes the refracted ray $n \sin\phi = n' \sin\phi'$

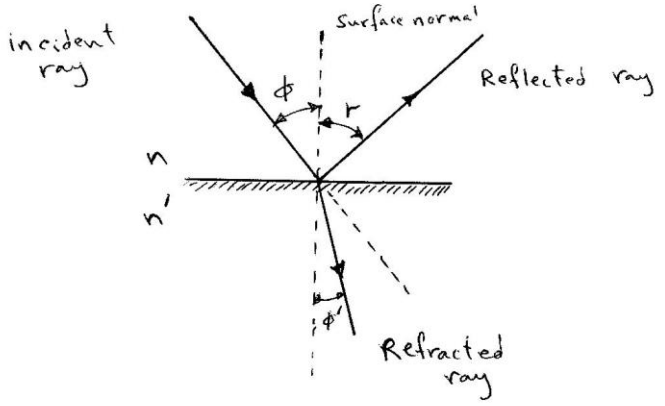


Figure (4-2) Reflected and Refracted rays

Where n and n' are the indices of refraction of the two media, ϕ is the angle of incidence, and ϕ' the angle of refraction. Thus the Product of the index of refraction and the Sine of the angle the ray makes with the normal to the Surface remain invariant across the interface.

The reflected ray is described by the law of Reflection $r = \phi$

This simply means that the angle of reflection equals the angle of incidence. In addition the incident, reflected and refracted rays are always coplanar (sometimes called the second law of Reflection).

With a refractive optical system, that is, one employing lenses, the reflected ray represents energy lost to the system and it is important to know the magnitude of this loss. An exact calculation requires knowledge of the angle of incidence, the index of refraction, and the polarization of the incident flux. However, for the special case of an unpolarized ray incident from air ($n=1$) and an angle of incidence of zero ($\phi = 0$), the fraction of the flux reflected by a single plane surface is

$$\rho_s = \left(\frac{n-1}{n+1}\right)^2$$

where ρ_s is the reflectance of the surface. The fraction of the flux transmitted by the surface is

$$\tau = 1 - \rho_s = 1 - \left(\frac{n-1}{n+1}\right)^2$$

In a plate having plane surfaces parallel to one another some of this flux is absorbed, some emerges from the second surface, and some is reflected back and forth between the surfaces of the plate.

considering all these effects, the transmittance of a parallel-sided plate is $\tau = \frac{(1-\rho_s)^2 e^{-ax}}{1-\rho_s^2 e^{-2ax}}$

where ρ_s is given above and a is the absorption coefficient, and x is the thickness of the material. There are several ways to simplify this expression. In many applications we can select an optical material having no absorption bands close to the spectral bandpass of the system. Under these conditions the absorption coefficient is very small in the spectral region of interest and the transmittance is approximately.

$$\tau = \frac{(1-\rho_s)^2}{1-\rho_s^2} = \frac{2n}{n^2+1}$$

Thus the transmittance of a germanium plate ($n=4$) is about 0.47. If the index of refraction is less than 1.9 or if reflection-reducing coatings have been applied so that the value of ρ_s is less than 0.1, the term $1-\rho_s^2 e^{-2ax}$ can be ignored and the transmittance is given with an error of less than 1 per cent by

$$\tau = (1-\rho_s)^2 e^{-ax}$$

under these conditions the absorption coefficient is 48 given with an error of less than 1 per cent by:

$$a = \frac{1}{x} \ln \frac{(1-f_s)^2}{T}$$

Let us examine more closely the relationship between the incident and refracted rays by rewriting as

$$\sin\phi' = \frac{n}{n'} \sin\phi ,$$

When the incident ray is in the medium having the lower index of refraction. $n/n' < 1$ and $\sin\phi' < \sin\phi$. Thus, even if the incident ray is at grazing incidence ($\phi \sim 90°$) there will still be a refracted ray. If the incident ray is in the medium having the higher index of refraction, $\sin\phi' > \sin\phi$, and there will not be a refracted ray when the incident angle exceeds a certain value. This occurs when $\sin\phi'$ is equal to unity and the refracted ray is tangent to the interface. For greater angles of incident there is no refracted ray, only a reflected one. This effect is called total internal reflection and the angle of incidence at which it occurs is called the critical angle ϕ_c where $\sin\phi_c = \frac{n'}{n}$

If the refracted ray is in air, this reduces to

$$\sin \phi_c = \frac{1}{n}.$$

Thus for a ray passing from germanium into air, the critical angle is about 14.5°, and all rays at higher angles of incidence are reflected back into the germanium.

4.3 Describing an Optical System

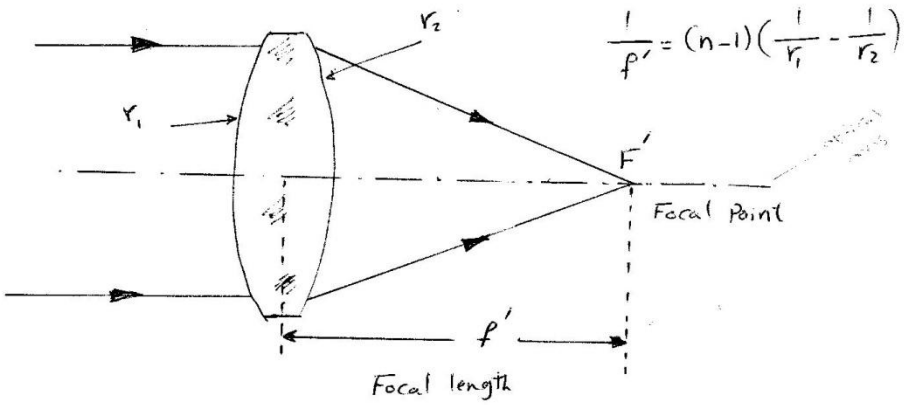
To the system engineer the purpose of the optics in his infrared system is to collect radiant flux and deliver it to the detector. Thus the optics are quite analogous to a radar antenna used to receive echoes from a target. He usually knows what field of view the optics must cover, the spectral region over which they will be used, any optical system consists of one or more reflecting or refracting elements. All elements are considered to be centered, that is, the centers of curvature of each of the surfaces all lie on the same straight line, called the optical axis.

Any departure from this condition because of poor manufacturing or careless mounting impairs the performance of the optics. Using various simplified formulas, the optical designer lays out a preliminary design and then examines it in detail to see how well it meets the specifications, his principal tool is ray tracing, that is, examining the paths followed by rays passing through the optics. This can be done by applying Snell's law, or the law of reflection, at each surface that the ray encounters. Such a procedure is inherently very accurate, and it is customary to retain figures to the fifth or sixth decimal place. Modern electronic computers can make such calculations at the rate of a few seconds per ray for a complex system.

Since Snell's law involves the sine of various angles, the ray-tracing equation can be simplified by replacing the sine by its series expansion

$$\sin\phi = \phi - \frac{\phi^3}{3!} + \frac{\phi^5}{5!} + \frac{\phi^7}{7!} + \text{-----}$$

A simple thin lens and concave mirror are shown in Figure (4-3). Rays from an object are brought to a focus at the point F'. An image of the object is formed on the focal plane, a plane that is normal to the optical axis and passes through F'. Higher-order theory shows that the focal plane may be a slightly curved surface rather than a plane. If the rays incident on the lens or mirror are parallel to the optical axis (from an axial object at an infinite distance), the focus at F' is on the axis and is called the focal point. The distance from the center of a lens (assumed to have negligible thickness) to the focal point is called the focal length and is the most important parameter used to describe a lens. Likewise, for a concave mirror, the focal length is the distance from the vertex of the mirror surface to the focal point. The formulas in Figure (4-3), show how the focal length can be computed. For a spherical mirror, focal length is equal to half the radius of curvature of the reflecting surface.



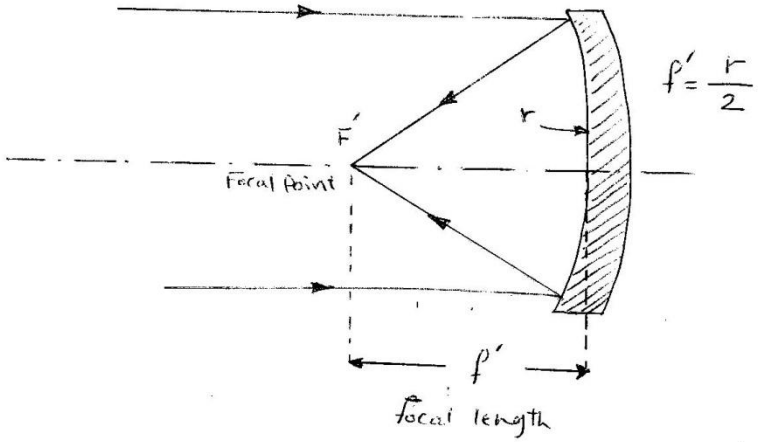


Fig 4-3 Focal length and Focal Point of a thin lens and of concave mirror.

* There are two important ways of describing the amount of radiant flux collected by an optical system. The first uses the f/number of the optics

$$\frac{f}{no.} = \frac{f}{D}$$

where f is the equivalent focal length and D is the diameter of the aperture stop or entrance pupil. The f/number concept is familiar to most photographers, who usually refer to it as the speed of the optics. It is unfortunate that the f/number is an inverse quantity; that is, the smaller the f/number the greater the radiant flux collected and the higher the speed of the optics.

* An alternative means of describing flux collection is the numerical aperture, which is given by

$$NA = n' \sin u'$$

where n' is the index of refraction of the medium between the final optical surface and the second focal point, and u' is the half-angle

of the cone of rays converging at the focal Point. The relationship between numerical aperature and f/number is

$$NA = \frac{1}{2(f/no.)}.$$

4.9 Factors affecting Image quality

Suppose that we use a microscope to view the image of a Point Source formed by a lens. Since a Point Source exists only in the world of mathematics, we must simulate one by using a star, a collimator, or an illuminated Pinhole Placed as far away as Possible. Examination of the Image Shows that it is a bright, somewhat diffuse disk and it is usually called a blur circle

* Two processes contribute to the size of the blur circle these are diffraction, which is a consequence of the wave nature of radiant energy and aberrations, which depend on the geometrical arrangement of the optical surfaces and on the dispersion of the optical materials.

* Aberration can be controlled by the optical designer, diffraction, on the other hand, is a Physical limitation over which he has no control, even in the absence of abeirations, diffraction still causes a Point to be imaged as a blur circle.

7.4.1

Diffraction: which is an interaction between a train of waves and an obstacle, is not peculiar to optics and is often observed in acoustics and in the operation of microwave antenna. In optical systems diffraction occurs at the edges of the optical elements and at the diaphragms used to limit the beam.

The image of a Point Source formed by diffraction-limited optics appears as a bright central disk surrounded by several alternately bright and dark rings. The distribution of radiant flux in the diffraction image is shown in fig 3-4. The central disk contains 84 Percent of the radiant

flux, and the rest is in the surrounding rings. Since Airy was on of the first to analyze the diffraction process, the central disk is usually called the Airy disk. The angular diameter of this disk, which is considered to be equal to the diameter of the first dark ring, is

$$\delta = \frac{0.244\lambda}{D}$$

where δ is expressed in milliradians, λ is in microns and the aperture diameter D is in centimeters.

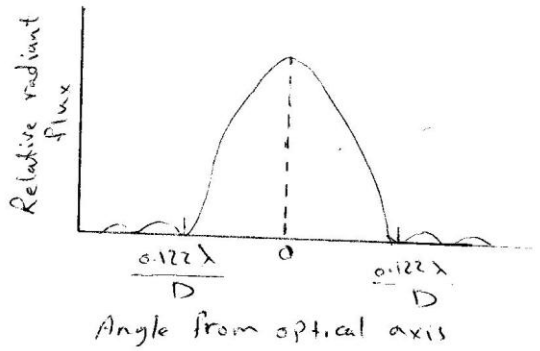


Figure 4-9 Distribution of radiant flux in the diffraction image.

4.4.2 Aberrations: Third-order theory predicts seven types of aberrations. Two of these called Chromatic aberrations, are caused by variation in the index of refraction of the lens material with wavelength. The rest, called monochromatic aberrations, occur even though only a single wavelength is involved. These seven aberrations can be briefly described as follows:

A- Monochromatic Aberrations

① Spherical - Rays from a common axial point which pass through the optics at different distances from the optical axis are not brough to a common focus.

② Coma - The image of an off-axis object is no longer symmetrical but becomes an enlarged, comet-shaped blur.

③ Astigmatism - The image of an off-axis point becomes a pair of lines that are at right angles to each other. The lines lie at different distances from the optics, and the smallest blur circle lies somwhere between them.

④ Curvature of Field – The Image of a Plane object lies on a curved rather than on a Plane Surface.

⑤ Distortion – straight lines, except those Passing through the center of the field, are Imaged as curved lines.

B. Chromatic Aberrations

① longitudinal – A variation in the Position of the focal Point as a function of wavelength.

② Lateral – The Size of an Image formed by the optics Varies as a function of wavelength.

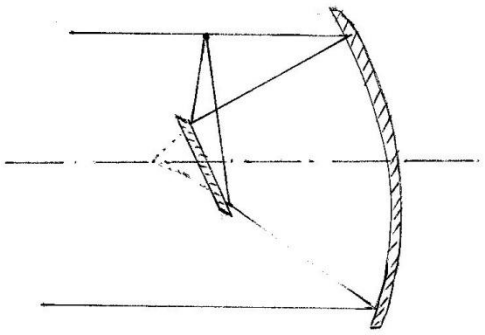
4.5 Typical optical systems for the Infrared

Having examined the limitations of simple lenses and mirrors, let us now consider some of the more complicated optical systems used in the infrared. Until recently most of these systems used mirrors rather than lenses because there were relatively few optical materials transparent in the infrared. This situation no longer exists, and one is as likely to find a lens as a mirror system used in a new design.

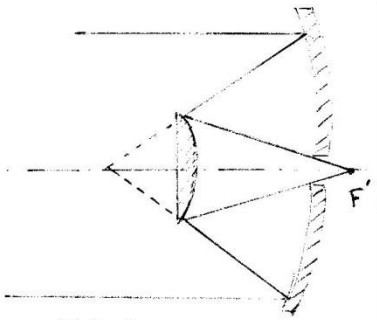
4.5.1 Reflective optics

Most of the mirror system have evolved from the classical reflective types developed by astronometers. Since the focus of a spherical or parabolic mirror lies in the direction of the incoming rays. Some of them must be blocked in order to place a detector at the focus. This is the prime focus and is rarely used except in large astronomical telescopes. Newton suggested that a flat secondary mirror be so placed as to bring the focus to the side of the telescope

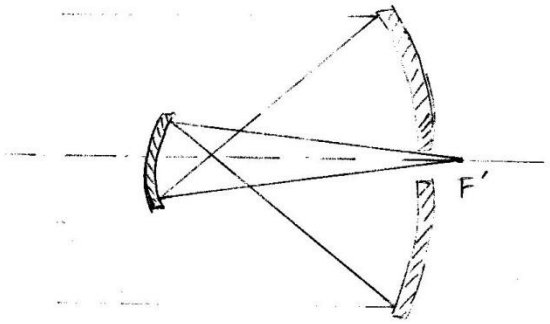
See fig (4-5a). This is a convenient location for the detector since it minimize blockage of the incoming rays. The Cassegrainian system fig (4-5-b) uses a convex secondary mirror placed inside the prime focus; it redirects the rays through a hole in the primary mirror to a new focus. The Gregorian syster Fig (4-5-c) is similar to the Cassegrainian, but it uses a concave secondary mirror placed outside the prime focus. In both of these cases the combination of the two mirrors has an effective focal length longer than that of the primary mirror. The Herschelian Fig (4-5-d) is occasionally used when it is desirable to minimize the number of optical surface in a system.



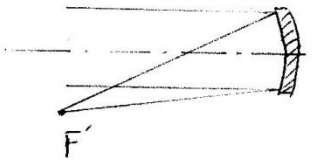
(a) Newtonian



(b) Cassegraian



(c) Gregorian



(d) Herschelian

Fig(4-5) Classical reflective optical systems.

4.6 Methods of generating Scan Patterns

Some infrared systems must seek a target by scanning a large search field. In essence, we start with a small instantaneous field of view and then find a way to move it so as to scan the search field completely. The frame time is the time required for one complete scan of the search field. Most of these systems generate a rectangular raster, that is a line-by-line scan of the search field formed in much the same way that a television picture is formed. The gimbaled optics can be moved in the desired scan pattern by a servomotor. Angular scanning rates as high as 250 deg sec^{-1} can be achieved. Means for mechanically generating a conical scan are described in Reference (Apparatus for Producing a conical Scan of Automatically Varying Apex Angle) by H. Blackstone. If mechanical means are not satisfactory for generating scan patterns, optical means can often be used.

They offer the advantages of higher angular scan rates better scan linearity, a variety of scan patterns, and reduced power consumption since the mass to be driven is usually much smaller. A means of generating a raster scan by a plane placed in front of the optical system is shown in fig (4-6). Continuous rotation of the mirror about the vertical axis provides a full 360 deg coverage in azimuth, and rotation about a horizontal axis provides elevation coverage. The scan element, which is the projection of the instantaneous field of view on the object plane, appears to spiral upward in the hemisphere above the system. Since the projected height of the mirror must be equal to the diameter of the optics, large elevation angles are impractical. Two-mirror system can be devised for scanning the entire hemisphere.

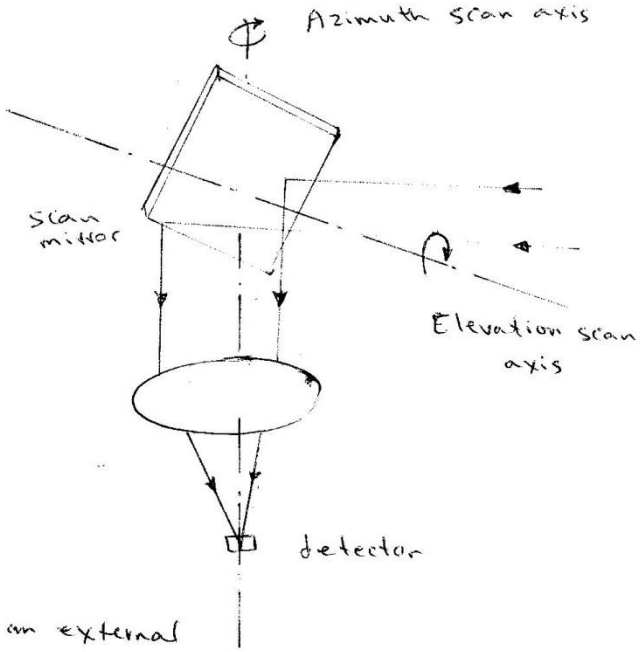


Fig (4-6) Use of an external mirror to generate a raster scan.

Another scanning means, the rotating eyeball, is show in fig (4-7). Four lenses rotates about a single fixed detector. A shield placed around the detector ensures that it "sees" only one set of optics at a Time and limits the azimuthal coverage to 60 deg. If each optical axis lies in the Plane of the Paper, each lens scans in turn the Same scan line. To Provide elevation Coverage,

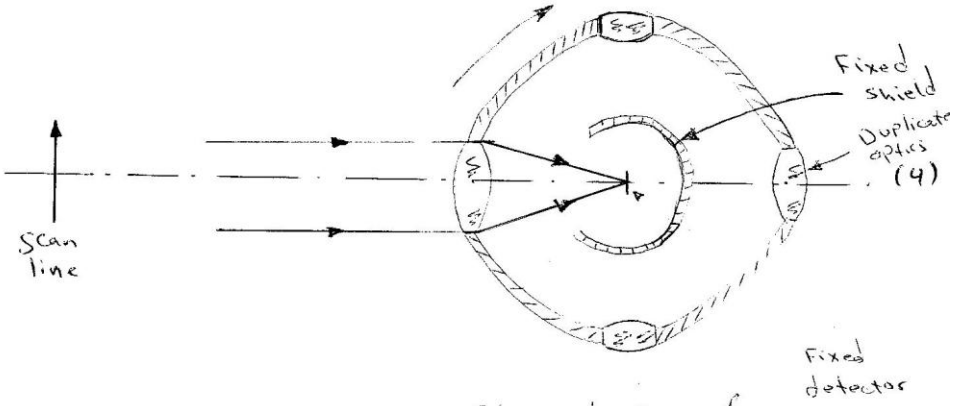


Fig (4-7) Use of rotating optics to generate a raster scan.

A pair of thin prisms placed in front of the optics as shown in fig (4-8) can be rotated to generate a variety of scan patterns. If the prisms rotate in opposite directions and if their angular velocities are equal, a linear scan results; if these velocities are not equal, a rosette scan is obtained. Similarly if they rotate in the same direction and if their angular velocities are equal, a circular scan results; if their velocities are unequal, a spiral scan is obtained.

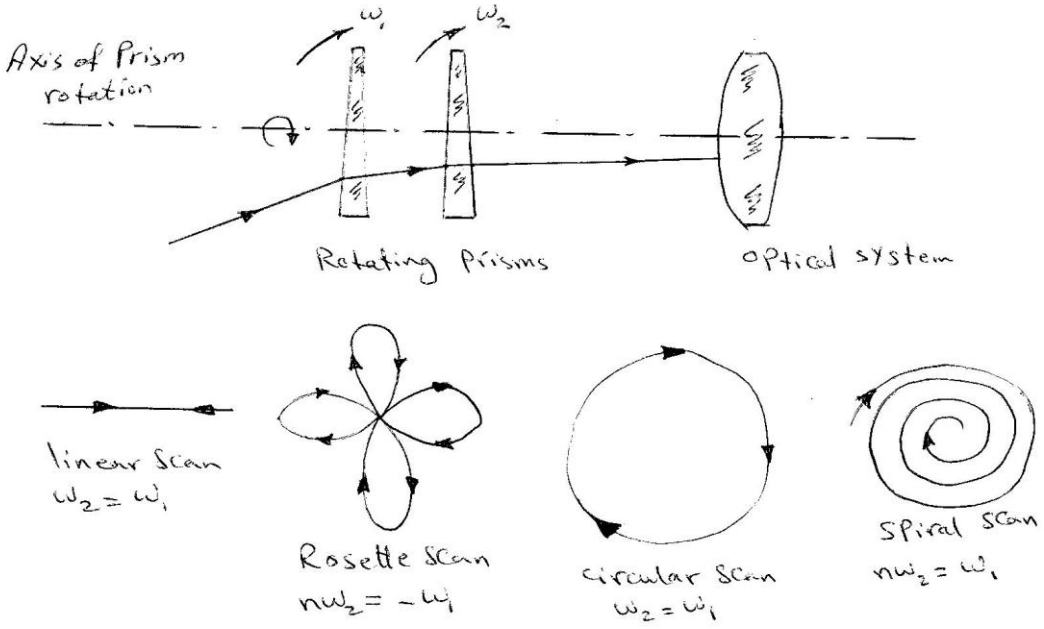


Fig (4-8) Scan generation with rotating prisms.

The efficiency with which the search field is covered can be increased by using multielement detectors, as shown in fig (4-9) with the linear array, a single scan generates several lines in a raster pattern. one for each element in the array. By using a mosaic, that is, a two-dimensional array of detector, it may be possible to cover the search field without any mechanical or optical scanning motion.

Because no scanning is involved, the entire search field is observed at all times. This is a great advantage in detecting a target that may appear only briefly and at an unknown time and location within the field. With any of the techniques previously described, a target that appears only briefly will be missed if the scan element is in some other part of the search field at that time. Although it simplifies the scanning system, the use of a linear array or mosaic increases the complexity of the electronics. It has not yet proved practical to apply conventional sampling or time-multiplexing methods at the low signal levels typical of infrared detectors. Instead each detector element must have its own preamplifier to bring the signal level up to the point where it is practical to use sampling. Multielement detectors are both difficult to manufacture and expensive.

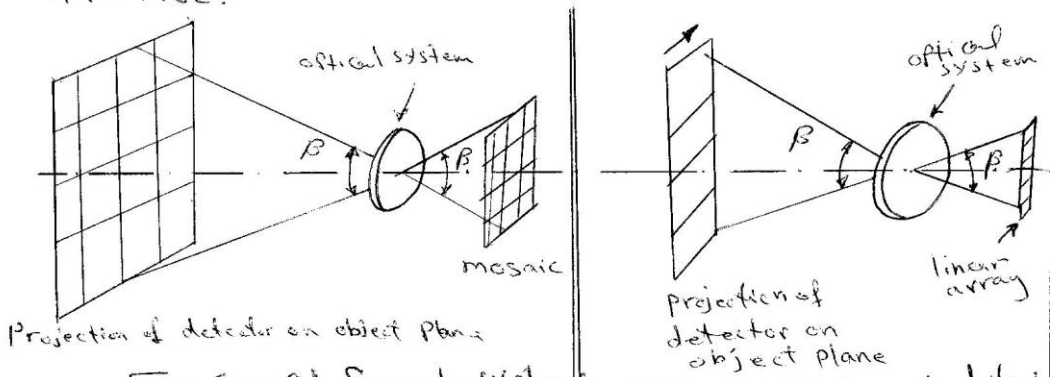


Fig (4-9) Search systems using multielement detecto

4.7 Optical materials for the Infrared.

Early workers in the infrared field rarely had to spend much time in choosing between refractive or reflective optics for their systems. Since so few infrared-transparent materials were available, the choice was almost invariably in favor of reflective optics. There are perhaps a hundred optical materials that transmit in some part of the infrared. The system engineer concerned with equipment that must operate in the field finds that most of these materials are of little use to him because of their undesirable Physical Properties. Before considering these materials, let us see what Physical Properties are important. The day when only the transmission and index of refraction of an optical material were significant has long since passed. Before selecting a material, the following properties should be examined in terms of the intended system application

1- Spectral transmittance and its variation with temperature

2- Index of refracting and its variation with temperature

3- Hardness
4- Resistance to surface attack by liquids.
5- Density
6- Thermal conductivity
7- Thermal expansion
8- Specific heat
9- Elastic moduli
10- Softening and melting temperatures.
11- RF Properties.

4.8 Antireflection Coatings.

The index of refraction of most of the preferred optical materials is high enough that significant amounts of the incident radiant flux are lost by reflection from the surfaces. A thin film or coating can be applied to the surface by vacuum evaporation in order to eliminate completely the reflection at a given wavelength. The reduction attainable over a band of wavelengths, such as an atmospheric window, although not complete, can still be very impressive. An antireflection coating for optical materials used in air must meet two criteria: its index of refraction must be equal to the square root of the index of the optical material to be coated, and its optical thickness must be equal to one-fourth of the wavelength at which minimum reflection is to occur.

optical thickness is the product of the index of reflection and the physical thickness of the coating. Since optical thickness changes with the angle of incidence, an antireflection coating will not be equally efficient for all of the rays in a converging bundle.

4.9 High-reflection Coatings

For many years chemically deposited silver was the traditional material for coating mirrors. Freshly applied films had a high reflectance but they tarnished rapidly when exposed to air. Since the introduction during world war II of the techniques for evaporating metal films, virtually all mirrors have been coated by this means. The reflectance of most metals increase at longer wavelengths. Table 4.1 shows the spectral reflectance of various evaporated metal films.

Wavelength (μ)	Reflectance (Percent)				
	Aluminum	Silver	Gold	Copper	Rhodium
0.5	90.4	97.7	47.7	60.0	77.4
1.0	93.2	98.9	98.2	98.5	85.0
3.0	97.3	98.9	98.3	98.6	92.5
5.0	97.7	98.9	98.5	98.7	94.5
8.0	98.0	98.9	98.4	98.7	95.2
10.0	98.1	98.9	98.4	98.8	96.0

Table (4.1) Reflection of evaporated metal films.

The characteristics of aluminum, the most commonly used material, occasionally the slightly higher reflectance of gold warrants using despite its higher cost. If a mirror is exposed to a dusty environment, it is wise to protect its surface with a thin evaporated coating of magnesium fluoride or silicon monoxide.

4.10 Optical filters

With the development of better photoconductive detectors it has become increasingly important to match the spectral bandpass of the system to one of the atmospheric transmission windows. The principal means of accomplishing this is to use optical filters. Two types are available. The absorption filter depends for its effectiveness on the absorbing characteristics of various dyes, plastics, and optical materials. The reflection or interference wavelengths. The relatively new interference filters can be made to have almost any desired spectral transmittance characteristics - Thus it desirable to have a more quantitative descriptive terminology for specifying filters.

(a) A bandpass filter : transmits a band of wavelengths sharply bounded by extended regions of low transmittance.

(b) Spectral bandwidth : describes the wavelength interval transmitted by the filter in terms of the center wavelength and the half-width.

(c) half-width : is the wavelength interval (in microns) over which the transmittance exceeds one-half the peak transmittance of the filter.

(d) Peak transmittance : is the maximum transmittance within the spectral bandwidth of the filter. It is expressed as a percentage of the transmittance of the uncoated substrate at the same wavelength.

(e) Substrate : is the optical material on which the filter is deposited.

(f) A long-wavelength Pass filter : transmits all wavelengths Longer than a specified cut-off wavelength.

(g) A Short-wavelength Pass filter : transmits all wavelengths shorter than a specified cut-off wavelength.

(H) Cut-on wavelength or cut-off wavelength λ_c is the wavelength at which the transmittance is 5 Percent of the Peak transmittance.

(I) slope : specifies the rate at which the transmittance increase between the cut-on or cut-off wavelength and the wavelength λ_c' where the transmittance is 80 Percent of the Peak transmittance. it is given by the wavelength difference between λ_c' and λ_c expressed as a decimal fraction of λ_c.

Absorption filters are rarely used as bandpass filters. because their spectral bandwidths are usually very wide. However, they are often used as short-wavelength or long-wavelength Pass filters. Because the absorption filter absorbs the radiant flux lying outside its spectral bandwidth, it may become warm and Perhaps even shatter. For this reason, a system employing such a filter should never be allowed to look directly at the Sun.

Interference filters are made by the vaccuum deposition of several layers of dielectric material onto a suitable substrate, the index of refraction and the thickness of each layer must be precisely controlled.

Infrared Technology

chapter one

Introduction to IR system engineering

1.1 The development of the IR Portion of the spectrum.

1.2 The market for IR devices

1.3 system engineering

chapter two

2.1 The electromagnetic spectrum

2.2 Terminology used in the measurment

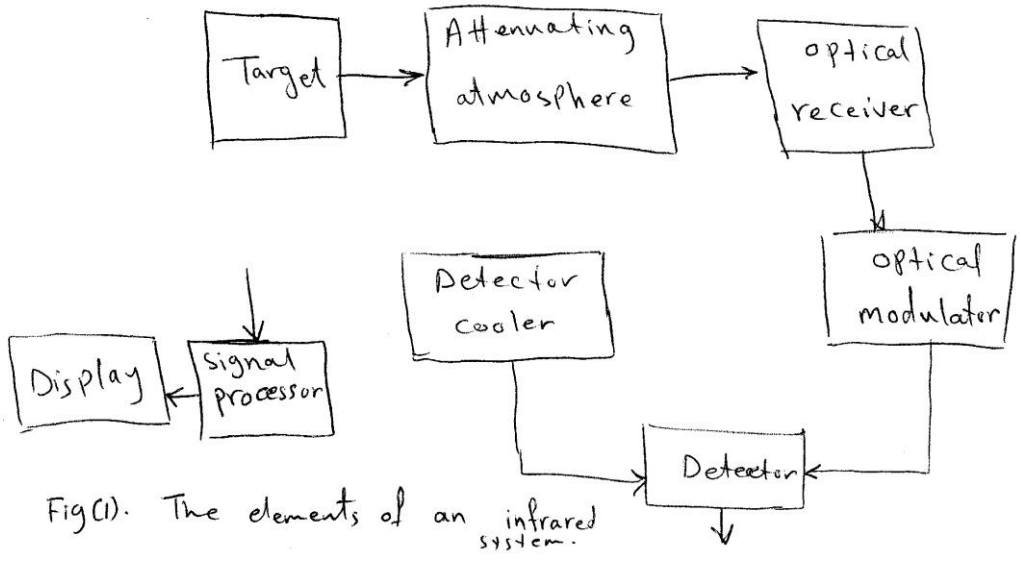
2.3 The measurment of radiant flux

2.4 Thermal radiation, Thermal radiation Laws

2.5 Emissivity and kirchoffs law.

The elements of an Infrared system are shown in block diagram form in fig (1). The target is the object of interset, usually the real reason for the existence of the system; it's assumed that the target radiates energy somwhere in the infrared portion of the spectrum. The system may be designed to detect the presence of the target, to track it as it moves; to glean information leading to its identity, or to measure its temperature. if the radiation from the target passes through any portion of the earth's atmosphere, it will be attenuated because the atmosphere is not perfectly transparent. The optical receiver which is closely analagous to a radar antenna, collects some of the radiation from the the target and delivers it to a detector which converts it into an electrical signal.

before reaching the detector, the radiation may pass through an optical modulator where it is coded with information concerning the direction to the target or information to assist in the differentiation of the target from unwanted details in the background. Since some detectors must be cooled, one of the system elements may be a means of providing such cooling. The electrical signal from the detector passes to the processor where it is amplified and the coded target information is extracted. The final step is the use of this information to interpretation by a human observer.



Fig(1). The elements of an infrared system.

1.3 The Electromagnetic spectrum.

Electromagnetic spectrum is an arrangement of the various radiations by wavelength or frequency, shown in fig (2). All of the radiations obey similar laws of reflection, refraction, diffraction and Polarization. The velocity of Propagation, Popularly called the "Velocity of light" is the same for all. They differ from one another only in wavelength and frequency.

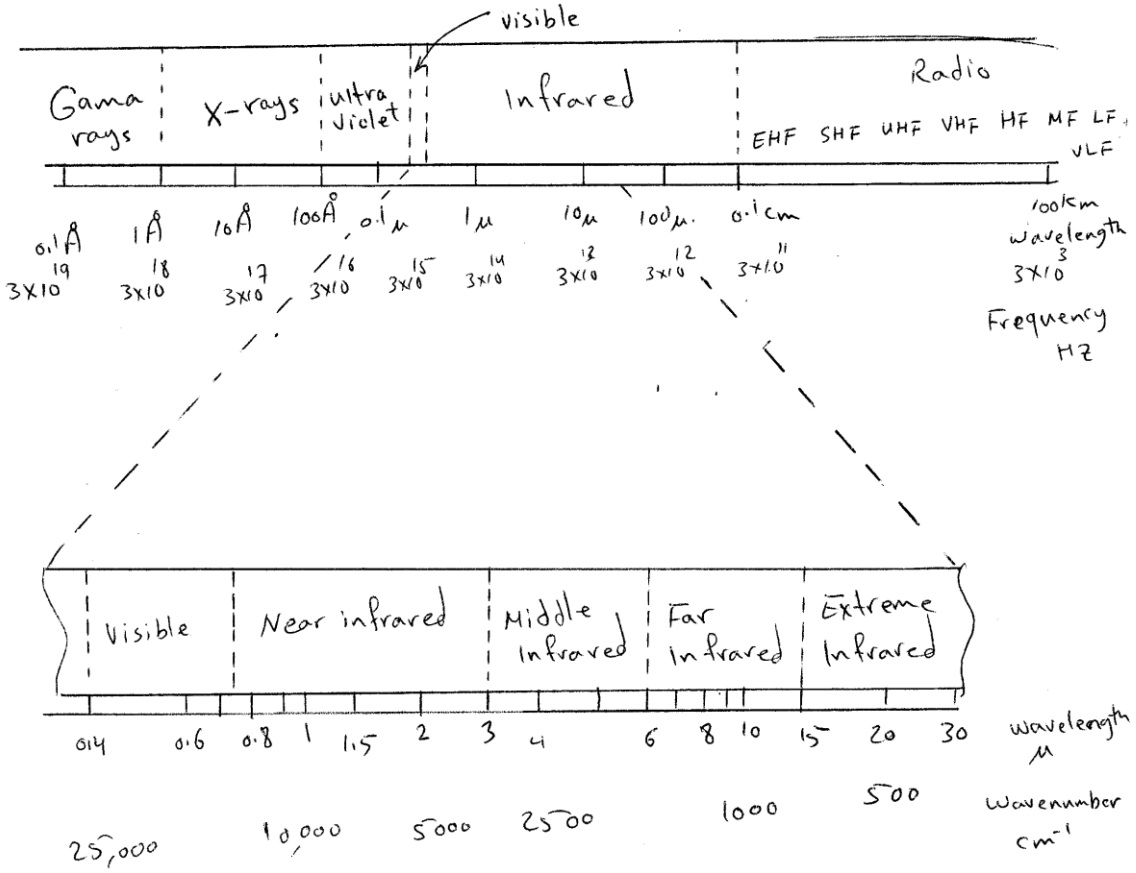


Fig (2) The electromagnetic spectrum.

where Near infrared 0.75 to 3 μ (NIR)

Middle Infrared 3 to 6 μ (MIR)

Far infrared 6 to 15 μ (FIR)

Extreme Infrared 15 to 1000 μ (XIR)

Since all of the types of radiation shown in fig(2) are considered to be a form of wave motion, they must obey the general equation

$$\lambda \nu = c$$

where λ is the wavelength, ν the frequency, and c the velocity. Wave length is the distance measured in the direction of Propagation, between any two successive Points on a wave having the same Phase. and it's in micron μ. where

$1\mu = 10^{-4}$ cm $= 10^{-6}$ m $= 3.937 \times 10^{-5}$ inch.

$1\mu = 10^{3}$ mμ $= 10^{4}$ Å Å angstrom.

* Frequency: is the number of waves Passing a Point Per unit time. the units, cycles Per Second are called the hertz (HZ).

chapter two

2.1 Terminology used in the measurment of radiant energy.

It is unfortunate that the concepts basic to the measurment of light evolved many years before it was known that light is not a separate entity but is instead merely the radiation in one narrow region of the electromagnetic spectrum. Photometry (قياس شدة الضوء), the measurment of light implicitly involves (يتضمن) the visual (بصري) sensations produced by light in the consciousness (شعور) of an observer. Thus the methods of Photometry are Psychophysical rather than Physical. and Photometry cannot be classed along with such familiar Physical measurments as those of mass, length or time.

ملاحظة: قياس الضوء البصري الذي يتناول الضوء على شعور ووعي المراقب لذلك فانه يختلف عن فرع علم الفيزياء

more pertinent to the infrared region are the methods of radiometry, the measurment of radiant energy which are based on a system of Physical measurments. In Principle, ~~radiation~~ radiometers absorb some of the radiant energy from the source and convert it to another form, such as electrical, thermal or chemical energy.

The energy transferred by electromagnetic waves is called radiant energy (U) and is measured in joules. This term is used to describe the entire amount of energy radiant from a source in a given time interval. it is also describes the energy received by an accumulative or integrating type of detectors, such as the Photographic plate.

Most detectors used in infrared equipment respond to the time rate of transfer of radiant energy rather than to the total amount of energy transferred.

Radiant flux (P) is the measure of the time rate of transfer of radiant energy and is given in watts: a watt is numerically equall to one joule per second. An equally acceptable equivalent term, preferred by some authors, is radiant power

Three terms (a) radiant emittance, (b) radiant intensity and (c) radiance may be used to describe the radiant flux from a source. They are usually determined by radiometric measurements made at a distance from the source.

The radiant flux emitted per unit area of a source is the radiant emittance (W). It is the limiting value of the expression $W = \frac{\partial P}{\partial A}$, where P is the radiant flux emitted by a source element having an area A. one way of measuring the flux is to place the source in a device that collects flux from all directions

with equall efficiency. After a correction is made for the efficiency of the collector, the flux collected is divided by the area of the source to obtain the radiant emittance. For an inaccessible source such as the Sun. we would operate as follows:

1- Measure the radiant flux at the detector.

2- Divide this value by the solid angle subtended by the receiver at the Sun.

3- correct for attenuation, and so forth, along the line of sight

4- Multiply by 4π

5- Divide this Product by the area of the Sun.

step 4 yields a value for total radiant from the Sun based on the assumption that the flux it emits in the same in all directions, dividing this value by the area of the Sun (step 5) gives the radiant emittance

Before discussing radiant intensity and radiance, it is necessary to differentiate between Point Sources and extended Sources

A true Point Source is of course, not physically realizable, but it can be closely approximated by a star, a small source at a great distance, or by an optical device called a collimator. what is important is not the physical size of the source but the angle that it subtends at the detector. The same source can at different times be classed as either point or extended. The tailpipe of a jet aircraft at a distance of 10 miles is effectively a point source, whereas at a distance of 10 ft it represents an extended source. if the image is smaller than the detector the source is considered to be a point source; if the image is larger than the detector, the source is considered an extended one.

The radiant flux emitted per unit solid angle is called the radiant intensity (J). it is the limiting value of the expression $J = \frac{\partial P}{\partial \omega}$, where ω is the solid angle subtended by the detector at the source.

Radiant intensity is used to describe a point source. The radiant intensity is found by measuring the radiant flux and dividing it by the solid angle subtended by the detector at the source.

if the source is an extended one, such as the sky, it is not possible to define the solid angle subtended by the detector at the source. This difficulty is resolved by describing an extended source by the radiance (N), the radiant flux per unit solid angle per unit area of source. it is the limiting value of the expression $N = \frac{\partial^2 P}{\partial A \, \partial \omega}$. In order to measure radiance it is necessary to use masks or optical means to

limit the measurement to a small area of the extended source.

The radiant Photon emittance (Q) is the number of Photons emitted Per second Per unit area.

Irradiance (H) is the radiant flux incident on a surface of unit area. The units in which it is measured, $W\,cm^{-2}$, are the same as those used for radiant emittance, the Irradiance at a distance d from the source is $H = \frac{J}{d^2}$ (with a Point source)

The error in assuming that the source is a Point is less than 1 Per cent, if the distance is at least ten times the largest dimension of the source. If the source is an extended one, the irradiance must be found by an integration.

Radiant emittance, radiant Photon emittance, radiant intensity radiante, and Irradiance refer to the flux contained in a Particular solid angle or Passing through a Particular area. Thus these quantities are differential with respect to solid angle or area.

Each is associated with a corresponding spectral quantity in which the radiant flux is that within a small wavelength interval centrated about a particular wavelength. Thus the spectral quantities are differential with respect to wavelength. As an example, spectral radiant flux (P_λ) is the radiant flux per unit wavelength interval evaluated at a particular wavelength: it is the limiting value of the expression $P_\lambda = \frac{\partial P}{\partial \lambda}$. use of the subscript λ to indicate a differential with respect to wave length has had wide acceptance and is followed in this chapter. In order to find the radiant flux between wavelength λ_1 and λ_2 it is necessary to integrate the expression

$$P = \int_{\lambda_1}^{\lambda_2} P_\lambda \, d\lambda .$$

If the limits extend from zero to infinity, the result of the integration is the total radiant flux.

2.2 The Measurment of radiant flux

A radiometer is a device for measuring radiant flux over a broad spectral interval.

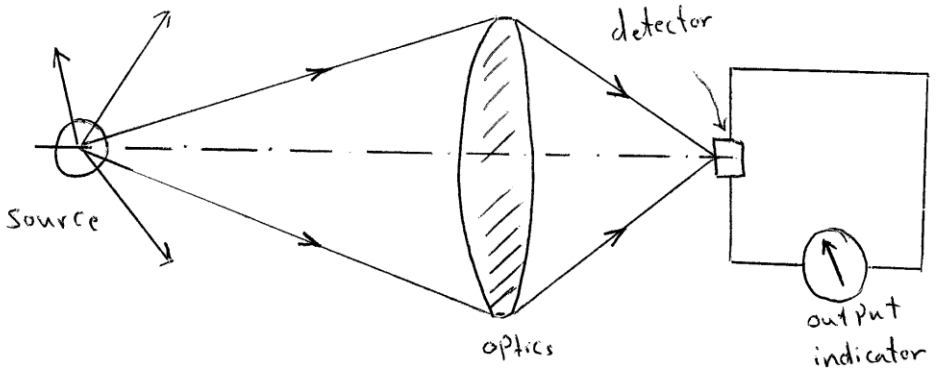
A spectroradiometer is a device for measuring the spectral radiant flux within a small spectral interval. Thus radiometry provides broad-band measurments, while spectroradiometry is used for narrow band measurments.

The basic elements of a radiometer are shown in fig below, some of radiant flux from the source is collected by the optics and focused onto the detector the detector produces an electrical signal that is proportional to the flux input, since such measurment are invariably made at a distance from the source. The radiometer responds to the irradiance, the areal density of the flux, at its input (the optics). Thus irradiance is the fundamental quantity involved in all radiometric measurments; other quantities, such as radiant emittance, radiant intensity, and radiance are

calculated from the measured value of Irradiance

In the radiometer, the spectral interval over which the measurement is made is controlled by the spectral response of the detector and the transparency of the optics. If the detector responds equally to all wavelengths and that the optics transmit all wavelengths without absorption, then the output indication will be proportional to the total Irradiance at the optics.

Radiometers can be constructed that come quite close to achieving these conditions, By using athermal detector and mirror optics, we can obtain a nearly equal response to wavelengths extending from 2 to 40 μ. If desired an optical filter can be placed in front of the detector to limit the response of the radiometer to any desired smaller spectral interval.



Elements of a radiometer

The elements of a spectroradiometer are shown in fig'l below consists of two major parts: a monochromator to provide radiant flux of a narrow band of wavelengths, and a radiometer to measure this flux. The flux from the source is dispersed or spread into a spectrum by a prism. A small portion of this flux passes through the ~~exit~~ exit slit to the radiometer. Rotation of the prism- and mirror combination varies the wavelength of the flux passing through the exit slit; the width of this slit determines the spectral interval passed by the monochromator. The monochromator and source are seen to be analogous to the signal generator used in electronics. The prism and exit slit are used to select a particular wavelength; in the signal generator a particular frequency is selected by varying a resonant circuit. Note, however, that the signal generator inherently provides a single-frequency signal, whereas the signal from the monochromator must always consist of

a narrow band of wavelengths. Perhaps a somewhat closer analogy is a broad-band noise generator followed by a variable band pass filter.

A spectroradiometer measures the spectral distribution of radiant flux, that is the variation in flux as a function of wavelength. If the source is a heated solid or liquid, the spectral distribution curve is continuous and shows a single maximum at a wavelength that varies with the temperature of the source. Such sources are called thermal radiators. If the source is a flame or an electrical discharge in a gas, the spectral distribution curve is not continuous, but instead the flux is concentrated in narrow spectral intervals. with a high-resolution monochromator these intervals may appear to be extremely narrow, sharply defined lines, and the distribution is called a line spectrum. Alternatively, the spectrum may consist of bands of narrow lines, and in this case it is called a band spectrum. Sources giving line or band spectra are called selective radiators.

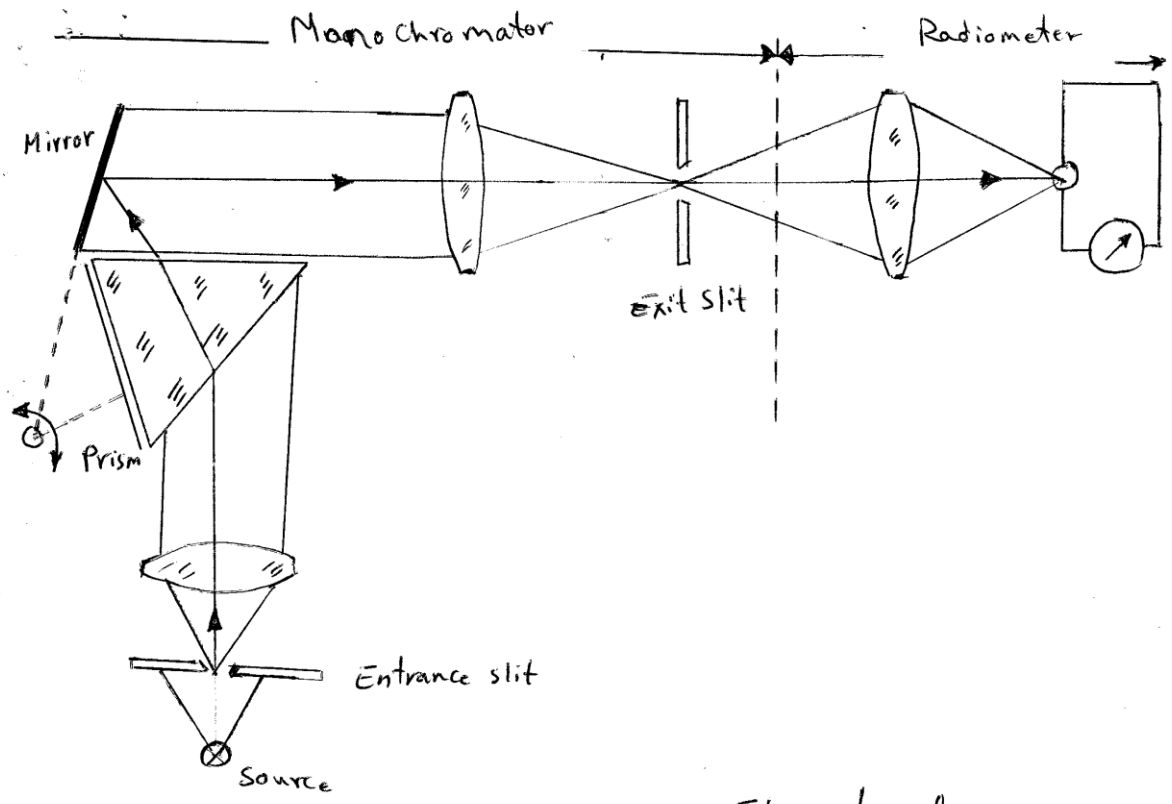


Fig (1) Elements of a spectroradiometer.

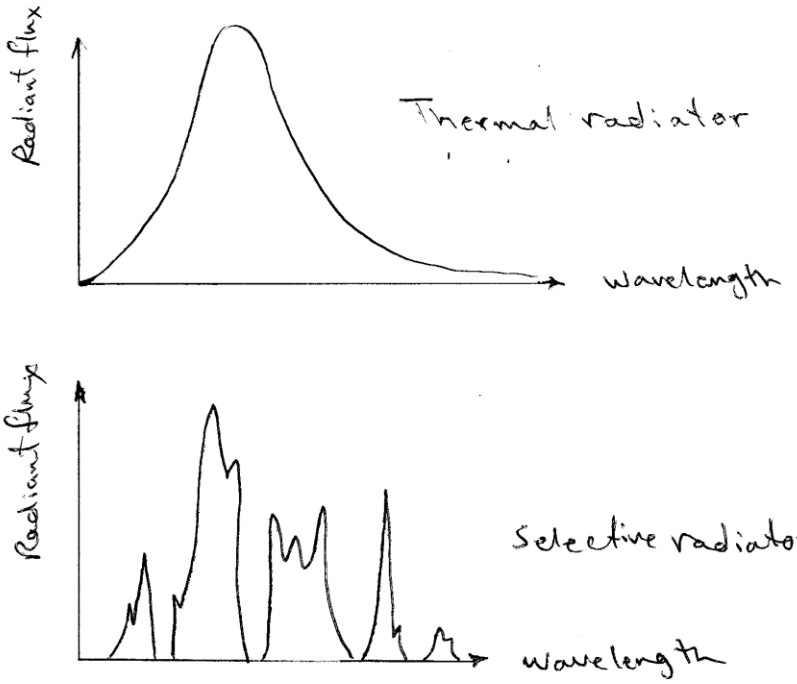


Fig (2) Spectral distribution of thermal and selective radiators.

2.4 Thermal radiation

one of the major Problems facing Physicists during the second half of the nineteenth century was to explain the energy distribution in the spectrum of a thermal radiator In 1860 Kirchhoff introduced his famous law that states in effect, that good absorber are also good radiators. This law is one of the keystones in the theory of radiation transfer. Kirchhoff also Proposed the term blackbody to describe a body that absorbs all of the incident radiant energy and that, as a consequence of his law, must also be the most efficient radiator. A black body, then Provides a standard of comparison; it is the ultimate thermal radiator with which we can compare any other source. In 1879 Stefan concluded from his experimental measurments that the total amount of energy radiated by a blackbody is proportional to the fourth power of

its absolute temperature. In 1884 Boltzmann reached the same conclusion by the application of thermodynamic relationships; the result has become known as the Stefan-Boltzmann law. In 1894 Wien published the displacement law that gives the general form of the equation for the spectral distribution of the radiation from a blackbody.

2-4-1 Thermal Radiation laws

Plank's law describes the spectral distribution of the radiation from a blackbody as

$$W_\lambda = \frac{2\pi h c^2}{\lambda^5} \frac{1}{\left(e^{ch/\lambda K T} - 1\right)}$$

which is usually written as

$$W_\lambda = \frac{C_1}{\lambda^5} \frac{1}{e^{C_2/\lambda T} - 1}$$

where

W_λ = spectral radiant emittance $w/cm^2.\mu$

λ = wavelength $.\mu$

h = plank's constant = $(6.6256 \pm 0.0005) \times 10^{-34}$ w·sec^2

T = absolute temperature K°

c = velocity of light = $(2.997925 \pm 0.000003) \times 10^{10}$ cm/sec

$c_1 = 2\pi hc^2$ = First radiation constant

$= (3.7415 \pm 0.0003) \times 10^{4}$ w/cm^2 μ^4

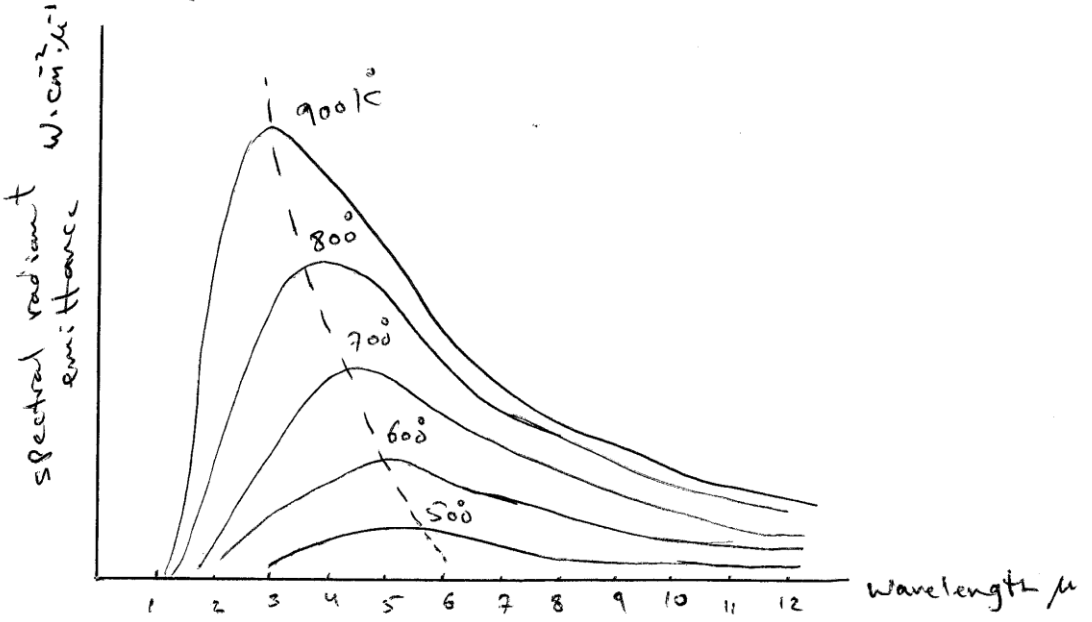
$c_2 = ch/K$ = second radiation constant

$= (1.43879 \pm 0.0019) \times 10^{4}$ μK°

K = Boltzmann's constant = $(1.38054 \pm 0.00018) \times 10^{-23}$ w·sec·K°$^{-1}$

The spectral radiant emittance of a blackbody at temperatures ranging from 500°K to 900°K is shown in figure below. This is an interesting range because it includes the temperature of the hot metal tailpipes of turbojet aircraft. Several characteristics of the radiation from a blackbody are evident from these curves. The total radiant emittance, which is proportional to the area under the curves, increases rapidly with temperature. The wavelength of maximum spectral radiant emittance shifts toward shorter wavelengths as the temperature increases. the individual curves never crosses one another; hence the higher the temperature, the higher the spectral radiant

emittance at all wavelengths.



spectral radiant emittance of a blackbody at various temperature.

* Integrating Planck's law over wavelength limits extending from zero to infinity gives an expression for the radiant emittance, the flux radiated into a hemisphere above a blackbody 1 cm² in area This is commonly known as the Stefan-Boltzmann law.

$$W = \frac{2\pi^5 k^4}{15 c^2 h^3} T^4 = \sigma T^4$$

where

W = radiant emittance $W.cm^{-2}$

σ = Stefan-Boltzmann constant

= $(5.6697 \pm 0.0029) \times 10^{-12}\ W.cm^{-2} K^{-4}$

in the above the rapid increase in radiant emittance with increasing temperature is evident; from the Stefan-Boltzmann law, this increase is proportional to the forth power of the absolute temperature. Thus relatively small changes in temperature can cause large changes in radiant emittance.

* Differentiating Plank's law and solving for the maximum gives Wien's displacement law.

$$\lambda_m T = a$$

where

λ_m = wavelength of maximum spectral radiant emittance

$a = 2897.8 \pm 0.4\ \mu^\circ K$

Thus the wavelength at which the maximum spectral radiant emittance occurs varies inversely with the absolute temperature. The dashed curve in figure above is the locus of these maxima. * An alternative form of Wien's displacement law gives the maximum value of the spectral radiant emittance as

$$W_{\lambda m} = 21.2014\mu \frac{C_1}{C_2^5} \cdot T^5 = b\, T^5$$

where

$W_{\lambda m}$ = maximum spectral radiant emittance $w \cdot cm^{-2} \cdot \mu^{-1}$

$b = 1.2862 \times 10^{-15}\ w\, cm^{-2} \mu^{-1} {}^\circ C^{-5}$

Hence the value of the maximum spectral radiant emittance from a blackbody is proportional to the fifth power of its absolute temperature.

The radiation laws can also be written in terms of photons, a form that will be useful in discussing the performance of photon detectors.* If Plank's expression is divided by hc/λ, which is the energy associated with one photon, the result is the spectral radiant photon emittance.

$$Q_\lambda = \frac{2\pi c}{\lambda^4} \frac{1}{e^{ch/\lambda KT} - 1}$$

$$= \frac{C_1'}{\lambda^4} \frac{1}{e^{C_2/\lambda T} - 1}$$

where

Q_λ = spectral radiant photon emittance

photon $sec^{-1} \cdot cm^{-2} \cdot \mu^{-1}$

$C_1' = 2\pi c = 1.88365 \times 10^{23}$ $sec^{-1} \cdot cm^{-2} \mu^3$

*

The Stefan – Boltzmann law becomes

$$Q = \frac{c_1'}{c_2^3} \frac{2\pi^3}{25.79436} = \sigma' T^3$$

where

Q = radiant Photon emittance, Photon $sec^{-1} cm^{-2}$

$\sigma' = 1.52041 \times 10^{11}$ $sec^{-1} cm^{-2} K^{-3}$

Therefore the rate at which Photons are emitted from a blackbody varies as the third Power of its absolute temperature, rather than as fourth – Power relationship observed with radiant flux.

*

The Wien displacement law becomes

$$\lambda_m' T = \frac{c_2}{3.92069} = a'$$

where

λ_m' = wavelength of maximum spectral radiant Photon emittance

$a' = 3669.73 \, \mu K$

This the displacement law has the same form for both flux and Photons. but the wavelength at which the maximum occurs is about 25 Per cent

greater for photons. The alternative form of Wien's law becomes

$$Q'_{\lambda_m} = 4.77984 \frac{C_1'}{C_2^4} T^4 = b' \cdot T^4$$

where

Q'_{λ_m} = maximum spectral radiant photon emittance

Photon $sec^{-1} cm^{-2} \mu^{-1}$

$b' = 2.10098 \times 10^7$ $sec^{-1} cm^{-2} \mu^{-1} {}^\circ K^{-4}$

Another pair of useful relationships is

$$\text{energy per photon} = \frac{hc}{\lambda} = \frac{1.9863 \times 10^{-19}}{\lambda} \quad \text{W.Sec}$$

where λ is in microns. The reciprocal of this expression gives the number of photons per second per watt.

$1 \text{ watt} = 5.035 \times 10^{18} \lambda$ Photon sec^{-1}.

2-5 Emissivity and Kirchoff's Law

The formulas in the previous section describe the radiation from a blackbody. A factor can be added to them so that they can also be applied to sources that are not blackbodies. This

factor, called the emissivity ϵ, is given by the ratio of the radiant emittance W' of the source to the radiant emittance of a blackbody at the same temperature.

$$\epsilon = \frac{W'}{W} .$$

Thus emissivity is a numeric whose value lies between the limits of zero for a nonradiating source and unity for a blackbody. Emissivity is a function of the type of material and its surface finish and it can vary with wavelength and with the temperature of the material. A more general expression in terms of the spectral emissivity $\epsilon(\lambda)$ is

$$\epsilon = \frac{\int_0^\infty \epsilon(\lambda) W_\lambda d\lambda}{\int_0^\infty W_\lambda d\lambda} = \frac{1}{\sigma T^4} \int_0^\infty \epsilon(\lambda) W_\lambda d\lambda$$

There are three of sources can be distingnished by the way that the spectral emissivity varies.

① A blackbody or Plankian radiator, for which $\epsilon(\lambda) = \epsilon = 1$

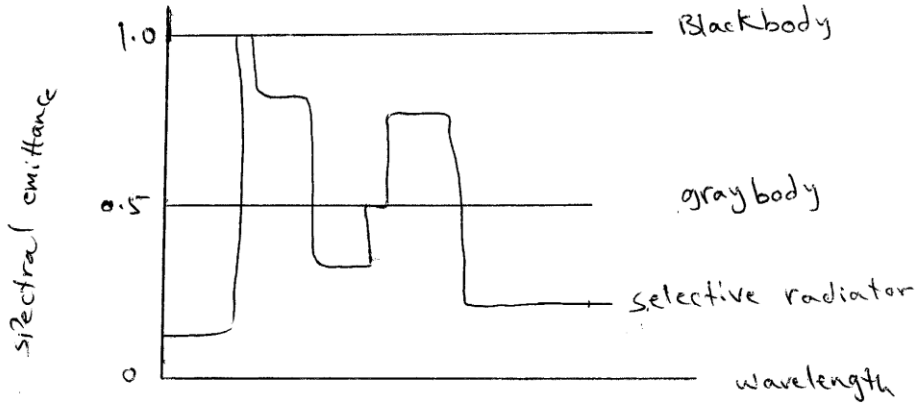
② A graybody, for which $\epsilon(\lambda) = \epsilon =$ constant (but less than unity)

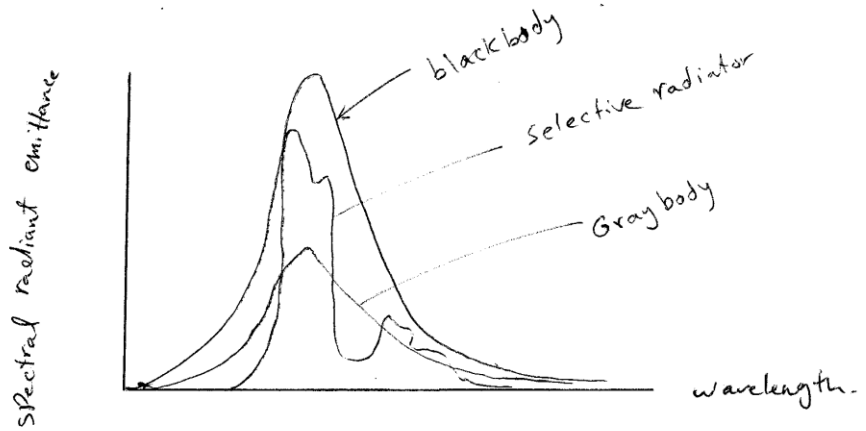
③ A selective radiator, for which $\epsilon(\lambda)$ varies with wavelength

From the figure below shown the spectral emissivity and spectral radiant emittance for each type of source

A blackbody, which has defined as the ultimate thermal

radiator, radiates more flux, either total or in an arbitrary spectral interval, than any other type of source at the same temperature. Thus the spectral distribution curve of a blackbody provides the limiting envelope for the other types of sources. A gray body, for which the emissivity is a constant fraction of that for a blackbody, is a particularly useful concept because such sources as jet tailpipes, aerodynamically heated surfaces, unpowered space vehicles personnel, and terrestrial and space backgrounds can be represented as graybodies with an accuracy sufficient for most engineering calculations.





spectral emissivity and spectral radiant emittance of three types of radiators.

As shown in fig above, a selective radiator can sometimes be considered to be a gray body over a limited spectral interval, thus simplifying calculations. When radiant energy is incident on a surface, three processes can occur: a fraction of the incident energy α may be absorbed, a fraction ρ may be reflected, and a fraction τ may be transmitted. Since energy must be conserved, the following relation ship can be written: $\alpha + \rho + \tau = 1$.

By definitation or definition, a blackbody absorbs all of the incidient radiant energy so that $\alpha = 1$ and $\rho = \tau = 0$ While studying the radiation transfer process, Kirchhoff observed that at a given temperature the ratio of radiant emittance to absorptance is a constant for all materials and that it is equal to the radiant emittance of a blackbody at that temperature known as Kirchhoff's law, it can be stated as

$\frac{W'}{\alpha} = W$. This law is often paraphrased as "good absorbers are good emitters". And we can rewritten as

$W' = \alpha W$, therefore

$$\frac{\epsilon \sigma T^4}{\alpha} = \sigma T^4$$

From this it follows that $\epsilon = \alpha$

Thus the emissivity of any material at a given temperature is numerically equal to its absorptance at that temperature. Since an opaque material does not transmit energy $(\alpha + \rho)$ equals 1 and $\epsilon = (1 - \rho)$. This is a particularly convenient relationship, since it is often easier to measure reflectance than to measure emissivity directly. Since emissivity can vary with the direction of measurement, particularly for polished metals, it is necessary to define several types of emissivity. Hemispherical emissivity ϵ_h is defined by

$\epsilon = \frac{1}{\sigma T^4} \int_0^\infty \epsilon(\lambda) W_\lambda d\lambda$, and gives the emissivity of a source radiating into a hemisphere. This type of emissivity is important in calculating the amount of heat transferred by radiation. Directional emissivity ϵ_θ is the emissivity measured in a small solid angle at an angle θ from the normal to the radiating surface. The particular case in which the angle θ is zero is called the normal emissivity ϵ_n. Each type may be either a total (implying that it is measured over all wavelengths) or a spectral quantity. Since most infrared systems respond to the flux contained in a small solid angle at a definite direction from the source, ϵ_θ and ϵ_n are of particular interest.

Fortunately the differences between ϵ_h, ϵ_θ and ϵ_n are usually small, and they can be ignored except for polished metals, for which the hemispherical emissivity is about 20 percent greater than the normal emissivity. For metals, emissivity is low, but it increases with temperature and may increase tenfold or more with the formation of an oxide layer on the surface. For nonmetals, emissivity is high, usually more than 0.8, and it decreases with increasing temperature. The radiation from a metal or other opaque material originates within a few microns of the surface; hence emissivity is a function of the surface state of a material rather than of its bulk properties. For this reason, the emissivity of a coated or painted surface is characteristic of the coating rather than of the underlying surface.

Optical detectors

1.1 Introduction
1.2 Device types
1.3 Optical detection principles
1.4 Absorption
1.5 Quantum efficiency
1.6 Responsivity
1.7 Long-wavelength cutoff
1.8 Semiconductor photodiodes without internal gain
1.9 Semiconductor photodiodes with internal gain
1.10 Mid-infrared and far-infrared photodiodes
1.11 Phototransistors

Lecture No. 1

- **1.1 Introduction:**
- **Photodetectors are classified according to their optical range which can be covered. The electromagnetic spectrum in the optical start regarding wavelength from UV, visible, and IR ranges (Buckner, 2008). But it is evident to notice that the visible range is small. So, it will be impeded some time into UV or/ and IR detectors. The main two types then are UV and IR photodetectors. The main problem is appearing at long wavelength optical range as the required energy gap should be small. As a result of the previous note, the required band gap materials which cover this region have very high cost. And, it is faces a complexity into the fabrication processes. The trend of the world was directed to nanotechnology (quantum) photodetectors to overcome the previous drawbacks. As a result, the review is directed to the quantum IR photodetectors like quantum well, dot, and wire infrared photodetectors respectively. They open the way to overcome main problems in commercial utilized IR photodetectors such as HgCdTephotodetectors.**

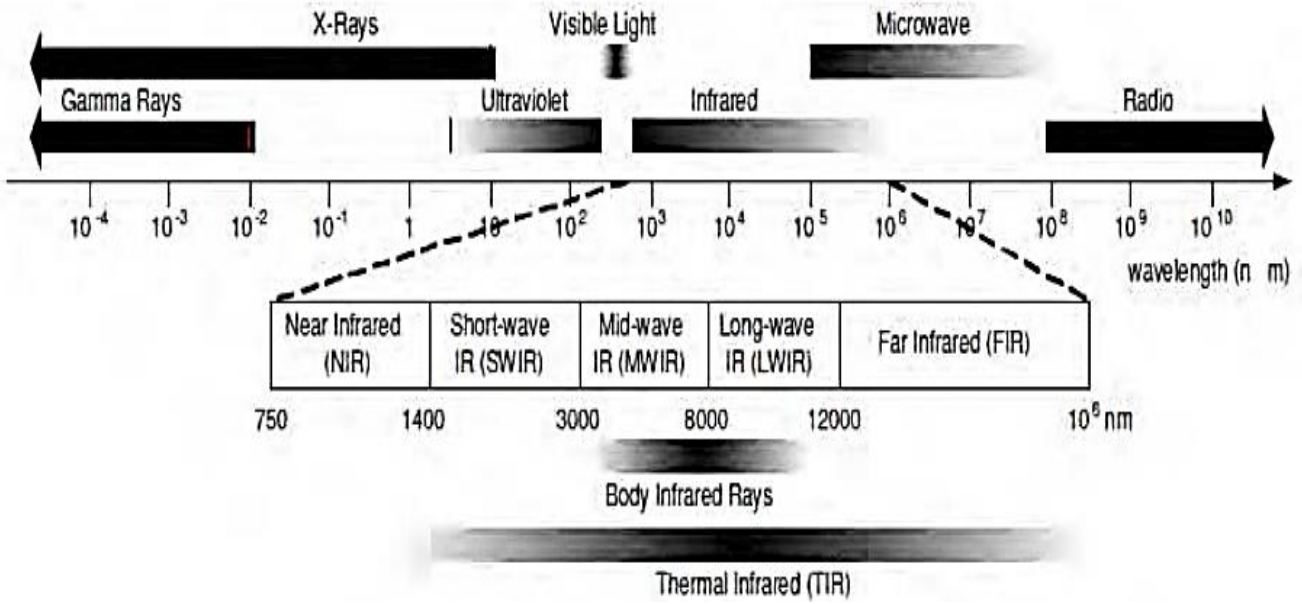


Fig. 1.1: The electromagnetic spectrum and the IR region.

- To understand the physical meaning of the detection process, we must mention how light emitting and how we choose the layer structure of any device (Detector or source of light). Fig. (1.2) shows the layer structure and circuit diagram for a typical electroluminescent device. The device consists of several epitaxial layers grown on top of a thick crystal substrate. The epitaxial layers consist of a p-n diode with a thin active region at the junction. The diode is operated in forward bias with a current flowing from the p-layer through to the n-layer underneath. The luminescence is generated in the active region by the recombination of electrons that flow in from the n-type layer with holes that flow in from the p-type side

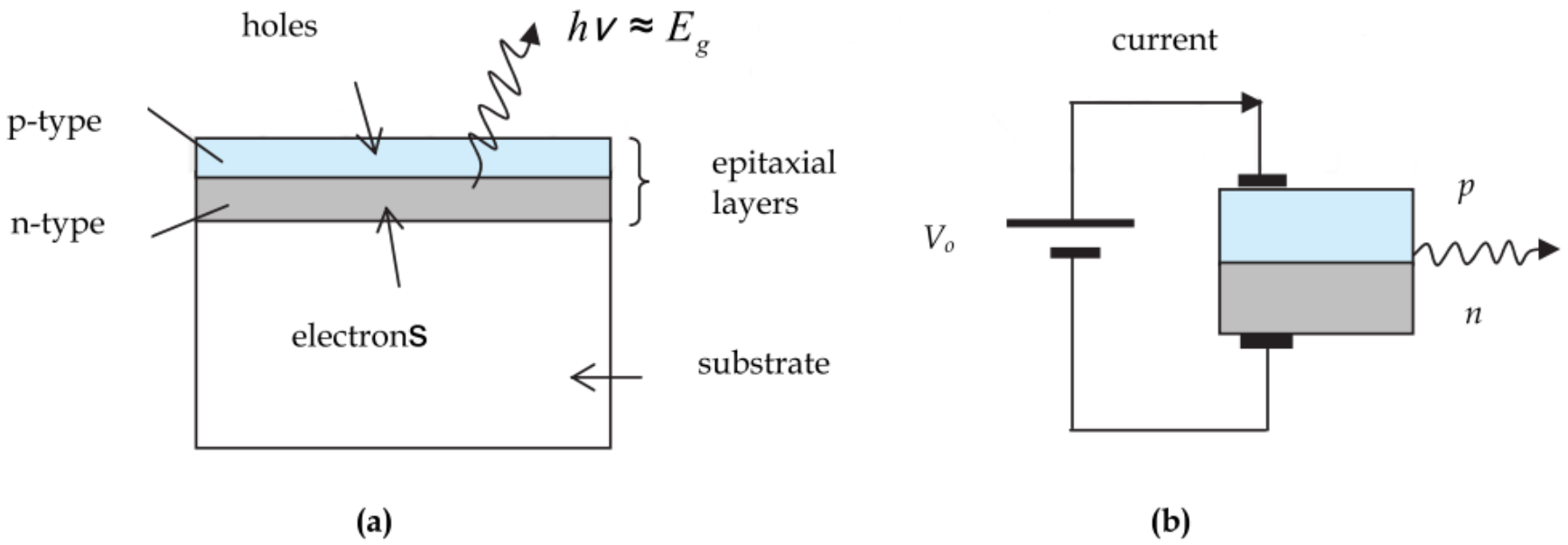


Fig. 1.2: (a) Layer structure and (b) circuit diagram for a typical electroluminescent device. The thin active region at the junction of the p-type and n-type is not shown, and the dimensions are not drawn to scale. The thickness of epitaxial layers will be only , whereas the substrate might be thick. The lateral dimensions of the device might be several millimetres.

- Any direct gap semiconductor can, in principle, be used for the active region, but in practice only a few materials are commonly employed. The main factors that determine the choice of the material are:
-
  - The size of the band gap;
  - Constraints related to lattice matching;
  - The ease of p-type doping.
-
  The first point is obvious: the band gap determines the emission wavelength. The second and third points are practical ones relating by which the way the devices are made.

1.1.1 Magnitude of the Energy Gap and concept of Conductors, Insulators, Semiconductors:

- Silicon and germanium have band gaps of 1 eV and 0.7 eV, respectively. At room temperature, a small fraction of the electrons are in the conduction band. Si and Ge are intrinsic semiconductors. Up to 4.0 eV the material becomes insulator. See Fig. (3)

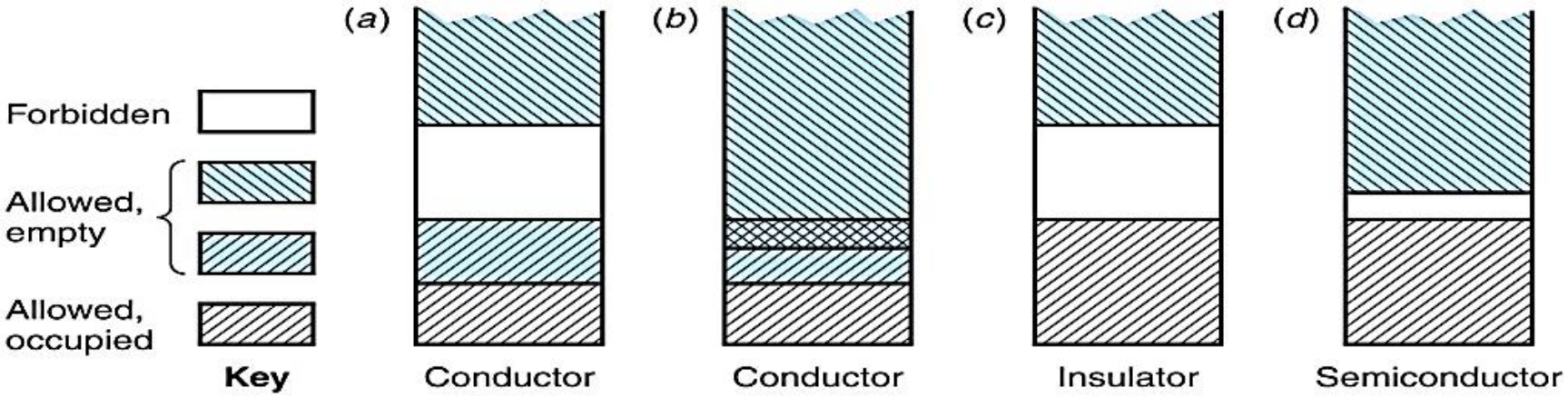


Fig.3: The difference between conductors, insulator and semiconductors.

- The role of an optical receiver is to convert the optical signal back into electrical form and recover the data transmitted through the lightwave system. Its main component is a photodetector that converts light into electricity through the photoelectric effect. The requirements for a photodetector are similar to those of an optical source. It should have high sensitivity, fast response, low noise, low cost, and high reliability. Its size should be compatible with the fiber-core size. These requirements are best met by photodetectors made of semiconductor materials.

Lecture No. 2

- **1.2 Device types**
- To detect optical radiation (photons) in the near-infrared region of the spectrum, both external and internal photoemission of electrons may be utilized. External photoemission devices typified by photomultiplier tubes and vacuum photodiodes meet some of the performance criteria but are too bulky, and require high voltages for operation. However, internal photoemission devices, especially semiconductor photodiodes with or without internal (avalanche) gain, provide good performance and compatibility with relatively low cost. These photodiodes are made from semiconductors such as silicon, germanium and an increasing number of III–V alloys, all of which satisfy in various ways most of the detector requirements. They are therefore used in all major current optical fiber communication systems.

- The internal photoemission process may take place in both intrinsic and extrinsic semiconductors. With intrinsic absorption, the received photons excite electrons from the valence to the conduction bands in the semiconductor, whereas extrinsic absorption involves impurity centers created within the material. However, for fast response coupled with efficient absorption of photons, the intrinsic absorption process is preferred and at presents all detectors for optical fiber communications use intrinsic photodetection.

- Silicon photodiodes have high sensitivity over the 0.8–0.9 μm wavelength band with adequate speed (tens of gigahertz), negligible shunt conductance, low dark current and long-term stability. They are therefore widely used in first-generation systems and are currently commercially available. Their usefulness is limited to the first-generation wavelength region as silicon has an indirect bandgap energy of 1.14 eV giving a loss in response above 1.09 μm. Thus for second-generation systems in the longer wavelength range 1.1 to 1.6 μm research is devoted to the investigation of semiconductor materials which have narrower bandgaps. Interest has focused on germanium and III–V alloys which give a good response at the longer wavelengths. Again, the performance characteristics of such devices have improved considerably over recent years and a wide selection of III–V alloy photodiodes as well as germanium photodiodes are now commercially available.

- In addition to the development of advanced photodiode structures fabricated from III–V semiconductor alloys for operation at wavelengths of 1.3 and 1.55 μm, similar material systems are under investigation for use at the even longer wavelengths required for mid-infrared and far-infrared transmission (2 to 12). Interest has also been maintained in other semiconductor detector types, namely the heterojunction phototransistor and the photoconductive detector, both of which can be usefully fabricated from III–V alloy material systems. In particular, the latter device type has more recently found favor as a potential detector over the 1.1 to 1.6 wavelength range. Nevertheless, at present the primary operating wavelength regions remain 0.8 to 0.9 , 1.3 μm and 1.55 , with the major device types being the p–i–n and avalanche photodiodes. We shall therefore consider these devices in greater detail before discussing mid-infrared photodiodes, phototransistors and photoconductive detectors.

1.3 Optical detection principles

- The basic detection process in an intrinsic absorber is illustrated in Figure 1.1 which shows a p–n photodiode. This device is reverse biased and the electric field developed across the p–n junction sweeps mobile carriers (holes and electrons) to their respective majority sides (p- and n-type material). A depletion region or layer is therefore created on either side of the junction. This barrier has the effect of stopping the majority carriers crossing the junction in the opposite direction to the field. However, the field accelerates minority carriers from both sides to the opposite side of the junction, forming the reverse leakage current of the diode. Thus intrinsic conditions are created in the depletion region.

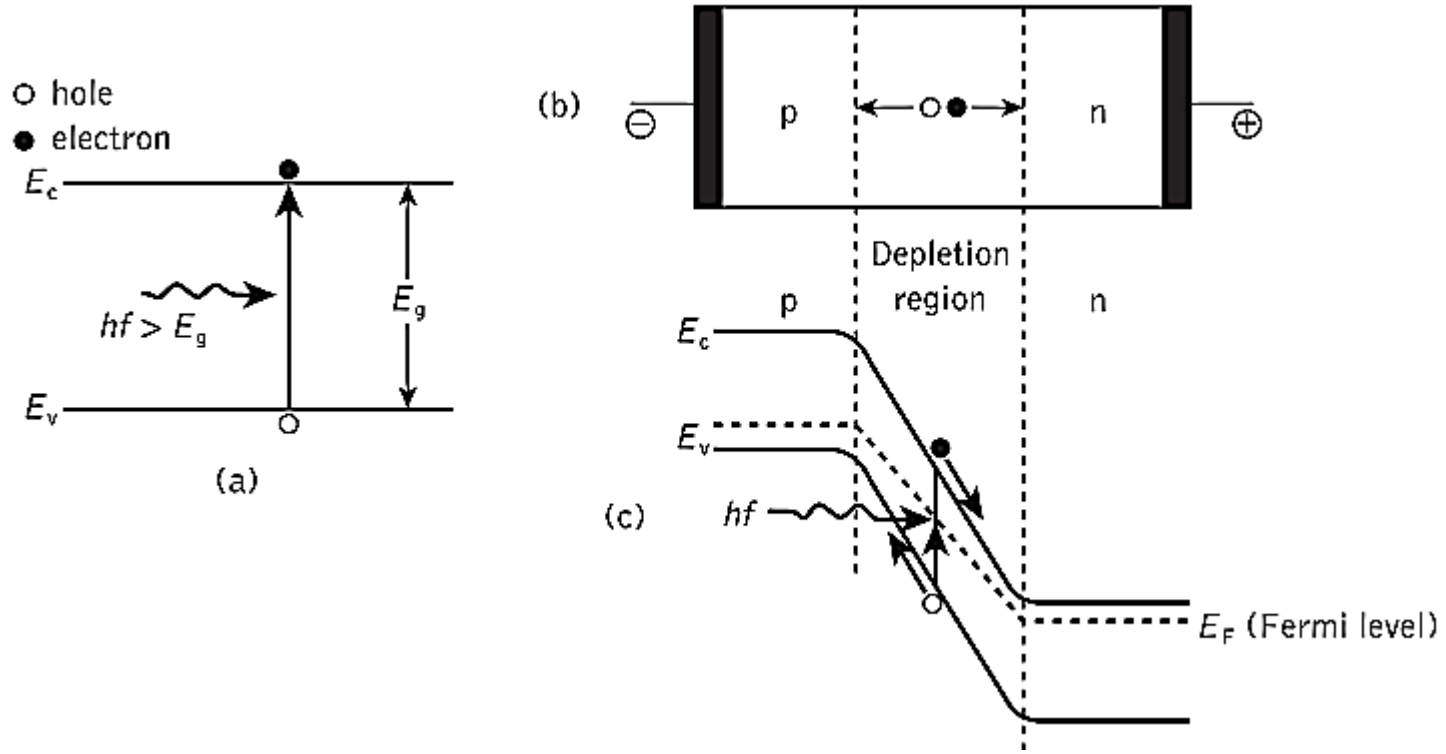


Figure 1.4: Operation of the *p–n* photodiode: (a) photogeneration of an electron–hole pair in an intrinsic semiconductor; (b) the structure of the reverse-biased *p–n* junction illustrating carrier drift in the depletion region; (c) the energy band diagram of the reverse-biased *p–n* junction showing photogeneration and the subsequent separation of an electron–hole pair.

- A photon incident in or near the depletion region of this device which has an energy greater than or equal to the bandgap energy of the fabricating material (i.e.) will excite an electron from the valence band into the conduction band. This process leaves an empty hole in the valence band and is known as the photogeneration of an electron–hole (carrier) pair, as shown in Figure 1.4(a). Carrier pairs so generated near the junction are separated and swept (drift) under the influence of the electric field to produce a displacement by current in the external circuit in excess of any reverse leakage current (Figure 1.4(b)). Photogeneration and the separation of a carrier pair in the depletion region of this reverse-biased p–n junction is illustrated in Figure 1.4 (c).

mit its width. Thus there is a trade-off between the number of photons absorbed (sensitivity) and the speed of response.

- The depletion region must be sufficiently thick to allow a large fraction of the incident light to be absorbed in order to achieve maximum carrier pair generation. However, since long carrier drift times in the depletion region restrict the speed of operation of the photodiode it is necessary to limit its width. Thus there is a trade-off between the number of photons absorbed (sensitivity) and the speed of response.

Lecture No. 3

- **1.4 Absorption**
- 1.4.1 Absorption coefficient
- The absorption of photons in a photodiode to produce carrier pairs and thus a photocurrent is dependent on the absorption coefficient of the light in the semiconductor used to fabricate the device. At a specific wavelength and assuming only bandgap transitions (i.e. intrinsic absorber) the photocurrent produced by incident light of optical power is given by:

$$I_p = \frac{P_0 e(1-r)}{hf}(1 - e^{-\alpha_0 W}) \qquad \ldots(1.1)$$

- where e is the charge on an electron, is the Fresnel reflection coefficient at the semiconductor–air interface and is the width of the absorption region.

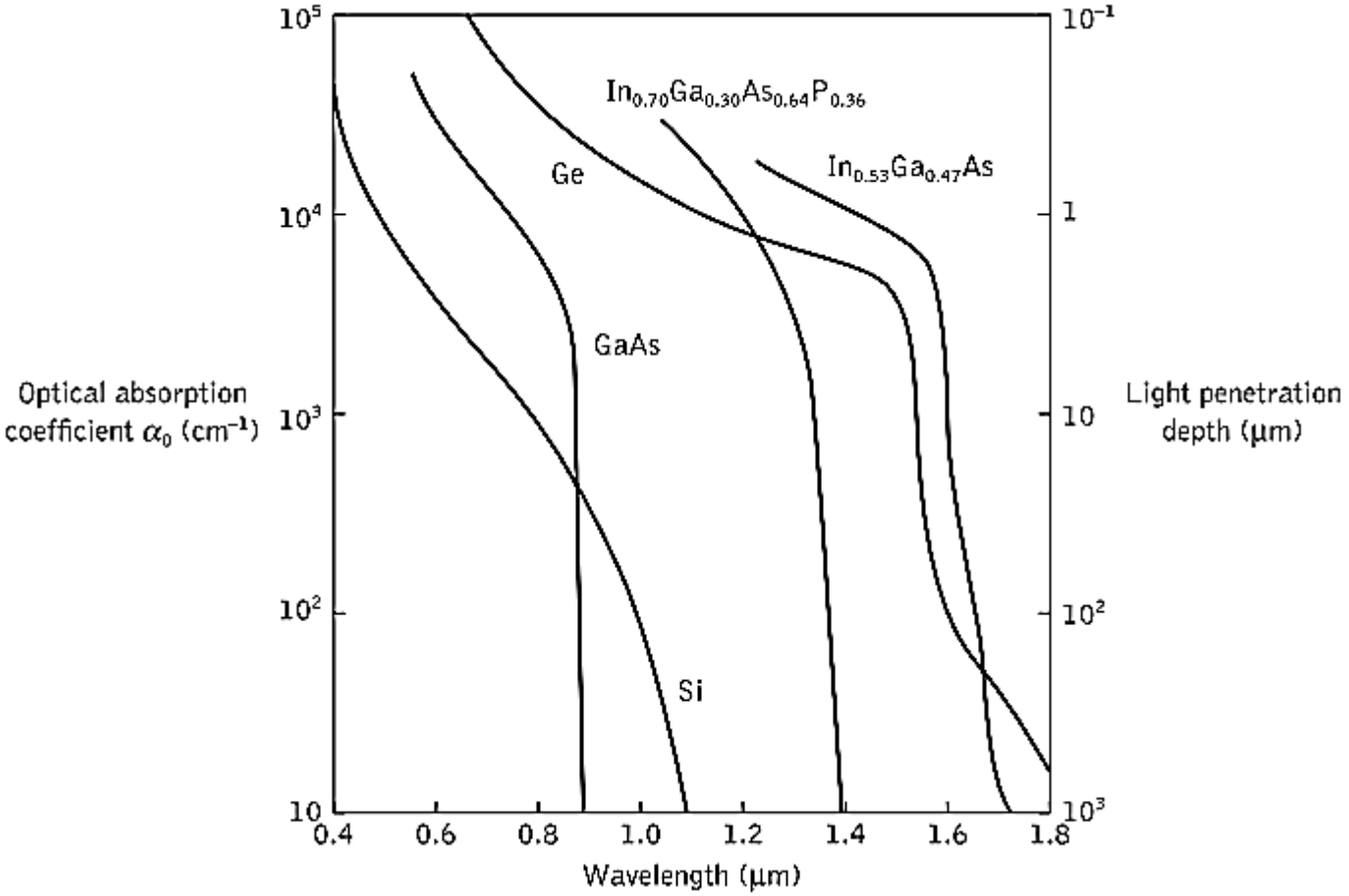


Figure 1.5: Optical absorption curves for some common semiconductor photodiode materials (silicon, germanium, gallium arsenide, indium gallium arsenide and indium gallium arsenide phosphide).

- The absorption coefficients of semiconductor materials are strongly dependent on wavelength. This is illustrated for some common semiconductors in Figure 1.5. It may be observed that there is a variation between the absorption curves for the materials shown and that they are each suitable for different wavelength applications. This results from their differing bandgaps energies, as shown in Table 1.1. However, it must be noted that the curves depicted in Figure 1.5 also vary with temperature.

1.4.2 Direct and indirect absorption: silicon and germanium

- Table 1.1 indicates that silicon and germanium absorb light by both direct and indirect optical transitions. Indirect absorption requires the assistance of a photon so that momentum as well as energy is conserved. This makes the transition probability less

	Bandgap (eV) at 300 K	
	Indirect	*Direct*
Si	1.14	4.10
Ge	0.67	0.81
GaAs	–	1.43
InAs	–	0.35
InP	–	1.35
GaSb	–	0.73
$In_{0.53}Ga_{0.47}As$	–	0.75
$In_{0.14}Ga_{0.86}As$	–	1.15
$GaAs_{0.88}Sb_{0.12}$	–	1.15

Table 1.1 Bandgaps for some semiconductor photodiode materials at 300 K

- likely for indirect absorption than for direct absorption where no photon is involved. In this context direct and indirect absorption may be contrasted with direct and indirect emission discussed in Sections. Therefore, as may be seen from Figure 1.5, silicon is only weakly absorbing over the wavelength band of interest in optical fiber communications (i.e. first-generation 0.8 to 0.9). This is because transitions over this wavelength band in silicon are due only to the indirect absorption mechanism. As mentioned previously (Section 1.2) the threshold for indirect absorption occurs at 1.09 . The bandgap for direct absorption in silicon is 4.10 eV, corresponding to a threshold of 0.30 in the ultraviolet, and thus is well outside the wavelength range of interest.

- Germanium is another semiconductor material for which the lowest energy absorption takes place by indirect optical transitions. However, the threshold for direct absorption occurs at 1.53 , below which germanium becomes strongly absorbing, corresponding to the kink in the characteristic shown in Figure 1.5. Thus germanium may be used in the fabrication of detectors over the whole of the wavelength range of interest (i.e. first- and second-generation 0.8 to 1.6), especially considering that indirect absorption will occur up to a threshold of 1.85 . Ideally, a photodiode material should be chosen with a bandgap energy slightly less than the photon energy corresponding to the longest operating wavelength of the system. This gives a sufficiently high absorption coefficient to ensure a good response, and yet limits the number of thermally generated carriers in order to achieve a low dark current (i.e. displacement current generated with no incident light (see Figure 1.9)). Germanium photodiodes have relatively large dark currents due to their narrow bandgaps in comparison with other semiconductor materials. This is a major disadvantage with the use of germanium photodiodes, especially at shorter wavelengths (below 1.1).

1.4.3 III–V alloys

- The drawback with germanium as a fabricating material for semiconductor photodiodes has led to increased investigation of direct bandgap III–V alloys for the longer wavelength region. These materials are potentially superior to germanium because their bandgaps can be tailored to the desired wavelength by changing the relative concentrations of their constituents, resulting in lower dark currents. They may also be fabricated in heterojunction structures which enhances their high-speed operations.

- Ternary alloys such as InGaAs and GaAlSb deposited on InP and GaSb substrates, respectively, have been used to fabricate photodiodes for the longer wavelength band. Although difficulties were experienced in the growth of these alloys, with lattice matching causing increased dark currents, these problems have now been reduced. In particular the alloy In0.53Ga0.47As lattice matched to InP, which responds to wavelengths up to around 1.7 (see Figure 8.2), has been extensively utilized in the fabrication of photodiodes for operation at both 1.3 and 1.55 μm. Quaternary alloys can also be used for detection at these wavelengths. Both InGaAsP grown on InP and GaAlAsSb grown on GaSb have been studied, with the former material system finding significant application within advanced photodiode structures.

Lecture No. 4

- **1.5 Quantum efficiency**
- The quantum efficiency is defined as the fraction of incident photons which are absorbed by the photodetector and generate electrons which are collected at the detector terminals:

$$\eta = \frac{\text{number of electrons collected}}{\text{number of incident photons}} \qquad \ldots (1.2)$$

- Hence:

$$\eta = \frac{\mathrm{r}_e}{\mathrm{r}_p} \qquad \ldots(1.3)$$

where is the incident photon rate (photons per second) and is the corresponding electron rate (electrons per second).The dependence of on enters through the absorption coefficient . If the facets of the semiconductor slab in Fig. (1.6) are assumed to have an antireflection coating, the power transmitted through the slab of width *W* is .

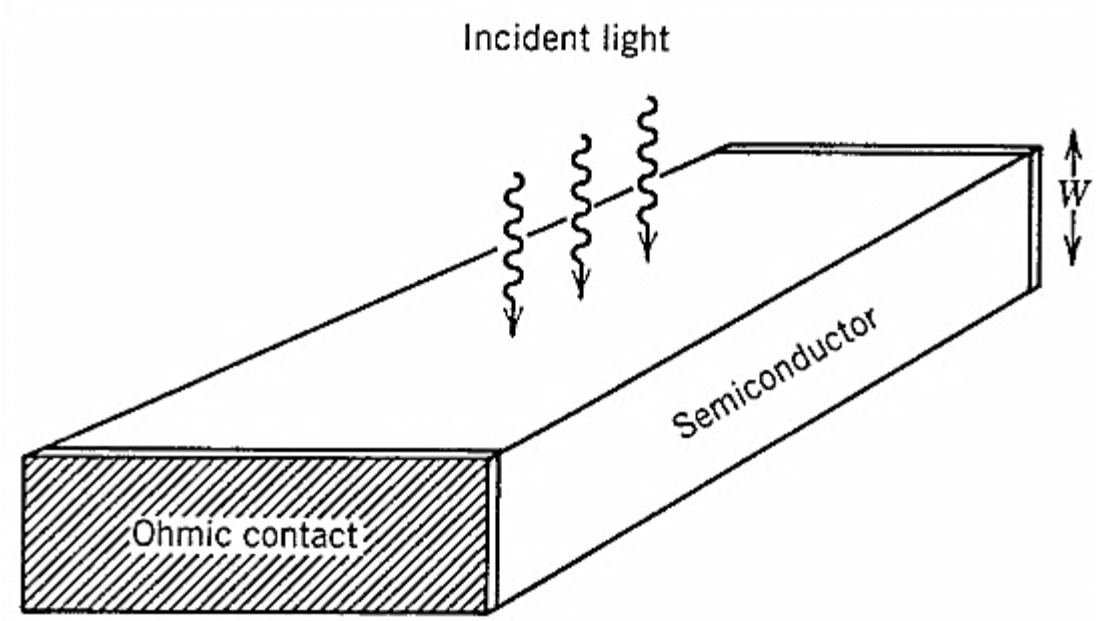


Figure 1.6: A semiconductor slab used as a photodetector.

- The absorbed power is thus is given by

$$P_{abs} = P_{in} - P_{tr}$$

$$= P_{in} - P_{in} e^{-\alpha_0 W}$$

$$= P_{in}[1 - e^{-\alpha_0 W}] \qquad \ldots (1.4)$$

- Since each absorbed photon creates an electron–hole pair, the quantum efficiency is given by

$$\eta = \frac{P_{abs}}{P_{in}} = 1 - e^{-\alpha_0 W} \quad \ldots (1.5)$$

Here, is called absorption factor. As expected, becomes zero when . On the other hand, approaches 1 if.

To prove the relation , we assume that have incident light itswavelength and incident power as shown in Fig. (1.7)

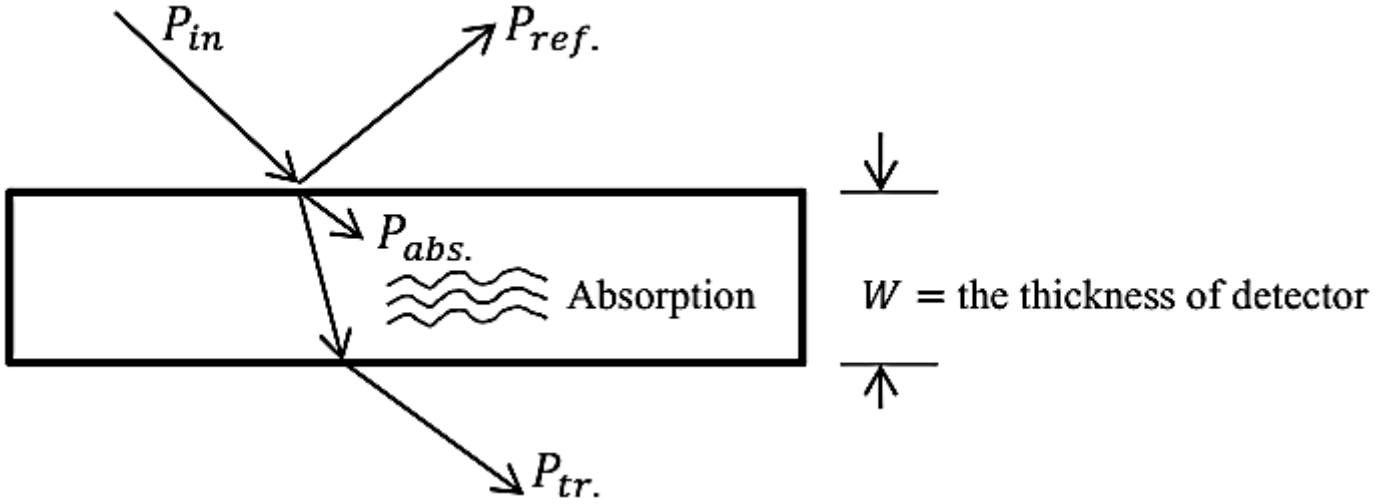


Fig. 1.7: shows the power incident, absorbed, and transmitted light of sample photodetector.

$$\frac{dP_{in}}{dW} = -\alpha_0(\lambda)P_{in}$$

$$\int \frac{dP_{in}}{P_{in}} = \int -\alpha_0(\lambda)\,dW$$

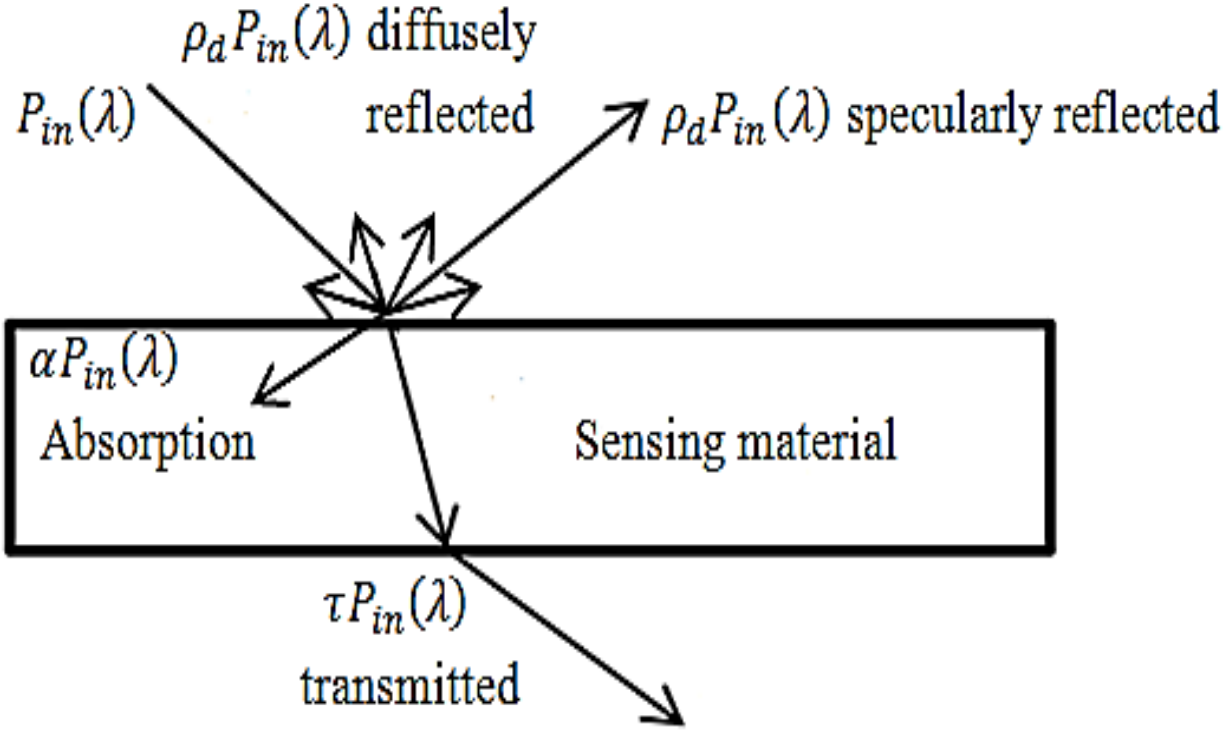
Let assume $P_{in}(\lambda) = \phi_\lambda$, So

$$\int_{\phi_0=P_{in}}^{=P_{tr.}} \frac{d\phi_\lambda}{\phi_\lambda} = -\alpha_0 W \quad \Rightarrow \quad \ln \phi_\lambda \Big|_{\Phi_0=P_{in}}^{\phi=P_{tr.}} = -\alpha_0 W$$

$$\ln \phi - \ln \phi_0 = -\alpha_0 W \quad \Rightarrow \quad \ln \frac{\phi}{\phi_0} = -\alpha_0 W$$

Taking *exponential* of both side of last equation, leads to

$$P_{tr} = P_{in} e^{-\alpha W}$$


$\rho_d P_{in}(\lambda)$ diffusely
reflected
$P_{in}(\lambda)$
$\rho_d P_{in}(\lambda)$ specularly reflected
$\alpha P_{in}(\lambda)$
Absorption
Sensing material
$\tau P_{in}(\lambda)$
transmitted

Lecture No. 5

- **1.6 Responsivity**
- The expression for quantum efficiency does not involve photon energy and therefore the responsivity R is often of more use when characterizing the performance of a photodetector. It is defined as:

$$R = \frac{I_p}{P_0} (A\,W^{-1}) \qquad \ldots (1.6)$$

- where is the output photocurrent in amperes and is the incident optical power in watts (i.e. output optical power from the fiber). The responsivity is a useful parameter as it gives the transfer characteristic of the detector (i.e. photocurrent per unit incident optical power). The relationship for responsivity (Eq. (1.6)) may be developed to include quantum efficiency as follows. Considering the energy of a photon , where h = 6.626 × J s is Planck's constant. Thus the incident photon rate may be written in terms of incident optical power and the photon energy as:

$$r_p = \frac{P_0}{hf} \qquad \ldots (1.7)$$

In Eq. (1.3) the electron rate is given by:

$$r_e = \eta r_p \qquad \ldots (1.8)$$

Substituting from Eq. (1.7) we obtain:

$$r_e = \frac{\eta P_0}{hf} \qquad \ldots (1.9)$$

Therefore, the output photocurrent is:

$$I_p = \frac{\eta P_0 e}{hf} \qquad \ldots (1.10)$$

where e is the charge on an electron. Thus from Eq. (1.6) the responsivity may be written as:

$$R = \frac{\eta e}{hf} \qquad \ldots (1.11)$$

Equation (1.11) is useful relationships for responsivity which may be developed a stage further to include the wavelength of the incident light. The frequency f of the incident photons is related to their wavelength λ and the velocity of light in air c, by:

$$f = \frac{c}{\lambda} \qquad \ldots (1.12)$$

Substituting into Eq. (1.11) a final expression for the responsivity is given by:

$$R = \frac{\eta e \lambda}{hc} \approx \frac{\eta \lambda}{1.24} \qquad \ldots (1.13)$$

It may be noted that the responsivity is directly proportional to the quantum efficiency at a particular wavelength.

The ideal responsivity against wavelength characteristic for a silicon photodiode with unit quantum efficiency is illustrated in Figure 1.7(a). Also shown is the typical responsivity of a practical silicon device.

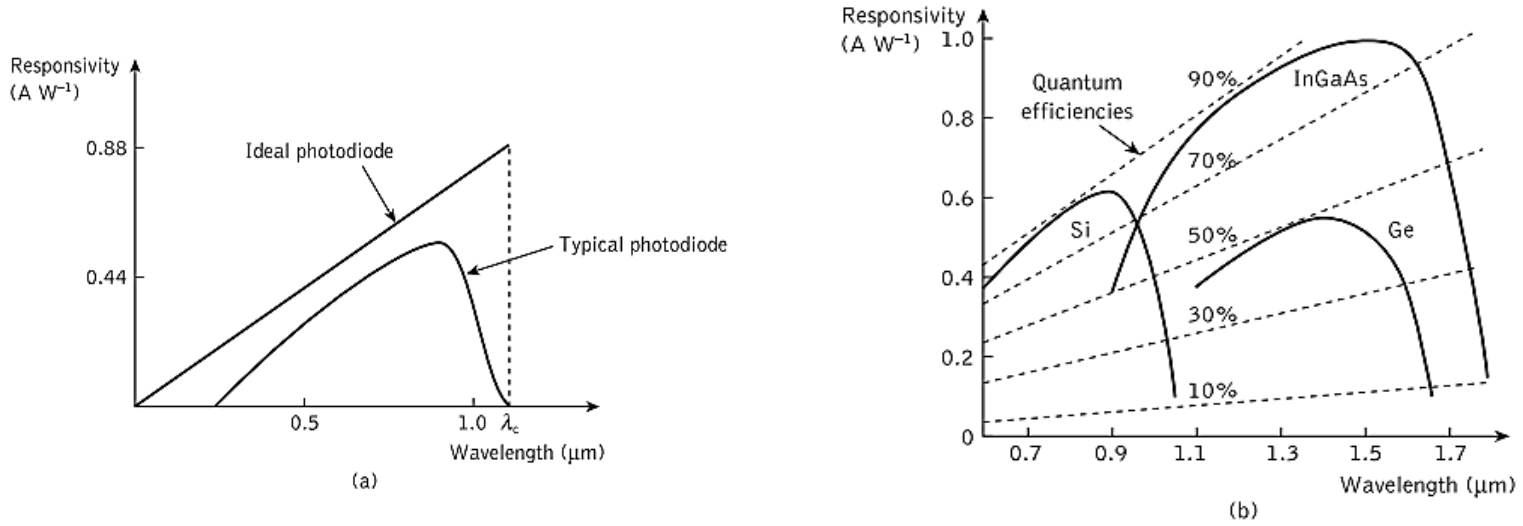


Figure 1.7: Responsivity against wavelength characteristics: (a) an ideal silicon photodiode with a typical device also shown; (b) silicon, germanium and InGaAs photodiodes with quantum efficiencies also shown.

- Figure 1.7(b), however, compares the responsivities and quantum efficiencies of the photodiodes based on silicon, germanium and the InGaAs ternary alloy.It shows the lower values of responsivity of 0.45 and at signal wavelengths of 0.90 and 1.30 , respectively, for silicon and germanium photodiodes. High responsivity values of 0.9 and 1.0 at signal wavelengths of 1.30 and 1.55 μm, respectively, for the photodiode from InGaAs alloy can also be observed.

- Moreover nearly 90% quantum efficiencies can be obtained for both the InGaAsand silicon photodiodes. It should also be noted that the responsivity drops rapidly at the cutoff wavelength for each of the photodiode materials. This factor is in accordance with Eq. (1.13) which provides the quantum efficiency as a function of signal wavelength which is critically dependent on the photodiode material bandgap energy. For a particular material, as the wavelength of the incident photon becomes longer the photon energy eventually is less than the energy required to excite an electron from the valance band to the conduction band and at this point the responsivity falls to zero.

- **Example 1.1**
- When photons each with a wavelength of 0.85 are incident on a photodiode, on average electrons are collected at the terminals of the device. Determine the quantum efficiency and the responsivity of the photodiode at 0.85 .
- *Solution*: From Eq. (1.2):

$$\eta = \frac{\text{number of electrons collected}}{\text{number of incident photons}}$$

$$= \frac{1.2 \times 10^{11}}{3 \times 10^{11}} = 0.4$$

The quantum efficiency of the photodiode at 0.85 is 40%. From Eq. (1.13): μm

$$\text{Responsivity } R = \frac{\eta e \lambda}{hc} = \frac{0.4 \times 1.602 \times 10^{-19} \times 0.85 \times 10^{-6}}{6.626 \times 10^{-34} \times 2.998 \times 10^{8}}$$

$$= 0.274\, A\, W^{-1}$$

The responsivity of the photodiode at 0.85 is $0.27\, A\, W^{-1}$ μm

- **Example 1.2**
- A photodiode has a quantum efficiency of 65% when photons of energy are incident upon it.
- (a) At what wavelength is the photodiode operating?
- (b) Calculate the incident optical power required to obtain a photocurrent of 2.5 when the photodiode is operating as described above.
- *Solution*: (a) From the photon energy Therefore:

$$\lambda = \frac{hc}{E} = \frac{6.626 \times 10^{-34} \times 2.998 \times 10^{8}}{1.5 \times 10^{-19}} = 1.32\,\mu m$$

The photodiode is operating at a wavelength of 1.32 μm.

(b) From Eq. (1.11):

$$\text{Responsivity} R = \frac{\eta e}{hf} = \frac{0.65 \times 1.602 \times 10^{-19}}{1.5 \times 10^{-19}}$$

$$= 0.694\ A\ W^{-1}$$

Also from Eq. (1.4):

$$P_o = \frac{I_p}{R} = \frac{25 \times 10^{-6}}{0.694} = 3.6\,\mu W$$

The incident optical power required is 3.60 μW

Lecture No. 6

- **1.7 Long-wavelength cutoff**
- It is essential when considering the intrinsic absorption process that the energy of incident photons be greater than or equal to the bandgap energy of the material used to fabricate the photodetector. Therefore, the photon energy:

$$\frac{hc}{\lambda} \geq E_g \qquad \ldots (1.14)$$

giving:

$$\lambda \leq \frac{hc}{E_g} \qquad \ldots (1.15)$$

Thus the threshold for detection, commonly known as the long-wavelength cutoff point λ_c, is:

$$\lambda_c = \frac{hc}{E_g} \qquad \ldots (1.16)$$

The expression given in Eq. (1.16) allows the calculation of the longest wavelength of light to give photodetection for the various semiconductor materials used in the fabrication of detectors.

It is important to note that the above criterion is only applicable to intrinsic photodetectors. Extrinsic photodetectors violate the expression given in Eq. (1.14), but are not currently used in optical fiber communications.

Example 1.3

GaAs has a bandgap energy of 1.43 eV at 300 K. Determine the wavelength above which an intrinsic photodetector fabricated from this material will cease to operate.

Solution: From Eq. (1.16), the long wavelength cutoff:

$$\lambda_c = \frac{hc}{E_g}$$

$$= \frac{6.626 \times 10^{-34} \times 2.998 \times 10^{8}}{1.43 \times 1.602 \times 10^{-19}} = 0.867\ \mu m.$$

The GaAs photodetector will cease to operate above 0.87 .

1.8 Semiconductor photodiodes without internal gain

- Semiconductor photodiodes without internal gain generate a single electron–hole pair per absorbed photon. This mechanism was outlined in Section 1.3, and in order to understand the development of this type of photodiode it is now necessary to elaborate upon it.

Lecture No. 7

- **1.8.1 The *p–n* photodiode**
- Figure 1.8 shows a reverse-biased p–n photodiode with both the depletion and diffusion regions. The depletion region is formed by immobile positively charged donor atoms in the n-type semiconductor material and immobile negatively charged acceptor atoms in the p-type material, when the mobile carriers are swept to their majority sides under the influence of the electric field. The width of the depletion region is therefore dependent upon the doping concentrations for a given applied reverse bias (i.e. the lower the doping, the wider the depletion region).

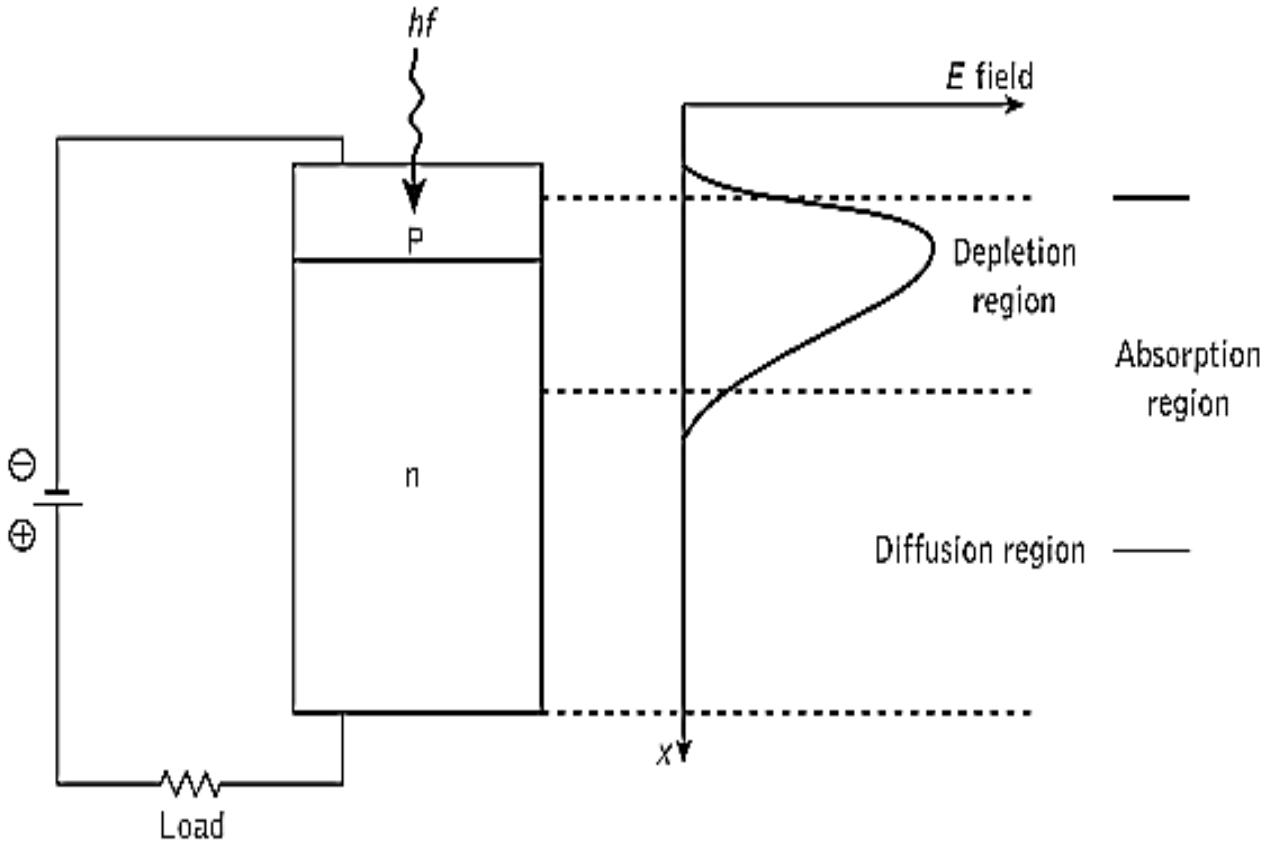


Figure 1.8 The *p–n* photodiode showing depletion and diffusion regions.

- Photons may be absorbed in both the depletion and diffusion regions, as indicated by the absorption region in Figure 1.8. The absorption region's position and width depend upon the energy of the incident photons and on the material from which the photodiode is fabricated. Thus in the case of the weak absorption of photons, the absorption region may extend completely throughout the device. Electron–hole pairs are therefore generated in both the depletion and diffusion regions. In the depletion region the carrier pairs separate and drift under the influence of the electric field, whereas outside this region the hole diffuses towards the depletion region in order to be collected. The diffusion process is very slow compared with drift and thus limits the response of the photodiode.

- It is therefore important that the photons are absorbed in the depletion region. Thus it is made as long as possible by decreasing the doping in the n-type material. The depletion region width in a p–n photodiode is normally 1 to 3 and is optimized for the efficient detection of light at a given wavelength. For silicon devices this is in the visible spectrum (0.4 to 0.7) and for germanium in the near infrared (0.7 to 0.9). Typical output characteristics for the reverse-biased p–n photodiode are illustrated in Figure 1.9. The different operating conditions may be noted moving from no light input to a high light level.

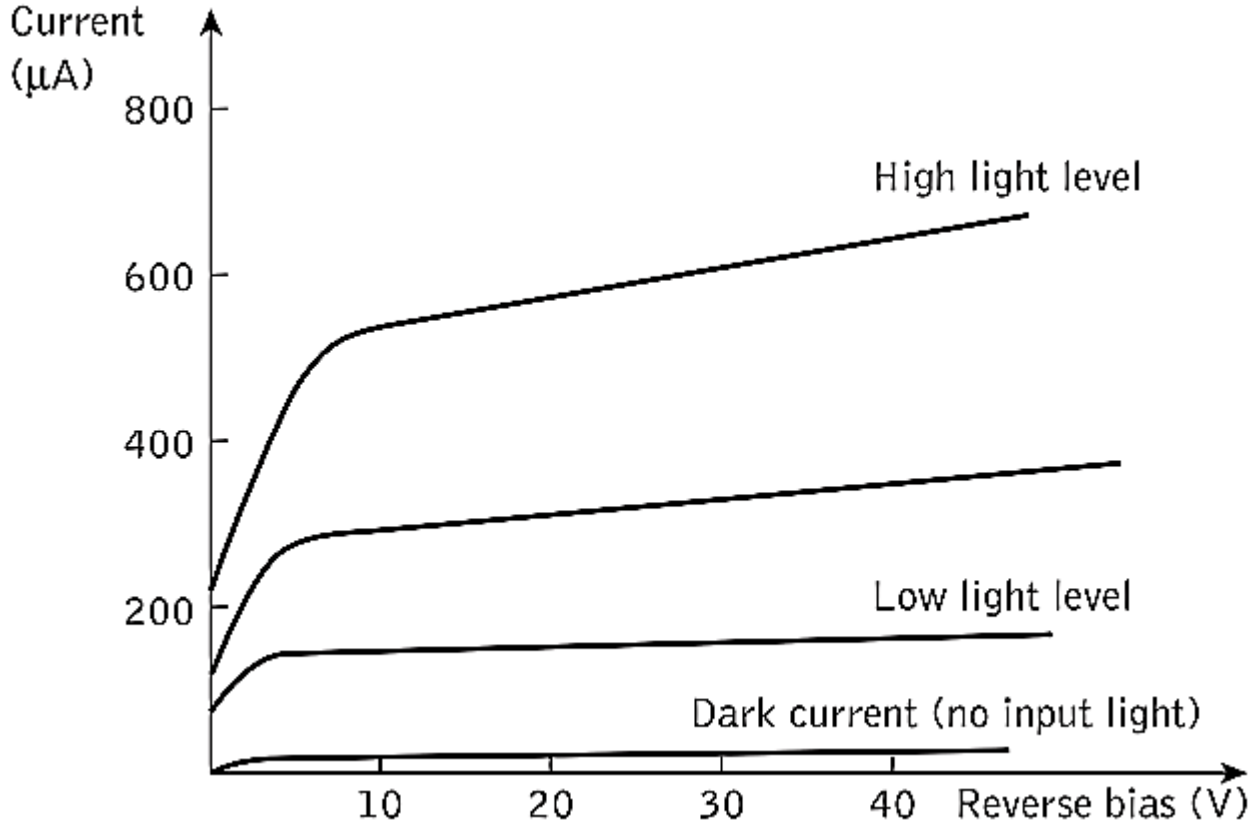


Figure 1.9 Typical *p–n* photodiode output characteristics

1.8.2 The *p–i–n* photodiode

- In order to allow operation at longer wavelengths where the light penetrates more deeply into the semiconductor material, a wider depletion region is necessary. To achieve this the n-type material is doped so lightly that it can be considered intrinsic, and to make a lowresistance contact a highly doped n-type () layer is added. This creates a p–i–n (or PIN) structure, as may be seen in Figure 1.10 where all the absorption takes place in the depletion region.

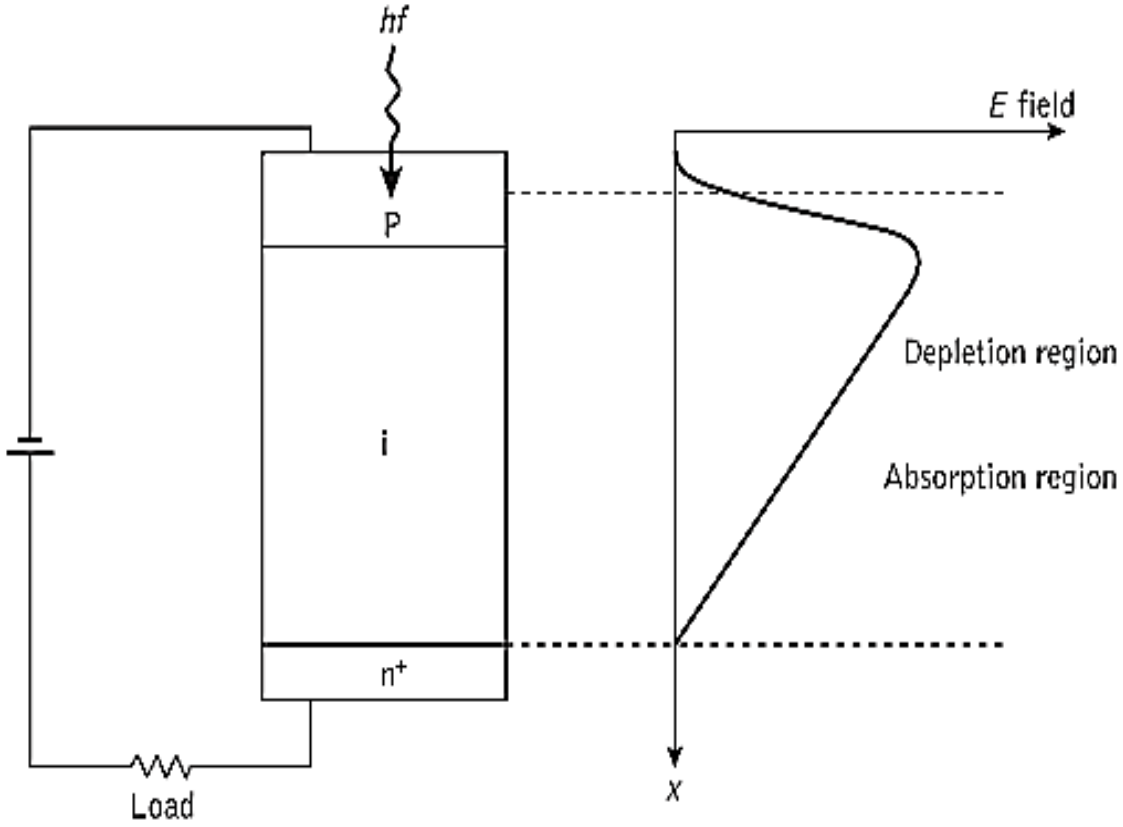
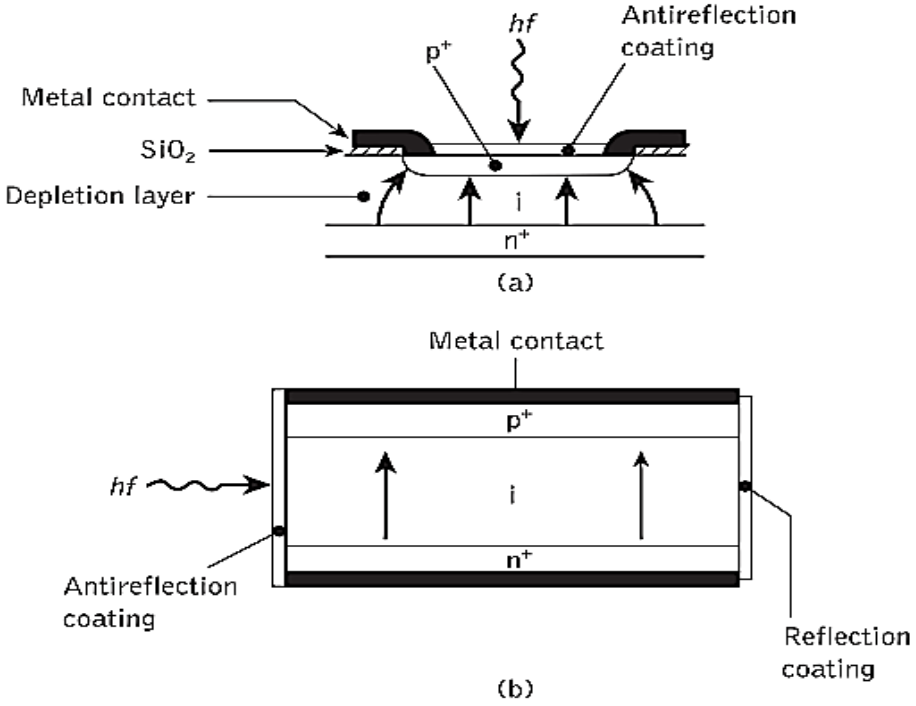


Figure 1.10 The *p–i–n* photodiode showing the combined absorption and depletion region.

- Figure 1.11 shows the structures of two types of silicon p–i–n photodiode for operation in the shorter wavelength band below 1.09 .



- **Figure 1.11**(a) Structure of a front-illuminated silicon *p–i–n* photodiode. (b) Structureof a side-illuminated (parallel to junction) *p–i–n* photodiode.

- The front-illuminated photodiode, when operating in the 0.8 to 0.9 band (Figure 1.11(a)), requires a depletion region of between 20 and 50 in order to attain high quantum efficiency (typically 85%) together with fast response (less than 1 ns) and low dark current (1 nA). Dark current arises from surface leakage currents as well as generation–recombination currents in the depletion region in the absence of illumination. The side-illuminated structure (Figure 1.11(b)), where light is injected parallel to the junction plane, exhibits a large absorption width (500) and hence is particularly sensitive at wavelengths close to the bandgap limit (1.09) where the absorption coefficient is relatively small.

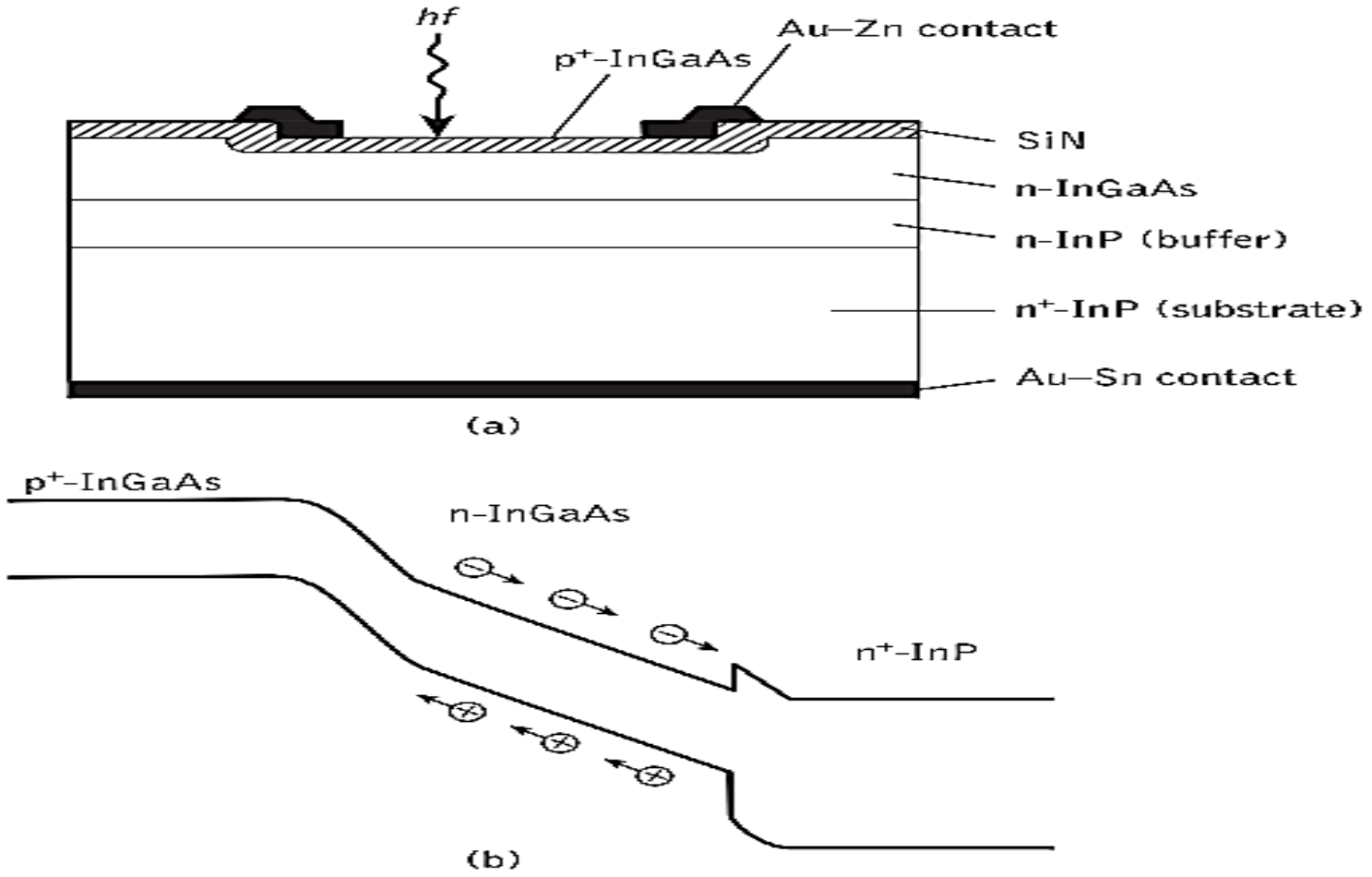


Figure 1.12 Planar InGaAs*p–i–n* photodiode: (a) structure; (b) energy band diagram showing homojunction associated with the conventional *p–i–n* structure.

- Germanium p–i–n photodiodes which span the entire wavelength range of interest are also commercially available, but as mentioned previously the relatively high dark currents are a problem (typically 100 nA at 20 °C increasing to 1 A at 40 °C).However, III–V semiconductor alloys have been employed in the fabrication of longer wavelength region detectors. The favored material is the lattice-matched In 0.53Ga0.47As/InP system which can detect at wavelengths up to 1.67 . A typical planar device structure is shown in Figure 1.12(a) which requires epitaxial growth of several layers on an n-type InP substrate. .The incident light is absorbed in the low-doped n-type InGaAs layer generating carriers, as illustrated in the energy band diagram Figure 1.12(b). The discontinuity due to the homojunction between the n+-InP substrate and the n-InGaAs absorption region may be noted. This can be reduced by the incorporation of an n-type InP buffer layer.

- The top entry (also referred to as front illumination) device shown in Figure 1.12(a) is the simplest structure, with the light being introduced through the upper -layer. However, a drawback with this structure is a quantum efficiency penalty which results from optical absorption in the undepleted- region. In addition, there is a limit to how small such a device can be fabricated as both light access and metallic contact are required on the top. To enable smaller devices with lower capacitances to be made, a substrate entry technique is employed. In this case light enters through a transparent InP substrate and the device area can be fabricated as small as may be practical for bonding.

- Conventional growth techniques for III–V semiconductors can be employed to fabricate these devices, although liquid-phase epitaxy (LPE) tends to be preferred because of the relative ease in obtaining the low doping levels needed (around) to obtain low capacitance (less than 0.2 pF). However, LPE does not easily allow low-impurity-level concentrations and it is necessary to use long baking procedures over several days to purify the source material. High-quality devices have been produced using metal oxide vapor-phase epitaxy (MOVPE), a technique which appears much more appropriate for large-scale production of such devices.

- A substrate entry (also referred to as back illumination) p–i–n photodiode is shown in Figure 1.13(a). This device incorporates a -InGaAsP layer to provide a heterojunction structure (Schottky barrier) which improves quantum efficiency. Moreover, it is fabricated as a mesa structure which reduces parasitic capacitances. Unfortunately, charge trapping can occur at the n–-InGaAs/ InGaAsP interface which may be observed in the energy band diagram of Figure 1.13(b). This may cause limitations in the response time of the device. However, small-area substrate entry devices can be produced with extremely low capacitance (less than 0.1 pF), quantum efficiency between 75% and 100% and dark currents less than 1 nA.

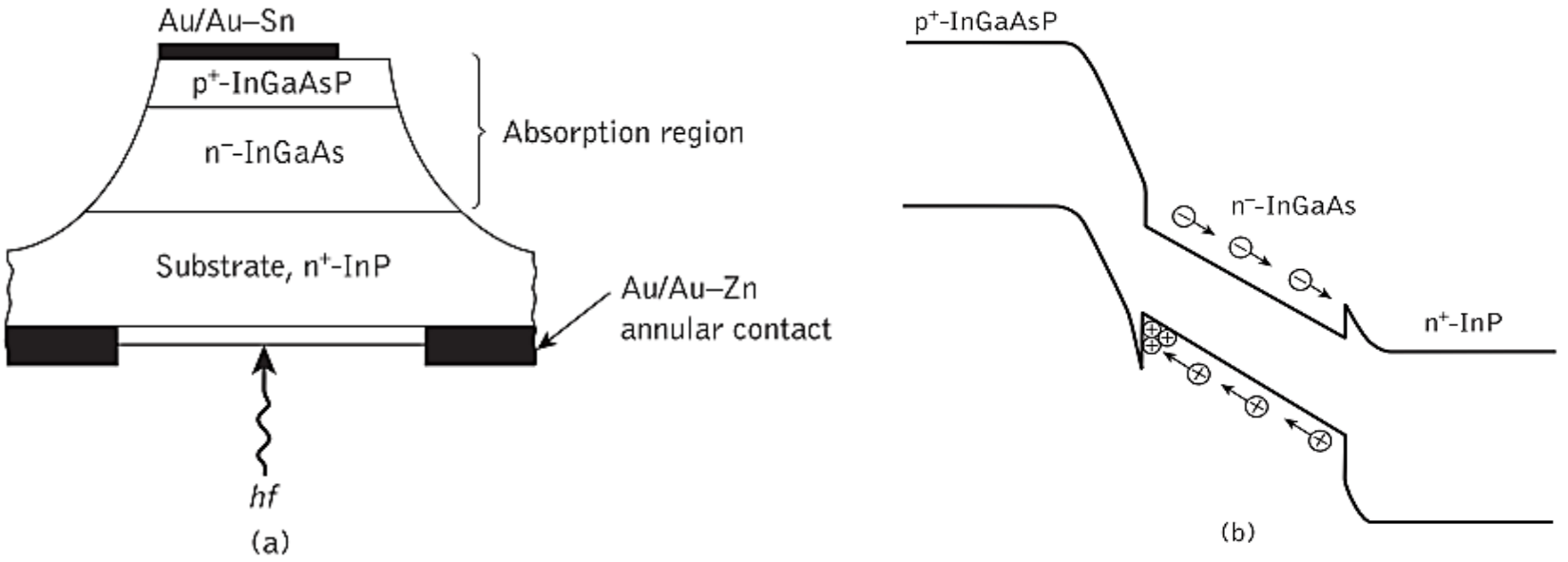


Figure 1.13 Substrate entry InGaAs*p–i–n* photodiode: (a) structure; (b) energy band diagram illustrating the heterojunction and charge trapping.

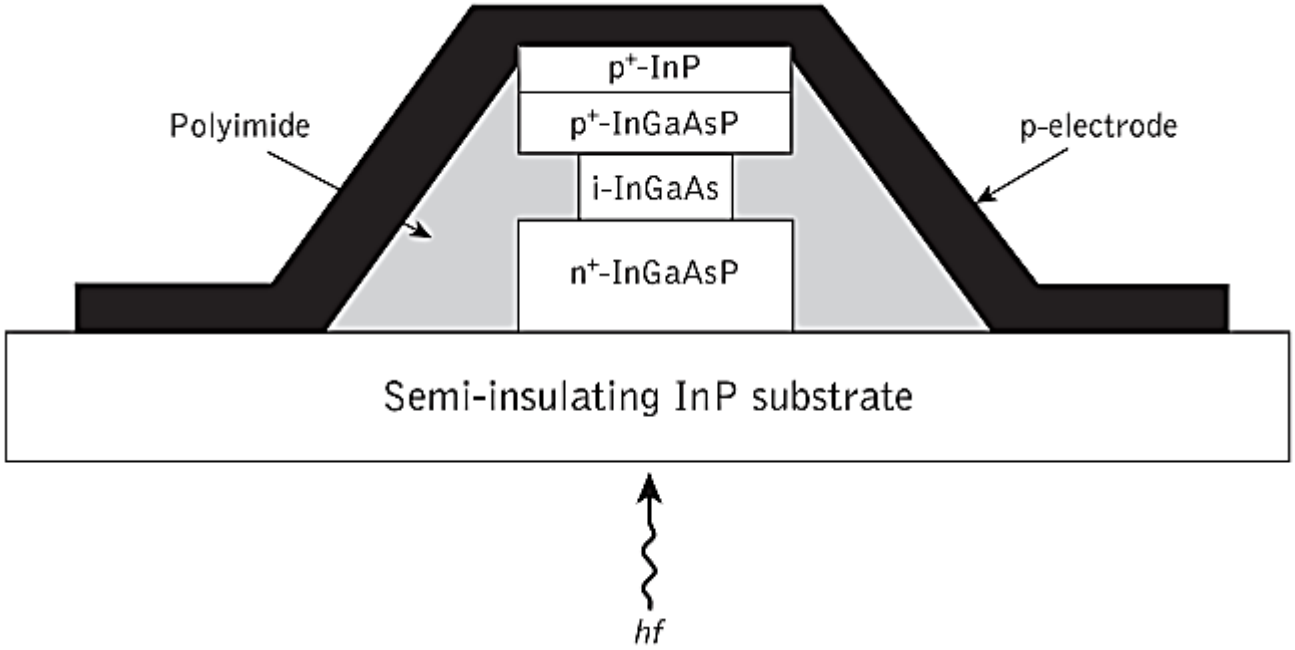


Figure 1.14 Structure of a mushroom waveguide photodiode

- In both device types a depleted InGaAs layer of around 3 is used which provides high quantum efficiency and bandwidth. Furthermore, low doping permits full depletion of the InGaAs layer at low voltage (5 V). The short transit times in the relatively narrow depletion layers give a theoretical bandwidth of approximately 15 GHz. However, the bandwidth of commercially available packaged detectors is usually between 1 and 2 GHz due to limitations of the packaging.

- A photodiode containing a waveguide structure, known as a mushroom waveguide, can, however, be used to overcome the bandwidth–quantum efficiency trade-off between the device capacitance and contact resistance. This structure, which is illustrated in Figure 1.14, comprises a thin layer of InGaAs (thickness of 0.20) used as the absorption material which is lattice matched to an InP substrate thus providing operation at a wavelength of 1.55 . It may be observed that two graded layers of InGaAsP material, each having a thickness of 0.80 , are also employed above and below the absorption layer to avoid charge trapping.

- Since the device is side illuminated its quantum efficiency is therefore a function of the length of the absorption layer and also the thickness of this layer determines the amount of electron drift time. Thus a long and thinabsorption layer provides both high quantum efficiency and fast response times. High-speed operation up to 110 GHz with 50% quantum efficiency using such structures been demonstrated. It should also be noted that in the mushroom waveguide structure the light and the carriers travel in different directions and therefore the device bandwidth and the quantum efficiency are not too dependent on each other. Hence quantum efficiencies of greater than 80% at a bandwidth of 10 GHz have been obtained using this waveguide structure.

Lecture No. 8

- **1.8.3 Speed of response and traveling-wave photodiodes**
- Three main factors limit the speed of response of a photodiode. These are:
- *Drift time of carriers through the depletion region*. The speed of response of a photodiode is fundamentally limited by the time it takes photogenerated carriers to drift across the depletion region. When the field in the depletion region exceeds a saturation value, the carriers may be assumed to travel at a constant (maximum) drift velocity . The longest transit time, , is for carriers which must traverse the full depletion layer width and is given by:

$$t_{\text{drift}} = \frac{W}{v_d} \qquad \ldots(1.17)$$

- A field strength above 2 × V in silicon gives maximum (saturated) carriermvelocities of approximately cm . Thus the transit time through a depletion layer width of 10 µm is around 0.1 ns.
- 2. *Diffusion time of carriers generated outside the depletion region*. Carrier diffusion is a comparatively slow process where the time taken, , for carriers to diffuse a distance may be written as:

$$t_{diff} = \frac{d^2}{2D_c} \qquad \ldots(1.18)$$

Where is the minority carrier diffusion coefficient. For example, the hole diffusion time through 10 m of silicon is 40 ns whereas the electron diffusion time over a similar distance is around 8 ns.

- 3. *Time constant incurred by the capacitance of the photodiode with its load.* A reversebiased photodiode exhibits a voltage-dependent capacitance caused by the variation in the stored charge at the junction. The junction capacitance is given by:

$$C_j = \frac{\mathcal{E}_s A}{w} \qquad \ldots(1.19)$$

Where is the permittivity of the semiconductor material and is the diode junction area. Hence, a small depletion layer width w increases the junction capacitance.

- The capacitance of the photodiode is that of the junction together with the capacitance of the leads and packaging. This capacitance must be minimized in
order to reduce the RC time constant which also limits the detector response time.

- **Example 1.4**
- A silicon p–i–n photodiode has an intrinsic region with a width of 20 m and a diameter of 500 μm in which the drift velocity of electrons is 105 m . When the permittivity of the device material is 10.5 × F , calculate: (a) the drift time of the carriers across the depletion region; (b) the junction capacitance of the photodiode.
- *Solution*: (a) The drift time for the carriers across the depletion region for the photodiode can be obtained using Eq. (1.17) as:

$$t_{\text{drift}} = \frac{w}{v_d}$$
$$= \frac{20 \times 10^{-6}}{1 \times 10^{5}}$$
$$= 2 \times 10^{-10} s$$

The drift time for the carriers across the depletion region is therefore 200 ps.

(b) The junction capacitance is given by Eq. (1.19) as:

$$C_j = \frac{\mathcal{E}_s A}{w}$$

where the area $A = \pi \times r^2 = 3.14 \times (500 \times 10^{-6})^2 = 0.79 \times 10^{-6} m^2$. Therefore:

$$C_j = \frac{10.5 \times 10^{-13} \times 0.79 \times 10^{-6}}{2 \times 10^{-6}}$$
$$= 0.41 \times 10^{-13}$$

So, the photodiode has a junction capacitance of 4 pF.

- Although all the above factors affect the response time of the photodiode, the ultimate bandwidth of the device is limited by the drift time of carriers through the depletion region . In this case, when assuming no carriers are generated outside the depletion region and that there is negligible junction capacitance, the maximum photodiode 3 dB bandwidth is given by:

$$B_m = \frac{1}{2\pi t_{drift}} = \frac{v_d}{2\pi w} \qquad \ldots.(1.20)$$

- Moreover, when there is no gain mechanism present within the device structure, the maximum possible quantum efficiency is 100%. Hence the value for the bandwidth given by Eq. (1.20) is also equivalent to the ultimate gain–bandwidth product for the photodiode.

Example 1.5

The carrier velocity in a silicon p–i–n photodiode with a 25 μm depletion layer width is 3×10^4 m s^{-1}. Determine the maximum response time for the device.

Solution: The maximum 3 dB bandwidth for the photodiode may be obtained from Eq. (1.18) where:

$$B_m = \frac{v_d}{2\pi w} = \frac{3 \times 10^4}{2\pi \times 25 \times 10^{-6}} = 1.91 \times 10^8 Hz$$

The maximum response time for the device is therefore:

$$\text{Max. response time} = \frac{1}{B_m} = 5.2 \text{ ns.}$$

It must be noted, however, that the above response time takes no account of the diffusion of carriers in the photodiode or the capacitance associated with the device junction and the external connections.

- The response of a photodiode to a rectangular optical input pulse for various device parameters is illustrated in Figure 1.15. Ideally, to obtain a high quantum efficiency for the photodiode the width of the depletion layer must be far greater than the reciprocal of the absorption coefficient (i.e. 1/) for the material used to fabricate the detector so that most of the incident light will be absorbed. Hence the response to a rectangular input pulse of a low-capacitance photodiode meeting this condition, and exhibiting negligible diffusion outside the depletion region, is shown in Figure 1.15(a). It may be observed in this case that the rising and falling edges of the photodiode output follow the input pulse quite well. When the detector capacitance is larger, however, the speed of response becomes limited by the RC time constant of this capacitance and the load resistor associated with the receiver circuit, and thus the output pulse appears as illustrated in Figure 1.15(b).

- Furthermore, when there is significant diffusion of carriers outside the depletion region, as is the case when the depletion layer is too narrow () and carriers are therefore created by absorption outside this region, then the output pulse displays a long tail caused by the diffusion component to the input optical pulse, as shown in Figure 1.15(c). Thus devices with very thin depletion layers have a tendency to exhibit distinctive fast response and slow response components to their output pulses, as may be observed in Figure 1.15(c), the former response resulting from absorption in the thin depletion layer.

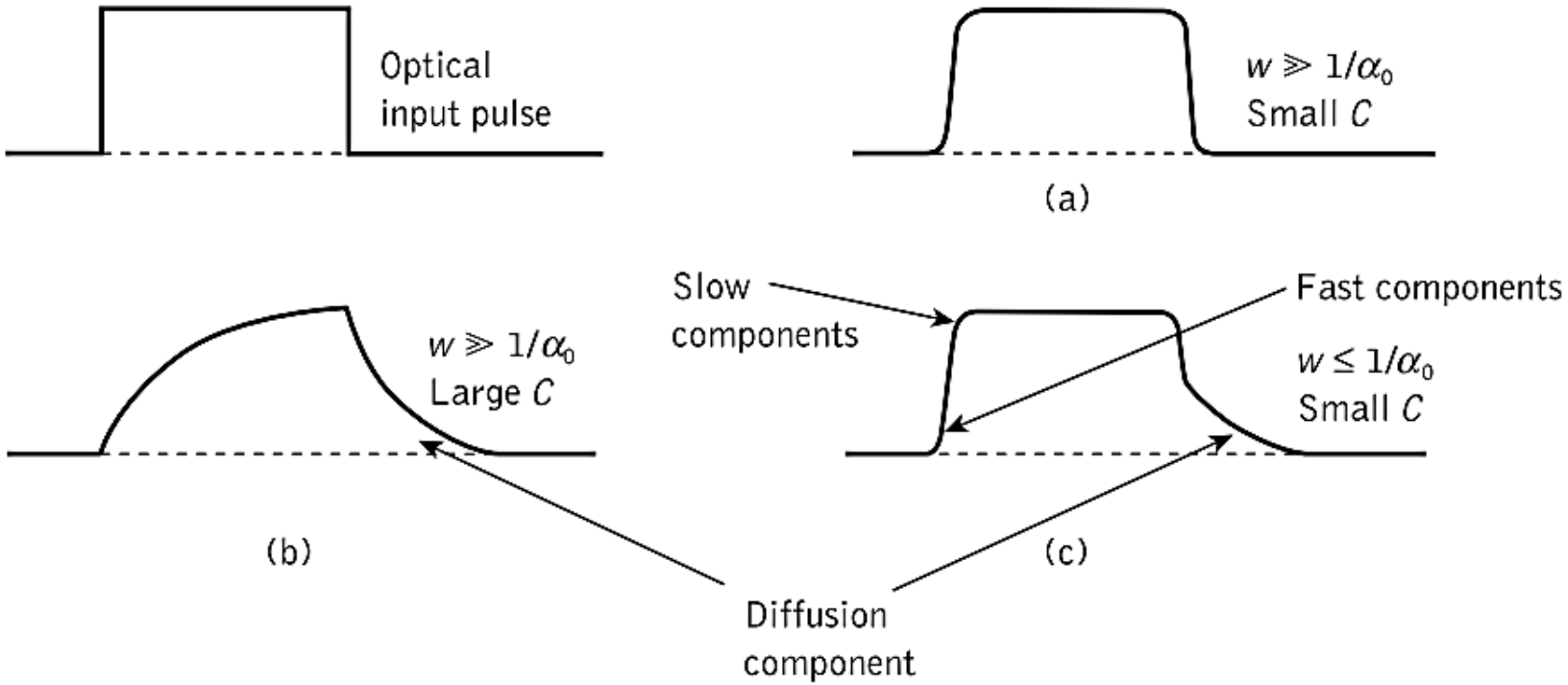


Figure 1.15: Photodiode responses to rectangular optical input pulses for various detector parameters

- A recent approach to reduce the RC time constant limitation is to use a traveling-wave (TW) photodiode structure in which the absorption and carrier drift regions are positioned orthogonally to each other. Such a p–i–n photodiode is illustrated in Figure 1.16(a) where the photogenerated carriers are controlled by the electrical transmission lines and the absorption occurs in an optical waveguide that collects the photogenerated carriers. This approach distributes the capacitance along the electrical transmission lines which can be terminated with a matching impedance thus rendering the bandwidth independent of capacitance. However, a shortcoming of the structure is that both electrical and optical signals do not arrive at the same time due to their mismatched velocities.

- Figure 1.16(b) represents a scheme to match the electrical and optical wave velocities. It consists of TW photodiodes and electrical transmission lines coupled to an optical waveguide. The waveguide geometry and its material composition determine the distribution of the incident power along the device. The group velocity of the optical traveling wave is fixed and therefore the only way to ensure that the velocity matching between optical and electrical waves can be achieved is by tuning the radio-frequency phase velocity (i.e. by varying the electrode dimensions). The same principle is implemented in the photodetector illustrated in Figure 1.16(c) where several photodiodes at regular intervals are produced above the waveguide, underneath the electrical transmission line.

- In this case the absorption occurs in a series of discrete photodiodes (i.e. instead of a single absorption layer) positioned periodically along the optical waveguide. This structure distributes the optical signal power into each high-speed photodiode which is then collected by bringing together the photocurrent from each photodiode on a low-loss electrical transmission line to reduce the microwave loss and to improve the output signal power. It is therefore possible to closely match the velocities of the optical and electrical waves using this generic structure. In addition, high-bandwidth performance up to 190 GHz has been demonstrated using a metal–semiconductor–metal traveling-wave photodetector.

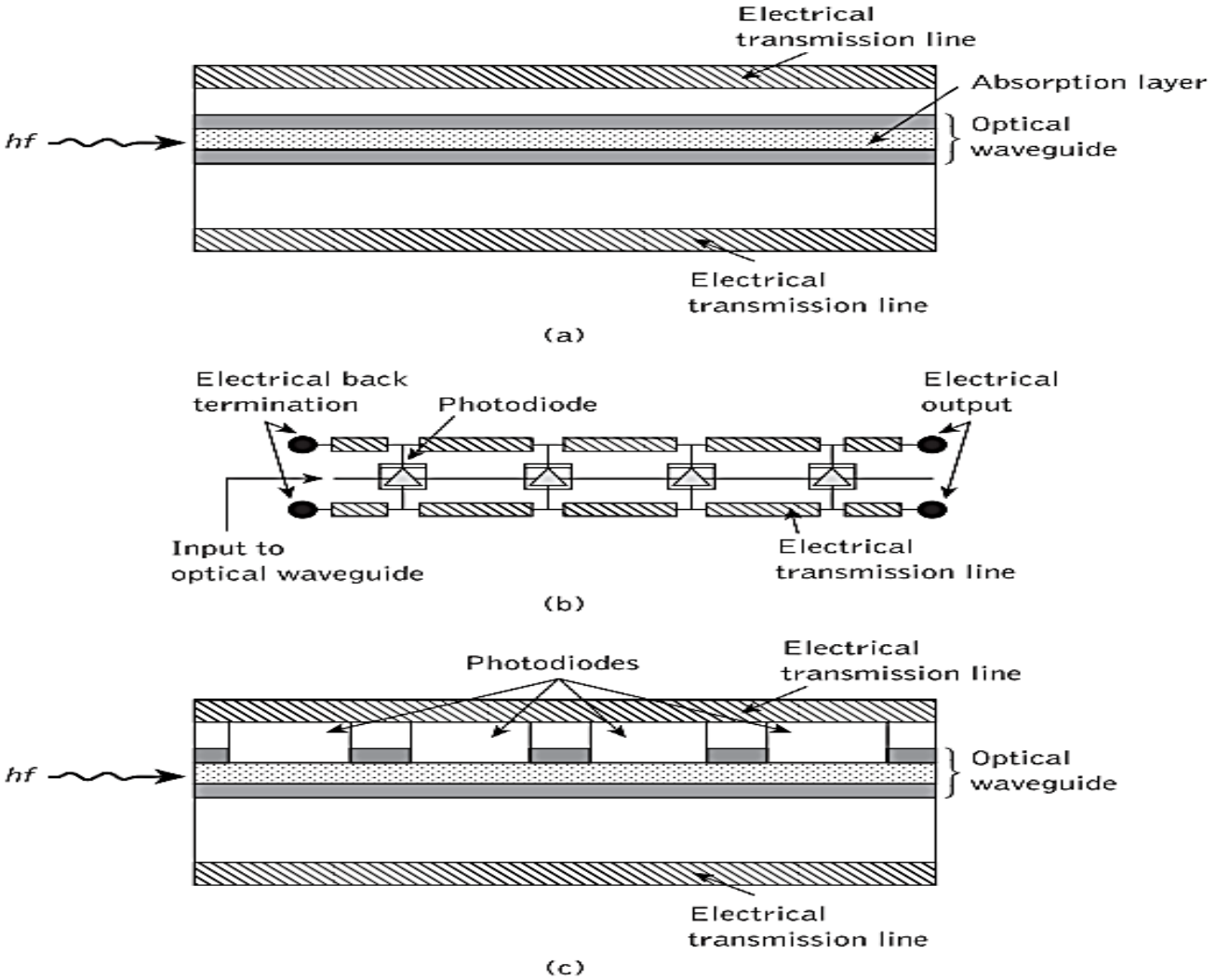


Figure 1.16 Traveling-wave photodiodes: (a) basic structure; (b) transmission line velocity matching scheme; (c) periodic traveling-wave photodiode.

- A further TW p–i–n photodiode device, known as unitraveling carrier (UTC) structure, is illustrated in Figure 1.17. Although the operation of a UTC photodiode is similar to a conventional p–i–n photodiode, absorption occurs in a thin p-type layer instead of the intrinsic i-region of the p–i–n photodiode. In the structure shown, based on the InGaAs/ InP material system, the photocarriers (i.e. electrons and holes pairs) are generated in the p-type absorption layer. However, when photons are absorbed to form the electron–hole pairs, the holes join these existing holes (instead of traveling) thus increasing the majority hole population. Hole carriers are replaced by electrons, which drift across the depletion region, and thus it is the electrons that generate the photocurrent. The bandwidth of the UTC photodiode is therefore determined by the electron diffusion time in the p-type absorption layer.

- When the absorption layer is thin the electrons can drift across it faster resulting in a higher bandwidth for the photodiode. Furthermore, using a small-gradient conduction band in the absorption layer can also speed up the diffusion time. Hence a photodiode structure with a 0.30 m thin absorption layer has provided a bandwidth as large as 310 GHz. Similar photodiodes have demonstrated high transmission rates of 100 Gbit and 160 Gbit with an output peak voltage of 0.8 V. Moreover, such wideband photodiodes have also been produced for optical wireless communication and a monolithic UTC photodiode operating at wavelength of 1.55 m was utilized for the purpose of generating a photonic CW signal transmitting over a bandwidth up to 1.5 THz.

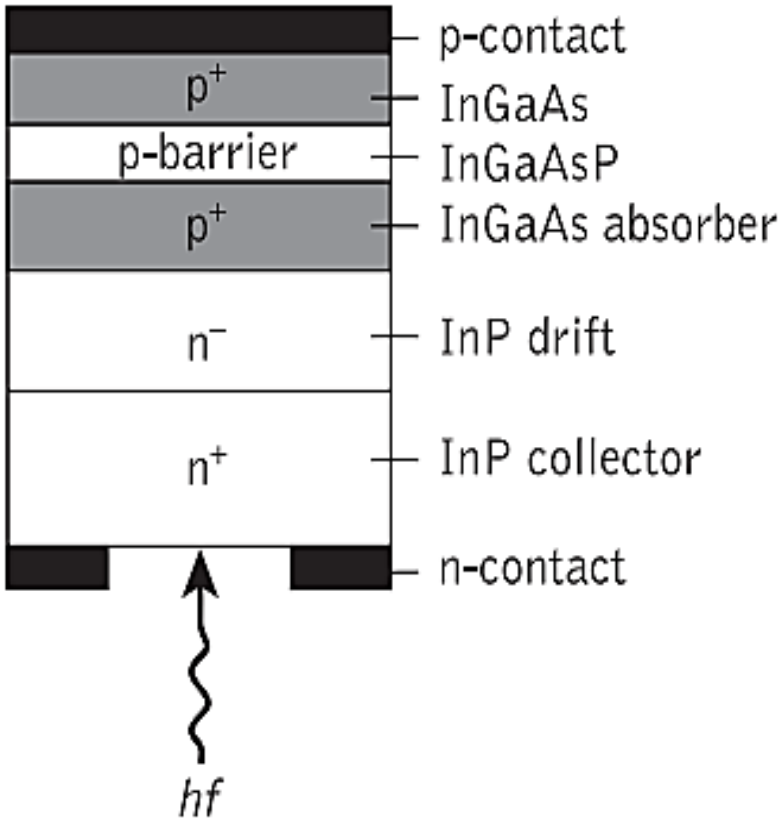


Figure 1.17: Unitraveling carrier (UTC) photodiode.

- In order to improve the absorption efficiency of high-speed photodiodes, an optical resonance cavity similar to the Fabry–Pérot cavity can also be employed. When resonance occurs in this structure the incoming photons are reflected back and forth between the two mirror face reflectors. If a thin absorption layer is placed within this mirror cavity the absorption efficiency is enhanced due to multiple passes of photons and the resultant device is known as a resonant cavity enhanced (RCE) photodiode. A simple RCE photodiode in which the absorption layer is placed in between two reflector mirrors of InGaAs n- and p-type material is shown in Figure 8.14.

- Various approaches can be used to produce the top and bottom distributed Bragg reflector (DBR) mirrors comprising several alternating layers of low-index and high-index material. For example, InGaAs/InP, AlGaAs/GaAs, InAlGaAs/InAlAs or InGaAsP/InAlAsmaterial systems can be employed to construct DBR mirrors operating at the wavelengths of 1.30 and 1.55 m. In addition, silicon-on-insulator technology can also be used to fabricate DBR mirrors using germanium on silicon substrates for an RCE photodiode to operate at long wavelength. For example, RCE p–i–n photodiodes using germanium material on a double silicon-on-insulator substrate operating at the wavelength of 1.55 m have exhibited high quantum efficiencies of 59%. These devices, which also provide a 3 dB bandwidth of around 13 GHz, are considered appropriate for reception at transmission rates of 10 Gbit. Furthermore, the small area of these devices (i.e. 10 to 70 m) could prove useful for photonic integration.

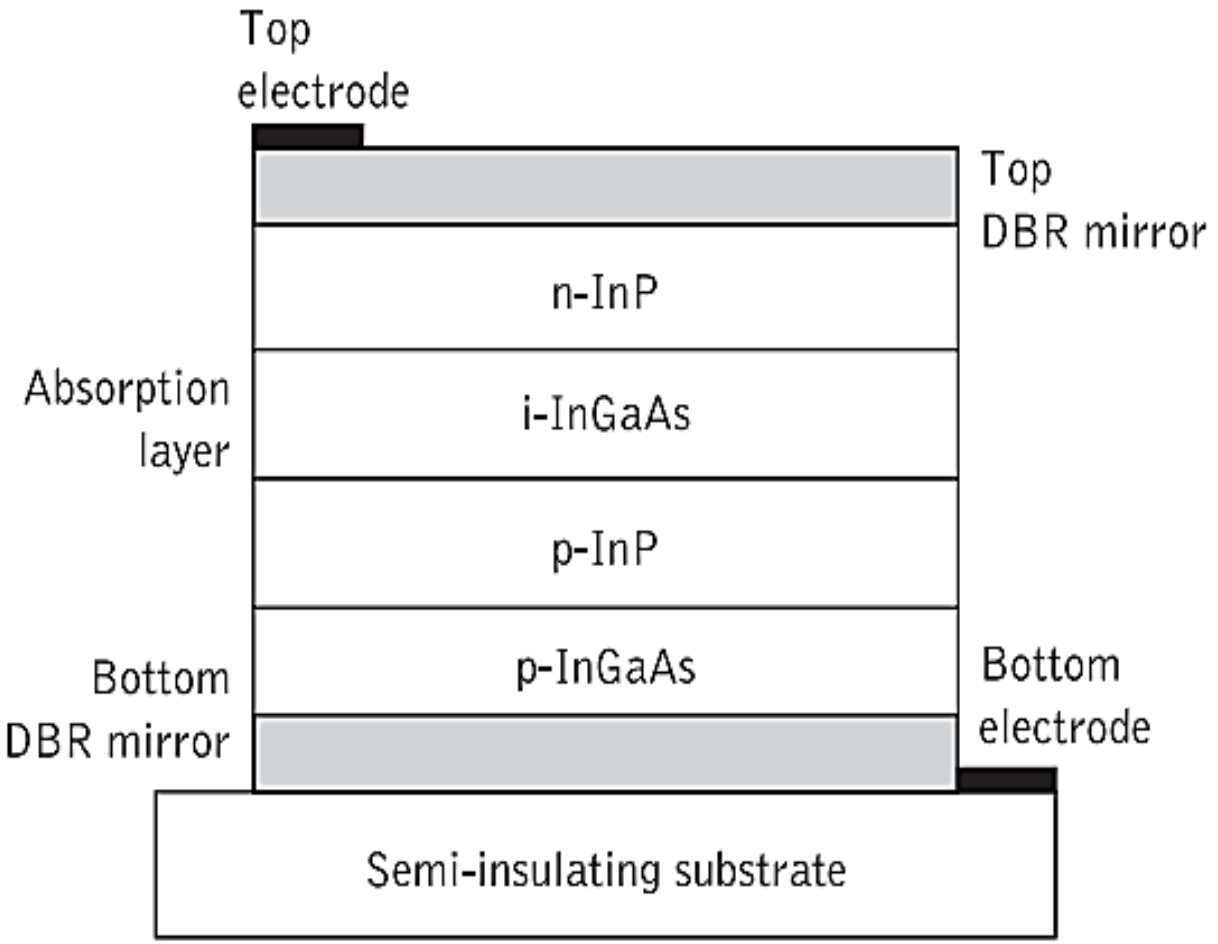


Figure 1.14 Schematic cross-section of a resonant cavity enhanced photodiode