Photonic Devices and Systems

Course Notes

Andrew Sarangan

Dept. of Electro-Optics and Photonics

University of Dayton

February 11, 2024

Contents

Chapter 1: A Basic Introduction to Optical Materials

Silica

The most widely used optical material is silica. The word “glass” and silica are often used interchangeably, although strictly speaking the term glass refers to a type of solid that is produced by rapid melt quenching (vitreous solid), and is not necessarily specific to silica. Silica is the oxide of silicon, SiO_2. Beach sand, as well as most of earth’s crust is mostly made up of silica. Optical glass is made by purifying and melting raw silica, and then rapidly cooling it to solidify it. The rapid cooling creates glass and prevents crystallization. The resulting material is known as fused silica. Fused silica has excellent optical transparency from the UV to the near infrared. However, fused silica is expensive due to its purity and high melting temperature (1700°C). Therefore, fused silica is only used in the most demanding applications that require high transparency and high temperature resistance.

The more widely used type of glass in optical components is BK7 – borosilicate glass. This is not pure SiO_2 as it contains boron oxide. BK7 has a very low thermal expansion coefficient so it is also used to make pyrex brand glassware. The melting temperature is 820°C, which is significantly lower than fused silica. As a result, it is also significantly less expensive. Soda-lime glass is another older form of glass which contains potassium oxide and calcium oxide, which are added to reduce the melting temperature down to 500°C. These are even cheaper than BK7, and are used in commercial windows, but the optical transmission is much lower, especially near the UV wavelengths.

Quartz is also SiO_2, but it is the crystalline form of SiO_2. It is anisotropic as well as piezoelectric. It is also far more expensive, so it is only used when ultra-high purity and crystallinity is required. The crystalline nature also make quartz more brittle than fused silica.

Other Common Dielectrics

CaF_2 (calcium fluoride) and MgF_2 (magnesium fluoride) are crystals used in some optical components due to their superior UV transmission properties. Al_2O_3 (aluminum oxide, or alumina) is also a transparent dielectric with a relatively high refractive index. The crystalline form of Al_2O_3 is sapphire. TiO_2 (titanium dioxide, or titania) is also used in optical coatings as a high index film. It’s crystalline form is known as rutile. The optical dispersion properties of a large number of materials can be found in public domain databases. One such database is . All of these materials are widely used in optical components and also in a thin film form for creating optical filters and anti-reflection coatings.

Semiconductors - Silicon

Silicon is by far the most commonly used semiconductor. Pure silicon is produced by purifying beach sand (SiO_2), and is the staple of the electronics industry. Unlike silica and quartz, the word silicon is used to describe the amorphous form as well as the crystalline form. Amorphous silicon is a black powdery material, similar to carbon. Crystalline silicon is grey, and is not transparent in the visible spectrum. However, it is very transparent in the infrared region for wavelengths longer than 1.1μm. Therefore, crystalline silicon is used for making refractive optical components in the infrared region where fused silica may not be transparent (for example in the mid-infrared 3-5μm spectrum). Since silicon is a semiconductor, absorbed light produces usable carriers that can be used in photon detection. Silicon photodetectors are used in virtually all consumer cameras.

Semiconductors – III-V compounds

Due to the bandstructure of silicon, it can be used to make photodetectors, but it cannot be used to make light emitters such as LED’s or laser diodes. III-V semiconductors are compound semiconductors, in contrast to elemental semiconductors such as Si and Ge. By combining a group III element with a group V element, III-V binary compounds can be created. GaAs, GaN, InP, InSb, InAs etc... are examples of III-V semiconductors. Most of these have direct bandgaps and can be used as photodetectors and also as photoemitters. Each material has a different refractive index, absorption and emission spectrum. GaAs is the most commonly used III-V semiconductor and is used to make LED’s and laser diodes. One can also combine two binary semiconductors to form ternaries like $Al_xGa_{1-x}As$. By adjusting the composition, their optical properties can be selected be fall anywhere between GaAs to AlAs. $Al_xGa_{1-x}As$ is used to make red laser diodes and LED’s. GaN is used in blue and UV laser diodes and LED’s. InP and $In_xGa_{1-x}As$ are used in infra-red sources and detectors. InSb and InAs are used in longer wavelength infrared applications.

Optical Spectrum

While the general electromagnetic spectrum has an infinite range, in the context of photonics what we consider as “light” has a few well-defined spectral windows. These windows are defined around the transmission properties of certain materials or the atmosphere.

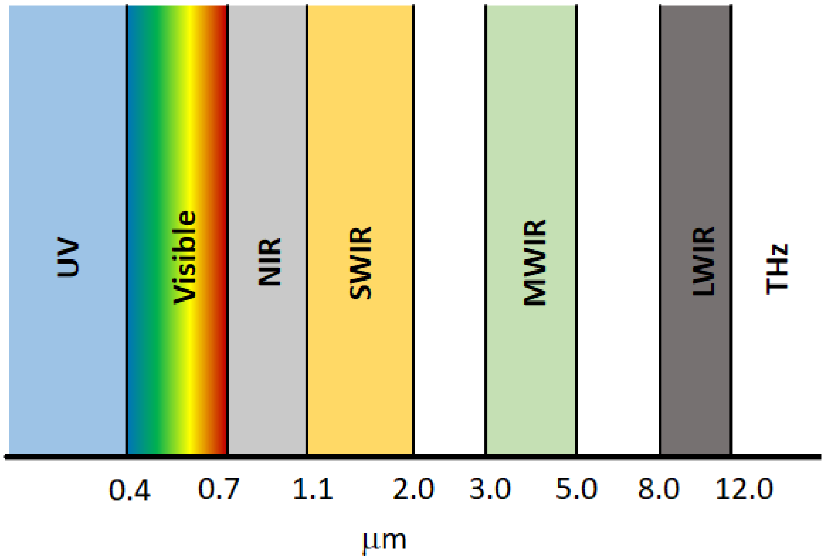

Figure 1: Commonly used optical spectral bands in photonics.

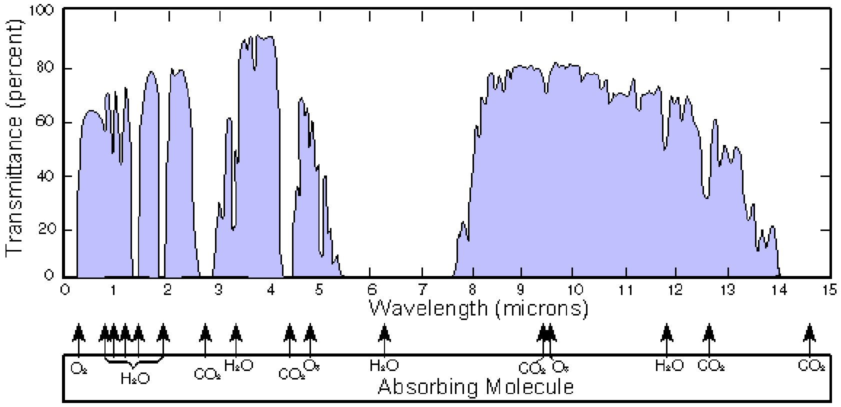

Figure 2: Atmospheric transmission spectrum. Source: Wikipedia

Homework 1

1. Show that the energy E of a photon (in eV) can be expressed as a function of its wavelength λ (in μm) as $E = \frac{1.24}{\lambda}$.

Qualitative Description of Electronic Bands in Semiconductors

Solitary atoms to crystals

Single atoms have discrete energy states due to the resonance of electron waves around the atom. This is a well-understood concept. A hydrogen atom only has one electron, but larger atoms have more. Even if an atom has a large number of electrons, its chemical and optical properties are mostly governed by the electrons in the highest energy states (i.e., the ones that are most loosely bonded to the nucleus). The lower energy electrons are tightly bounded to the nucleus, and rarely play a role except in high energy phenomena.

Silicon has 14 electrons, which are arranged as $1s^2 2s^2 2p^6 3s^2 3p^2$. These are discrete energy states of a single silicon atom. If we consider a collection of N silicon atoms that are isolated from each other, all of them will have the same discrete energy levels. In other words, we will have 14 states with N electrons in each state. Now, if the atoms are brought closer together such that they start to interact, the repulsion between electrons will produce a splitting of the energy levels. This will cause each of the 14 energy levels to split N-ways. As the atoms are brought closer and closer together, the energy levels will split further away from each other. This is how energy bands form in a solid. In other words, each discrete energy state of a solitary atom spreads out to create a band. If N is a large number, the separation between these split energy levels will be small. As a result, the band can be treated as a continuum rather than N discrete levels. The band that corresponds to the top most state ($3p^2$) is called the valence band. The band above that (corresponding to $4s^2$) is the conduction band. The difference between the highest point of the valence band and the lowest point of the conduction band is the energy bandgap.

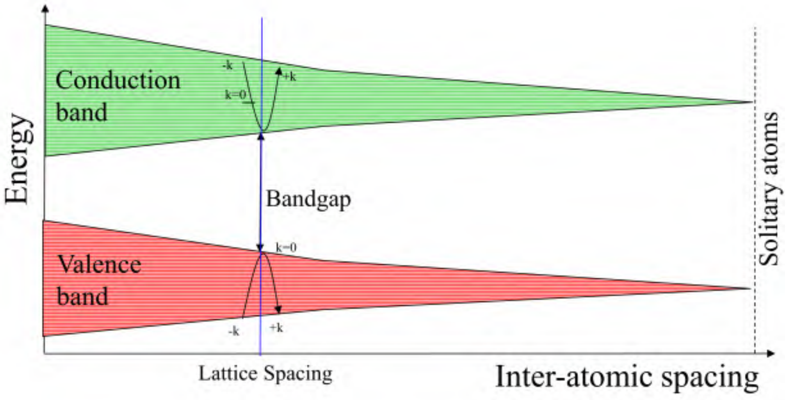

Figure 1: Evolution of the band structure due to inter-atomic interactions

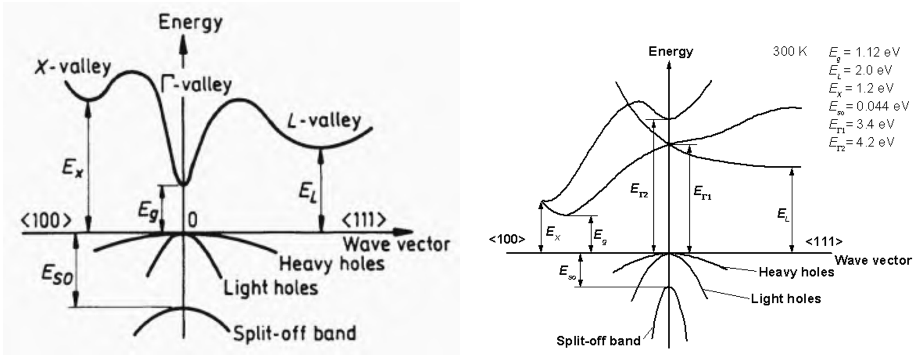

Figure 2: Band Structure of GaAs (left), Silicon (right)

Thermal excitation

Since 3p^2 was the highest state of the solitary silicon atom, one would expect the highest occupied band to be the valence band, and the conduction band to be empty. This is true only at 0K temperature. At this temperate, the material will be a perfect insulator. In order to make an electron move (hence increase its kinetic energy), its total energy has to be increased. This is only possible if the electron is able to accept some energy and move to a slightly higher energy state. If every single state in the valence band is occupied, then none of the electrons can change their energies, so they will not be able to respond to an external electric field.

At higher temperatures, some electrons in the valence band can acquire sufficient energy to become elevated into the conduction band. The bandgap of silicon is 1.1eV. At room temperature the average thermal energy is kT = 25meV. Even though 25meV is significantly smaller than 1.1eV, we should remember that this is the *average energy*; it does not mean every single electron will have 25meV. A fraction of electrons will have energy larger than 25meV, and a very small fraction will have energy greater than 1.1eV. Electrons with energy greater than 1.1eV will be capable of moving up into the conduction band. This creates vacancies in the valence band, so the electrons in the valence band can now respond to an external electric field. Additionally, electrons in the conduction band can also respond to the same electric field because they are also in a sea of empty states. This makes silicon a "semi" conductor at room temperature.

The conductivity of a semiconductor is directly related to the number of electrons in the conduction band and the number of vacancies (holes) in the valence band. As a result, the conductivity of silicon will increase with increasing temperature.

One would also expect that as band gap gets larger, the conductivity should decrease. Quartz has a band gap of 9eV, and consequently, behaves as an insulator at room temperature. At sufficiently high temperatures, even these crystals will start to conduct current, assuming the material does not melt or evaporate before reaching those temperatures.

At the other end of the spectrum, materials with narrow band gaps will have very high conductivities even at low temperature. InSb has a bandgap of 0.17eV. This is not too far from the average thermal energy of 25meV. As a result, a large number of electrons will be elevated into the conduction band. As a result, its conductivity at room temperature is very close to that of a metal. At sufficiently low temperatures, such as 77K, InSb behaves similar to silicon at room temperature.

Metals

Metals are different from semiconductors in the sense that they do not have a bandgap. They actually have a negative bandgap because the conduction and valence band overlap each other allowing the electrons to exist in both bands without having to traverse a gap. As a result, a metal will have mobile electrons in the conduction band even at 0K temperature.

Carrier Mobility

Conductivity depends not just on the concentration of electrons and holes, but also on their mobilities. The expression for conductivity is

$$\sigma = q\left(\mu_n n + \mu_p p\right). \tag{1}$$

Mobility is the ratio between the average drift velocity of the carrier and the electric field. Electrons and holes have different mobility values. They are also different from one material to another. Generally speaking, mobility decreases with increasing temperature. This is due to the increasing lattice vibrations which impedes the motion of electrons and holes.

The combined effect of temperature on carrier concentration and carrier mobility results in opposite trends in a metal and a semiconductor. In a metal, as temperature increases the carrier concentration is not affected because there is no band gap. However, mobility decreases. As a result, the conductivity of a metal will decrease as temperature is increased. In a semiconductor, as temperature increases the carrier concentration will increase dramatically. This more than offsets the decrease in mobility. As a result, the conductivity will increase with temperature. The conductivity of a semiconductor is more sensitive temperature than a metal. This effect can be used as a temperature sensor, such as in a thermistor. This is a semiconductor device whose conductivity is calibrated to allow small changes in temperature to be detected.

As an interesting side note, consider a filament light bulb. If you measure the resistance of the filament at room temperature, it will be significantly smaller than the value required to produce the rated output of the bulb. For example, a 25W, 120V light bulb should have a resistance of 576Ω, but if you measure it with an ohmmeter it will probably read something like 100Ω. This means, when the light bulb is turned on, the instantaneous power dissipation will be 144W, falling to 25W as the filament reaches its operating temperature. This should explain why filament light bulbs fail mostly when they are turned on rather than during a steady operation.

Thermistors

Shown here is a typical resistance vs temperature curve of a commercial NTC (Negative temperature coefficient) thermistor. Thermistors have a greater sensitivity compared to thermocouples, but they have a smaller range.

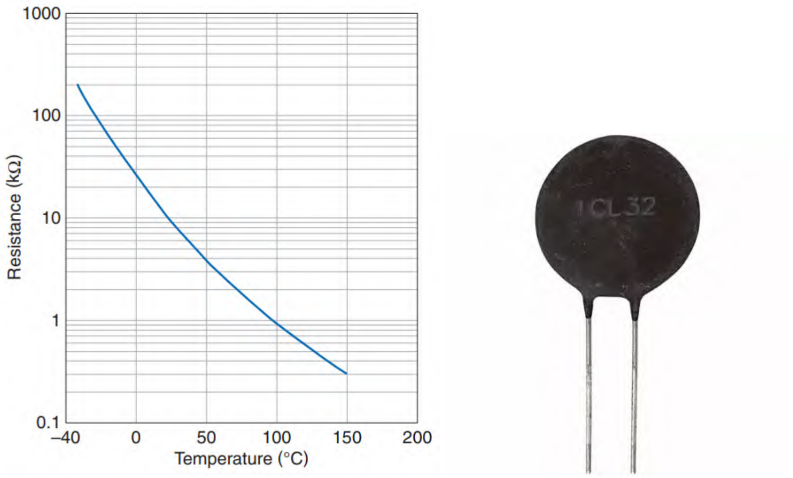

Figure 3: Resistance vs temperature of a commercial thermister. Source: TDK.

Photoconductors

In the case of thermistors, thermal energy elevates an electron from the valence band to the conduction band. Photoconductors function almost identically to thermistors, except they respond to light instead of temperature. They are also known as photoresistors, or simply as photocells. Photons with energy greater than the bandgap will elevate electrons from the valence band to the conduction band, resulting in an increased conductance (or reduced resistance).

CdS is a popular semiconductor used in photocells. Its bandgap of 2.42eV corresponds to a wavelength of 0.87μm. As a result, it is able to absorb the entire visible spectrum. Silicon, whose bandgap is 1.1eV can sense all wavelengths up to 1.1μm, which includes the visible wavelength range as well as the near infrared.

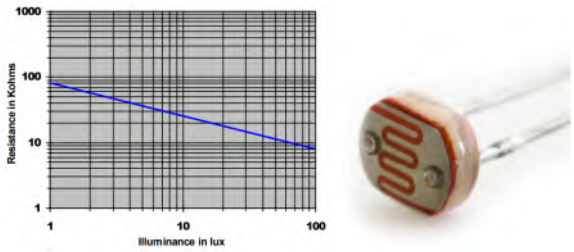

Figure 4: Resistance vs illumination of a commercial CdS photocell. Source: Luna Optoelectronics.

Photoconductors made from narrow bandgap semiconductors can respond to longer wavelengths, however, they may have to be cooled to a low enough temperature to ensure they are behave as semiconductors and not like metals. Silicon does not have to be cooled because the average thermal energy at room temperature is 25meV which is about 44 times smaller than its bandgap energy of 1.12eV.

Electron and Hole Densities in Semiconductors

Electrons in a Solid vs Other Particles

Fermi level

Fermi level is best understood in the context of metals. At T=0K, the highest electron energy is the Fermi energy. Think of a column of water in a vertical tube. The height of the water/air interface can be thought of as the Fermi level. Now, as temperature is increased, some of the electrons at the highest level will acquire enough energy to jump to an even higher level. This will leave a vacancy at the site it jumped from, and an occupancy at the site it is jumping to. As a result, the electron concentration will slightly decline below the Fermi level and slightly increase above the Fermi level. In the analogy of a water column, one can think of this as similar to the water vapor molecules jumping up from the liquid into air. The end result is, the electrons will have a smeared energy distribution rather than a sharp step-like distribution at the Fermi level.

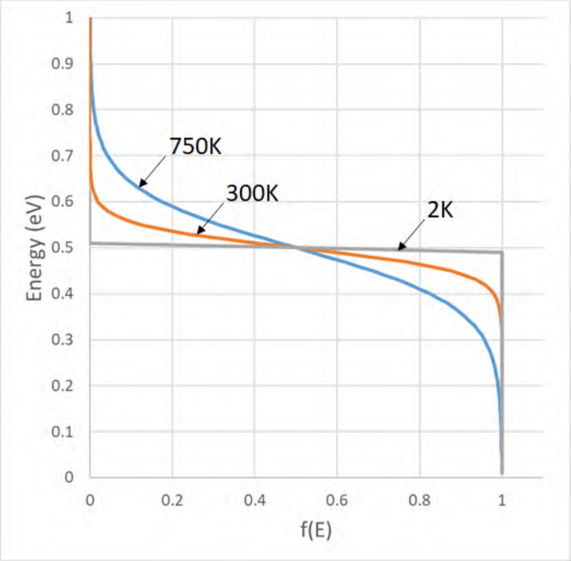

Figure 1: Fermi Distribution Function assuming $E_F = 0.5$eV.

Therefore, Fermi level lies halfway between the fully-occupied states and the fully-unoccupied states. Strictly speaking, Fermi level is the energy level where the occupation probability is 50%. This distribution can be mathematically written as:

$$f\left(E\right) = \frac{1}{e^{\left(\left(E-E_F\right)/kT\right)}+1}. \tag{1}$$

We should be able to verify that $f\left(0\right) \rightarrow 1$, $f\left(\infty\right) \rightarrow 0$ and $f\left(E_F\right) = 0.5$. We can also verify that this smearing around the Fermi level becomes more pronounced at higher temperatures.

Density of states

The Fermi function only describes the probability of occupation of electrons. If there are no allowed sites, they will not be occupied regardless of the probability of occupation. The mathematical distribution of electronic sites is the density of states (DOS) function. This is given in the units of number of electronic states per unit volume per unit energy.

$$\rho_c(E) = \frac{\sqrt{2}m_c^{3/2}}{\pi^2\hbar^3}(E-E_c)^{1/2} \text{ for } E > E_c \quad \text{(conduction band)} \tag{2}$$

$$\rho_v(E) = \frac{\sqrt{2}m_v^{3/2}}{\pi^2\hbar^3}(E_v-E)^{1/2} \text{ for } E < E_v \quad \text{(valence band)} \tag{3}$$

The density of states is zero at the band edges (top of the valence band and bottom of the conduction band), but increases as a square root of energy from there. The densities are roughly the same moving up into the conduction band and moving down into the valance band, except for the differences in the electron effective masses.

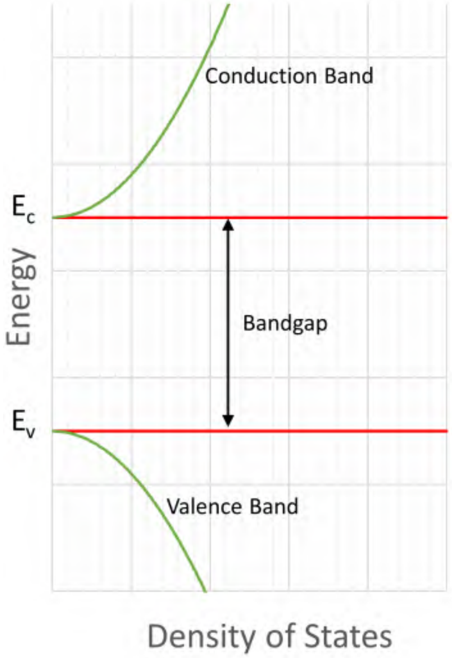

Figure 2: Density of States in the conduction and valence bands

Electron Density

The electron density in the conduction band can be evaluated by multiplying the density of available states by the Fermi function, which is the probability of occupation. This can be written as:

$$n(E) = \rho_c(E)\, f(E)\,. \tag{4}$$

This is the number of electrons per unit volume per unit energy. The total number of electrons per unit volume can be obtained by integrating over all energies in the conduction band. Since conduction band starts at $E = E_c$, this becomes:

$$n = \int_{E_c}^{\infty} n(E)\,dE = \int_{E_c}^{\infty} \rho_c(E)\, f(E)\,dE \tag{5}$$

$$= \int_{E_c}^{\infty} \left(\frac{\sqrt{2}m_c^{3/2}}{\pi^2\hbar^3}(E-E_c)^{1/2}\right)\left(\frac{1}{e^{((E-E_F)/kT}+1}\right)dE. \tag{6}$$

For energy levels much higher than the Fermi level ($E >> E_F$), we can simplify the Fermi function by dropping the 1 in the denominator. Because the integration limit begins at E_c, this is a valid assumption as long as E_F is significantly lower than E_c (i.e., near the middle of the bandgap). This assumption allows us to analytically integrate the function. The result is:

$$n = \frac{4\sqrt{2}}{h^3} \left(\pi m_c kT\right)^{3/2} e^{(E_F - E_c)/kT}. \tag{7}$$

The behavior of $n\left(E\right)$, $\rho_c\left(E\right)$ and $f\left(E\right)$ are shown in Fig 3.

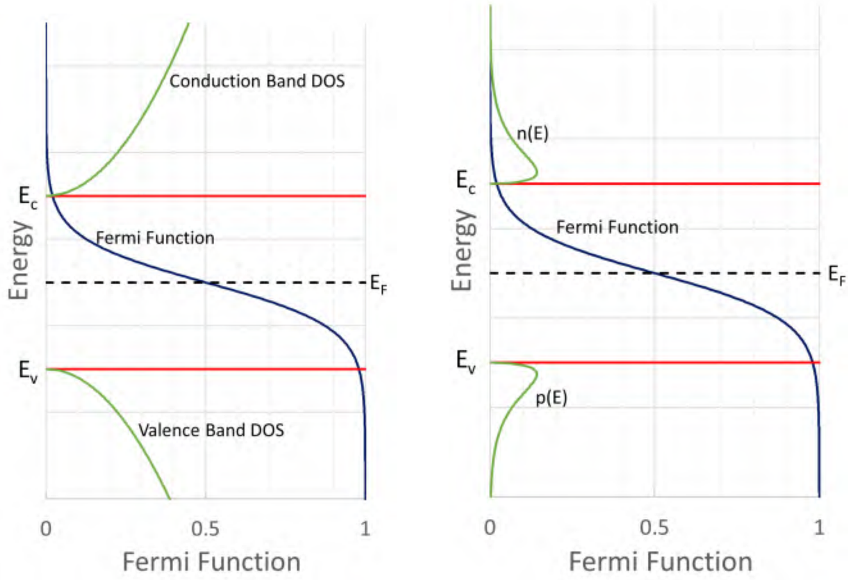

Figure 3: (Left) Density of States and Fermi distribution; (Right) Carrier Density and Fermi distribution

Hole Density

Similarly, the hole density in the valance band can be obtained by multiplying the density of state by the probability of having a vacancy (which is $1 - f\left(E\right)$):

$$\begin{aligned} p &= \int_{-\infty}^{E_v} p\left(E\right) dE = \int_{-\infty}^{E_v} \rho_v\left(E\right)\left(1 - f\left(E\right)\right) dE & (8) \\ &= \int_{-\infty}^{E_v} \left(\frac{\sqrt{2} m_v^{3/2}}{\pi^2 \hbar^3} \left(E_v - E\right)^{1/2}\right) \left(1 - \frac{1}{e^{((E-E_F)/kT)} + 1}\right) dE & (9) \\ &= \int_{-\infty}^{E_v} \left(\frac{\sqrt{2} m_v^{3/2}}{\pi^2 \hbar^3} \left(E_v - E\right)^{1/2}\right) \left(\frac{e^{((E-E_F)/kT)}}{e^{((E-E_F)/kT)} + 1}\right) dE. & (10) \end{aligned}$$

Similar to before, if the Fermi level is significantly above the upper integration limit of E_v, we can drop the bottom exponent, resulting in:

$$p = \frac{4\sqrt{2}}{h^3} \left(\pi m_v kT\right)^{3/2} e^{(E_v - E_F)/kT}. \tag{11}$$

Intrinsic Semiconductors

In an intrinsic (undoped) semiconductor, the number of electrons in the conduction band should be equal to the number of vacancies in the valence band. We will represent this as n_i (for intrinsic carrier concentration). In other words, $n = p = n_i$. We can set equation (7) and (11) to each other, which will allow us to move all the known variables to the right and the only unknown (E_F) to the left. E_F in this case works out to be nearly halfway inside the bandgap. It is nearly halfway and not exactly halfway because of the differences in the effective masses. If we work out the expression, it becomes:

$$E_F = E_i = E_v + \frac{E_c - E_v}{2} + \frac{3}{4} kT \ln \left(\frac{m_v}{m_c} \right). \tag{12}$$

The interesting thing about this result is that the Fermi level falls at a place where the density of states is zero, i.e., within the bandgap. However, that should not be a cause for alarm because Fermi level is just a parameter in the probability distribution function. It is not necessarily an energy level that an electron can occupy. Using this expression for E_F, the value of n_i can be evaluated. This expression results in

$$n_i = \frac{4\sqrt{2}}{h^3} \left(\pi kT\right)^{3/2} \left(m_c m_v\right)^{3/4} e^{-E_g/2kT} \tag{13}$$

where E_g is the bandgap $E_c - Ev$. Alternatively, we could have also taken a product $np = n_i^2$ by multiplying equations (7) and (11) together and then taking a square root. Both approaches will give the same result. For silicon at room temperature the value for n_i works out to approximately 1×10^{10}cm^{-3}. For GaAs, this value works out to be 1×10^{6}cm^{-3}. These are the average concentration of electrons and holes in these materials at room temperature.

Doped Semiconductors

Even when grown as an ultra-pure crystal, semiconductors will have trace amounts of unintentional impurities that significantly affect the electron and hole concentrations. One can also intentionally add impurities to raise the electron or hole concentrations. This process is known as doping. When doping an intrinsic semiconductor, it is not feasible to control the intentional impurity concentration to levels smaller than about 1 part per billion (ppb). The atomic density of silicon is about 5×10^{22}/cm^{-3}. As a result, an impurity concentration of 10^{14}cm^{-3}, despite its large magnitude, is still considered an extremely doping value. Doping levels smaller than that are not easy to achieve.

Impurities are classified as either donors or acceptors. When donor impurities are added to a semiconductor, the number of electrons in the conduction band will increase and the number of holes in the valence band will decrease. This is known as n-type semiconductors. The donor atoms are chosen such that they introduce occupied energy levels just below the conduction band edge, as shown in Fig 4. For silicon, these are typically from group V, such as phosphorous or arsenic. Because their energy levels are so close to the conduction band, it takes very little energy to excite those electrons into the conduction band. At room temperature, nearly all of those donor atoms would release their electrons into the conduction band.

Just like with intrinsic semiconductors where we set $n = p = n_i$, in this case we will have

$$n = p + N_D^+ \tag{14}$$

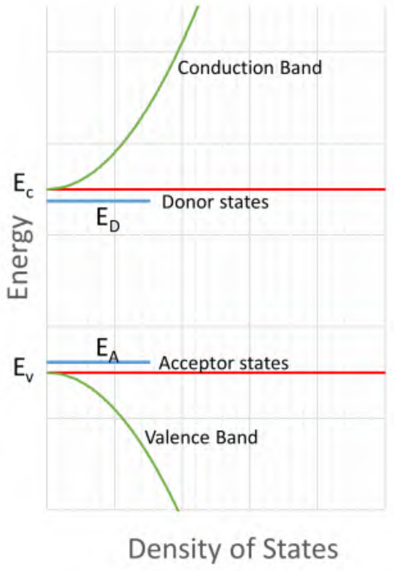

Figure 4: Donor and acceptor states

where N_D^+ is the number of dopants that are ionized (i.e., those that have released electrons into the conduction band). The fraction of ionized dopants (N_D^+) compared to the total dopants (N_D) will be governed by the Fermi function:

$$N_D^+ \left(E\right) = N_D \left(E\right) \left(1 - f \left(E\right)\right). \tag{15}$$

Using equations (6) and (10) from before, we can write this as:

$$\int_{E_c}^{\infty} \rho_c \left(E\right) f \left(E\right) dE = \int_{-\infty}^{E_v} \rho_v \left(E\right) \left(1 - f \left(E\right)\right) dE + \int_{-\infty}^{+\infty} N_D \left(E\right) \left(1 - f \left(E\right)\right). \tag{16}$$

Unlike with the conduction and valence bands where the density of states were a continuum above and below the band edges, the donor state is at a discrete energy level E_D, such as:

$$N_D \left(E\right) = N_D \delta \left(E - E_D\right). \tag{17}$$

When integrated, this becomes:

$$\int_{-\infty}^{+\infty} N_D \left(E\right) f \left(E\right) = \int_{-\infty}^{+\infty} N_D \delta \left(E - E_D\right) \left(1 - f \left(E\right)\right) = N_D \left(1 - f \left(E_D\right)\right). \tag{18}$$

Therefore, equation (16) becomes:

$$\begin{aligned} &\int_{E_c}^{\infty} \left(\frac{\sqrt{2} m_c^{3/2}}{\pi^2 \hbar^3} \left(E - E_c\right)^{1/2} \right) \left(\frac{1}{e^{((E-E_F)/kT} + 1} \right) dE = \\ &\int_{-\infty}^{E_v} \left(\frac{\sqrt{2} m_v^{3/2}}{\pi^2 \hbar^3} \left(E_v - E\right)^{1/2} \right) \left(1 - \frac{1}{e^{((E-E_F)/kT)} + 1} \right) dE \\ &\qquad + N_D \left(1 - \left(\frac{1}{e^{((E_D-E_F)/kT)} + 1} \right) \right). \end{aligned} \tag{19}$$

Just like with the earlier case, assuming E_F is sufficiently far from the band edges as well as from E_D, we can make some simplifications and rewrite equation (19) as

$$\frac{4\sqrt{2}}{h^3}\left(\pi m_c kT\right)^{3/2} e^{(E_F-E_c)/kT} = \frac{4\sqrt{2}}{h^3}\left(\pi m_v kT\right)^{3/2} e^{(E_v-E_F)/kT} + N_D, \tag{20}$$

from which we can solve for E_F. Once E_F is found, it is straightforward to calculate n and p. For example, in the case of silicon, assuming $N_D = 1 \times 10^{15}$cm^{-3}, we can calculate $E_F - E_v = 0.898$ eV. Since the band gap is 1.12eV, we can verify that this level is higher than the mid-level of the intrinsic (undoped) case. We can also verify that $\left(E_c - E_F\right)/kT = 8.6$ and $\left(E_F - E_v\right)/kT = 34.7$ which are sufficiently large to justify the approximation made in the Fermi function. The resulting carrier densities are $n = 1 \times 10^{15}$cm^{-3} and $p = 1.52 \times 10^4$cm^{-3}. In other words, the electron concentration is nearly the same as the donor density.

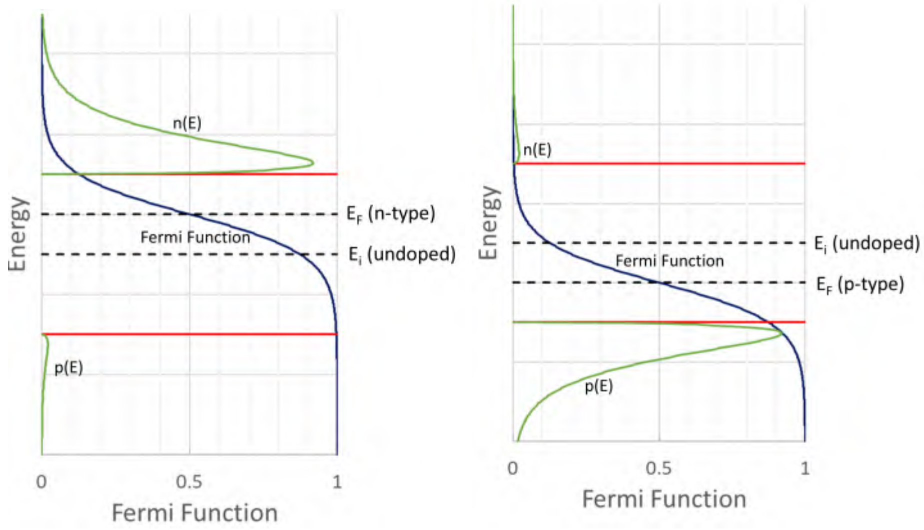

Figure 5: Density of States and Fermi distribution for n-type (left), and p-type (right) semiconductor.

Similarly, p-type semiconductors are created by adding acceptor atoms. These are atoms with a vacant energy level close to the valence band. In silicon, these are typically from group III, such as boron or aluminum. This will result in nearly all of those vacant acceptor states being filled by electrons in the valence bands, resulting in the same number of vacancies in the valence band. If N_A is the acceptor concentration, there will be nearly N_A number of vacancies (holes) in the valence band.

The equations corresponding to (14) and (19) are:

$$p = n + N_A^-, \tag{21}$$

and

$$\int_{-\infty}^{E_v}\left(\frac{\sqrt{2}m_v^{3/2}}{\pi^2\hbar^3}\left(E_v - E\right)^{1/2}\right)\left(1 - \frac{1}{e^{((E-E_F)/kT)}+1}\right)dE =$$

$$\int_{E_c}^{\infty}\left(\frac{\sqrt{2}m_c^{3/2}}{\pi^2\hbar^3}\left(E - E_c\right)^{1/2}\right)\left(\frac{1}{e^{((E-E_F)/kT}+1}\right)dE + N_A\left(\frac{1}{e^{((E_A-E_F)/kT)}+1}\right) \tag{22}$$

which, assuming E_F is sufficiently far from the band edges and E_A, can be simplified to

$$\frac{4\sqrt{2}}{h^3}\left(\pi m_v kT\right)^{3/2} e^{(E_v-E_F)/kT} = \frac{4\sqrt{2}}{h^3}\left(\pi m_c kT\right)^{3/2} e^{(E_F-E_c)/kT} + N_A. \tag{23}$$

For example, in silicon, assuming $N_A = 1\times10^{15}$cm^{-3} and $E_A - E_v = 0.044$eV (for boron doping), we can calculate $E_F - E_v = 0.254$ eV, which is lower than the intrinsic level. The carrier densities are $p = 1\times10^{15}$cm^{-3} and $p = 1.52\times10^4$cm^{-3}. Just like with the donor doping, the hole concentration in this case is almost equal to the acceptor density. The values for $(E_c - E_F)/kT$ $(E_F - E_v)/kT$ are 33.5 and 9.8, respectively, which satisfy the requirement for the approximation made in the Fermi function.

Since we define the intrinsic Fermi level as E_i, substituting equation (12) into (7) (and equation (12) into (11)) we can get an approximate relationship between the Fermi level and the carrier concentration in reference to the intrinsic condition. These expressions are applicable to doped semiconductors or any semiconductors under carrier injection:

$$n = n_i e^{(E_F-E_i)/kT} \tag{24}$$
$$p = n_i e^{(E_i-E_F)/kT}. \tag{25}$$

We can also reverse equations (24) and (25) and express the Fermi level in reference to the carrrier concentrations:

$$E_F = E_i + kT\ln\left(\frac{n}{n_i}\right) \tag{26}$$
$$E_F = E_i - kT\ln\left(\frac{p}{n_i}\right). \tag{27}$$

From the examples and calculations discussed above, we can see that E_F has to increase in an n-type semiconductor. For lightly doped n-type semiconductors, the Fermi level will be slightly above the midpoint of the bandgap. For heavily doped n-type semiconductors, the Fermi level can be higher. Similarly, for p-type semiconductors, the Fermi level will be below the intrinsic level. However, we need to be cautious about the approximation we made for the Fermi function; i.e., that the Fermi level E_F is far from the band edges and the dopant energy levels. This approximation may break down at high levels of doping concentrations, and we may have to revert back to the original Fermi functions without that approximation (equations (19) and (22)).

Mass Action Effect

As stated earlier, equation (13) can also be obtained by taking the product of n and p from equations (7) and (11) and then taking the square root. This product np also exhibits an interesting aspect: it is independent of E_F. This is an important observation because what it means is that np is a constant regardless of whether the material is doped or not. This is known as the mass action law. Therefore, if we know the value of n, we can evaluate the value of p, and vice versa. If a semiconductor is doped with a donor concentration of N_D, the electron concentration will be $n \approx N_D$, and the hole concentration will be $p = n_i^2/n$. If GaAs is doped with a donor concentration of 1×10^{15}cm^{-3}, we can get an electron concentration of $n \approx 1\times10^{15}$cm^{-3} and $p = n_i^2/n = \left(2.1\times10^6\right)^2/\left(1\times10^{15}\right) = 4\times10^{-3}$cm^{-3}. The dominant carrier type is known as majority carriers, and the other is known as minority carriers. In this example, electrons are the majority carriers, and holes are the minority carriers.

Conductivity and Resistivity

In a semiconductor, both electrons and holes contribute to current flow. The conductivity can be written as:

$$\sigma = q\left(\mu_n n + \mu_p p\right), \tag{28}$$

where μ_n is the mobility of electrons in the conduction band, and μ_p is the mobility of holes in the valence band. For Si, $\mu_n = 1400$ cm^{2}V·s and $\mu_p = 450$ cm^{2}V·s. From this, we can get the conductivity of intrinsic silicon as 4.4×10^{-6} S/cm. Alternatively, we can also express it as resistivity, which is $\rho = 1/\sigma = 225$ kΩ·cm. Similarly, the resistivity of intrinsic GaAs is 334 MΩ·cm. As stated earlier, these resistivities are extremely unlikely scenarios because it is not easy to produce silicon or GaAs with impurity concentrations to that level of purity. Hence, commonly produced silicon substrates, even when undoped, have resistivities on the order $\rho \approx 1$ kΩ·cm due to unintentional impurities.

When doped, these semiconductors will have much smaller resistivity values. For example, GaAs doped with a donor concentration of 1×10^{15}cm^{-3} will have a resistivity of 4.4 Ω·cm.

Homework 2

1. Calculate the intrinsic carrier concentration of silicon at room temperature using equation (13). Look up the relevant material parameters from `http://www.matprop.ru`.

   Run this code.

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val m0 = 9.1e-31
    val q = 1.602e-19
    val h = 6.62607004e-34
    val k = 1.38064852e-23
    val mc = 0.36*m0
    val mv = 0.81*m0
    val Eg = 1.12*q
    val T = 300.0
    val ni = (4.0*2.0.pow(0.5)/h.pow(3))*(PI*k*T).pow(1.5)*(mc*mv).pow(0.75)*exp(-Eg
    /(2.0*k*T))
    println("%.2e".format(ni/1.0e6))
}

>>3.90e+09
```

2. Using equation (20), solve for E_F when the the donor doping concentration is 1×10^{17} cm^{-3} in silicon. Then find the corresponding hole concentration. Verify the validity of the approximation made in simplifying the Fermi function.

   Run this code.

```
import kotlin.math.*
//Andrew Sarangan

fun NewtonRaphson(f: (input:Double) -> Double, initialX:Double):Double{
    val dx = initialX*1.0e-10
    var x1 = initialX
    var fprime:Double
    var x2:Double
    var diff:Double
    do {
      fprime = (f(x1) - f(x1-dx))/dx
      x2 = x1 - f(x1)/fprime
        diff = abs((x2-x1)/x1)
        x1 = x2
    } while(diff > 1.0e-12)
    return x1
}

fun main() {
    val m0 = 9.1e-31
    val q = 1.602e-19
    val h = 6.62607004e-34
    val k = 1.38064852e-23
    val mc = 0.36*m0
    val mv = 0.81*m0
    val Eg = 1.12*q
    val T = 300.0
    val Ev = 0.0
    val Ec = Ev+Eg
    val Nd = 1.0e17*1.0e6
    fun n(Ef:Double) = 4.0*2.0.pow(0.5)/h.pow(3)*(PI*mc*k*T).pow(1.5)*exp((Ef-Ec)/(k
    *T))
    fun p(Ef:Double) = 4.0*2.0.pow(0.5)/h.pow(3)*(PI*mv*k*T).pow(1.5)*exp((Ev-Ef)/(k
    *T))
    fun fn(Ef:Double) = n(Ef) - p(Ef) - Nd

    val Ef = NewtonRaphson((::fn), 0.9*q)

    println(”Fermi Level (Ef-Ev) = ${”%.3f”.format(Ef/q)} eV”)
    println(”n = ${”%.2e”.format(n(Ef)*1.0e-6)} /cm3”)
    println(”p = ${”%.2e”.format(p(Ef)*1.0e-6)} /cm3”)
    println(”(Ec-Ef)/kT = ${”%.1f”.format((Ec-Ef)/(k*T))}”)
    println(”(Ef-Ev)/kT = ${”%.1f”.format((Ef-Ev)/(k*T))}”)
}

>>Fermi Level (Ef-Ev) = 1.017 eV
>>n = 1.00e+17 /cm3
>>p = 1.52e+02 /cm3
>>(Ec-Ef)/kT = 4.0
>>(Ef-Ev)/kT = 39.3
```

The value of e^4 is 54.4, so the approximation of dropping the 1 in the Fermi expression is still valid.

3. Consider a donor doping density of 10^{19}cm^{-3} in silicon. Using equations (20) and (23) calculate E_F, n and p. Verify the validity of the approximations made in the Fermi function. If necessary, repeat the calculations with the full integrals without the approximation (equations (19) and (22)).

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun NewtonRaphson(f: (input:Double) -> Double, initialX:Double):Double{
    val dx = initialX*1.0e-10
    var x1 = initialX
    var fprime:Double
    var x2:Double
    var diff:Double
    do {
      fprime = (f(x1) - f(x1-dx))/dx
      x2 = x1 - f(x1)/fprime
        diff = abs((x2-x1)/x1)
        x1 = x2
    } while(diff > 1.0e-12)
    return x1
}

fun main() {
    val m0 = 9.1e-31
    val q = 1.602e-19
    val h = 6.62607004e-34
    val k = 1.38064852e-23
    val mc = 0.36*m0
    val mv = 0.81*m0
    val Eg = 1.12*q
    val T = 300.0
    val Ev = 0.0
    val Ec = Ev+Eg
    val Nd = 1.0e19*1.0e6

    fun n(Ef:Double) = 4.0*2.0.pow(0.5)/h.pow(3)*(PI*mc*k*T).pow(1.5)*exp((Ef-Ec)/(k
    *T))
    fun p(Ef:Double) = 4.0*2.0.pow(0.5)/h.pow(3)*(PI*mv*k*T).pow(1.5)*exp((Ev-Ef)/(k*T))
    fun fn(Ef:Double) = n(Ef) - p(Ef) - Nd

    val Ef = NewtonRaphson((::fn), 1.1*q)

    println("Fermi Level (Ef-Ev) = ${"%.3f".format(Ef/q)} eV")
    println("n = ${"%.2e".format(n(Ef)*1.0e-6)} /cm3")
    println("p = ${"%.2e".format(p(Ef)*1.0e-6)} /cm3")
    println("(Ec-Ef)/kT = ${"%.1f".format((Ec-Ef)/(k*T))}")
    println("(Ef-Ev)/kT = ${"%.1f".format((Ef-Ev)/(k*T))}")
}

>>Fermi Level (Ef-Ev) = 1.136 eV
>>n = 1.00e+19 /cm3
>>p = 1.52e+00 /cm3
>>(Ec-Ef)/kT = -0.6
>>(Ef-Ev)/kT = 43.9
```

The value of $e^{-0.6}$ is 0.54, so dropping the 1 in the Fermi expression is not a valid approximation. We need to repeat the calculation with the full integrals.

Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun NewtonRaphson(f: (input:Double) -> Double, initialX:Double):Double{
    val dx = initialX*1.0e-10
    var x1 = initialX
    var fprime:Double
    var x2:Double
    var diff:Double
    do {
      fprime = (f(x1) - f(x1-dx))/dx
      x2 = x1 - f(x1)/fprime
        diff = abs((x2-x1)/x1)
        x1 = x2
    } while (diff > 1.0e-12)
    return x1
}

fun main() {
    val m0 = 9.1e-31
    val q = 1.602e-19
    val h = 6.62607004e-34
    val hbar = h/(2.0*PI)
    val k = 1.38064852e-23
    val mc = 0.36*m0
    val mv = 0.81*m0
    val Eg = 1.12*q
    val T = 300.0
    val Ev = 0.0
    val Ec = Ev+Eg
    val Nd = 1.0e19*1.0e6
    val Ed = Ec - 0.046*q

    fun n(Ef:Double):Double{
        val dE = 1.0e-5*q
        val EcUpper = 1.0*q
        return DoubleArray((EcUpper/dE).toInt()){Ec+it*dE}.map{
                2.0.pow(0.5)*mc.pow(1.5)/(PI.pow(2)*hbar.pow(3))*(it-Ec).pow(0.5)/(exp((it-Ef)/(k*T))+1.0)*dE}.sum()
    }
    fun p(Ef:Double):Double{
        val dE = 1.0e-5*q
        val EvLower = 1.0*q
        return DoubleArray((EvLower/dE).toInt()){Ev-it*dE}.map{
                   2.0.pow(0.5)*mv.pow(1.5)/(PI.pow(2)*hbar.pow(3))*(Ev-it).pow(0.5)
    *(exp((it-Ef)/(k*T))/(exp((it-Ef)/(k*T))+1.0))*dE}.sum()
    }
    fun NdPlus(Ef:Double) = Nd*(1.0 - 1.0/(exp((Ed-Ef)/(k*T)) + 1.0))
    fun fn(Ef:Double) = n(Ef) - p(Ef) - NdPlus(Ef)

    val Ef = NewtonRaphson((::fn), 1.0*q)
    println("Fermi Level (Ef-Ev) = ${"%.3f".format(Ef/q)} eV")
    println("n = ${"%.2e".format(n(Ef)*1.0e-6)} /cm3")
    println("p = ${"%.2e".format(p(Ef)*1.0e-6)} /cm3")
}

>>Fermi Level (Ef-Ev) = 1.103 eV
>>n = 2.43e+18 /cm3
>>p = 5.34e+00 /cm3
```

Using the full integral, we can find that only 25% of the donors are ionized.

4. Plot the conductivity vs temperature for an undoped silicon. Compare this with a commercial thermistor, such as for example:
   https://www.vishay.com/docs/29050/ntclg100.pdf

5. Commercially manufactured silicon wafers have a number of different specifications, one of which is its resistivity. 3-inch silicon wafers with a resistivity greater than $20\text{k}\Omega\cdot$cm are significantly more expensive than those with resistivity values smaller than $10\Omega\cdot$cm. Explain the reasons for this.

6. Calculate the Fermi level of an n-type GaAs substrate that has been doped with 10^{17}cm^{-3} relative to either of the band edges (E_c or E_v). Provide your answers in eV.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val m0 = 9.1e-31
    val q = 1.602e-19
    val h = 6.62607004e-34
    val k = 1.38064852e-23
    val mc = 0.063*m0
    val T = 300.0
    val n = 1.0e17*1.0e6
    println ("Ef-Ec = %.4f".format(
        ln(n/(4.0*2.0.pow(0.5)/h.pow(3) * (PI*mc*k*T).pow(1.5)))*(k*T)/q))
}

>>Ef-Ec = -0.0356
```

7. Look up the bandgaps, cut-on wavelengths, and intrinsic carrier concentrations of InSb, Si, GaAs and GaN. Assume room temperature for all cases. Discuss the trends in these values.

```
InSb = 0.17eV Lambda = 7.29um     (ni = 2e16)
Si = 1.12eV Lambda = 1.107um      (ni = 1e10)
GaAs = 1.42eV Lambda= 0.873um     (ni = 2e6)
GaN = 3.2eV, Lambda = 0.387um     (ni = 1e-10)
```

8. A photocell consists of a thin film of CdS with an interdigitated electrode structure. The resistivity of CdS (with no illumination) is $10^3\Omega-$cm. The film thickness 10μm. Referring to Fig 8, the gap between the electrodes is 1 mm, and the length of the interdigitated trace is 1 cm. Calculate the dark resistance of this photocell.

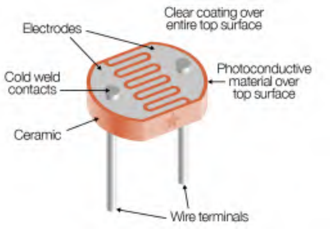
Electrodes
Clear coating over entire top surface
Cold weld contacts
Photoconductive material over top surface
Ceramic
Wire terminals

Figure 6: CdS Photocell

Run this code

```
import kotlin.math.∗
//Andrew Sarangan

fun main() {

    val rho = 1.0e3       //Ohm–cm
    val thickness = 10.0e–4 //cm
    val distance = 0.1     //cm
    val width = 1.0       //cm
    val R:Double = rho∗distance/(thickness∗width)
    println(”${”%.1f”.format(R/1000.0)} kOhms”)
}

>>100.0 kOhms
```

Basic Theory of PN Junction Diodes

Current Flow

Before describing diodes, we need to explain the mechanisms of current flow. The ordinary current flow that we know from everyday experience is the drift current. This current is produced by applying a voltage across a conductor. The current density J (in Amps/cm^2) will be

$$J_{\text{drift}} = q\mu_n nE \tag{1}$$

where μ_n is the electron mobility, n is the electron density, E is the applied electric field (which is voltage divided by distance) and q is the unit charge of an electron (assumed to be a positive value). This is the current flow that we are all familiar with through metal wires such as copper and aluminum as well as through uniformly doped semiconductors.

The second mechanism of current flow is diffusion current. This is less common in everyday experience, but it is an important mechanism in diodes. This current is produced when there is a concentration gradient of electrons or holes. The electrons (or holes) will flow from high concentration to low concentration through a diffusive process without the assistance of an electric field. The current density J (in Amps/cm^2) due to the diffusion of electrons can be written as

$$J_{n,\text{diff}} = qD_n\frac{dn}{dx}. \tag{2}$$

Similarly, the current due to hole diffusion can be written as

$$J_{p,\text{diff}} = -qD_p\frac{dp}{dx}. \tag{3}$$

μ and D are actually related, which should not be surprising because both are related to the transport of these carriers. This relationship is:

$$D = \mu kT/q. \tag{4}$$

Diffusion current does not typically exist in a metal wire because the electron concentration is nearly the same everywhere inside the metal. We encounter diffusion current only when there are large gradient in concentration. This can occur at the interface between two metals, or more importantly, at the interface between an n-type and p-type semiconductor.

Built-in Voltage of a Junction Diode

If an n-type region and a p-type region are adjacent in the same semiconductor crystal, there will be a large concentration gradient for electrons and holes across the junction. In the n-type region, the electron concentration will be equal to N_D, and in the p-side the electron concentration will be equal to n_i^2/N_A. Substituting typical values of N_A and N_D of 10^{16} cm^{-3}, we can see that the electron concentration will vary from 10^{16} on the n-side to 10^4 on the p-side.

This is a difference of 12 orders of magnitude. This concentration difference will force diffusion current to flow. As electrons leave the n-side and move to the p-side, it will make the n-side of the junction more positive and p-side of the junction more negative. As holes leave the p-side and move to the n-side, that too will leave the p-side more negative and n-side more positive. The net result is that an internal electric field will build which points from the n-side (positively charged) to the p-side (negatively charged). The associated voltage difference is called the built-in potential. This electric field will oppose the flow of diffusion current. For this reason, the built-in potential is also called a barrier voltage. The charges that create this potential difference reside within a narrow region on either side of the junction, and this is known as the space-charge region.

The built-in voltage can be calculated by setting the drift and diffusion current across the junction to be equal and opposite to each other.

$$J = J_{\text{drift}} + J_{\text{diff}} = 0 \tag{5}$$

$$= q\mu_n n E + qD_n \frac{dn}{dx} = 0. \tag{6}$$

From this, we can get an expression for E, which we can also write as $-\frac{dV}{dx}$ where V is the voltage:

$$\frac{dV}{dx} = \frac{D_n}{\mu_n}\frac{dn/dx}{n}. \tag{7}$$

Since we know that $\frac{D_n}{\mu_n} = \frac{kT}{q}$, which we can represent as a voltage V_t, equation (7) becomes

$$dV = V_t \frac{dn}{n}. \tag{8}$$

This function can be integrated from one side the junction to the other side where the electron concentration goes from n_i^2/N_A (minority carriers) to N_D (majority carriers), such as

$$\int_0^{V_{bi}} dV = \int_{n_i^2/N_A}^{N_D} V_t \frac{dn}{n} \tag{9}$$

resulting in:

$$V_{bi} = V_t \ln\left(\frac{N_D N_A}{n_i^2}\right). \tag{10}$$

V_{bi} is known as the built-in voltage of a junction diode. The internal voltage builds up to prevent the continuous flow of carriers from one side of the junction to the other. For example, if N_A and N_D are both equal to 10^{16}cm^{-3}, we can get $V_{bi} = 0.70V$ for silicon, which is the frequently used built-in voltage for silicon diodes.

The built-in potential, therefore, raises the potential of the n-side compared to the p-side. Since conventional definition of current and potential is with respect to positive charges, we can also view the *electron potential* as being higher on the p-side compared to the n-side, as shown in Fig 1. The potential profile can be viewed as a barrier that prevents the carriers from spilling over to the other side (or as a levee that prevents water from flooding the other side). One important

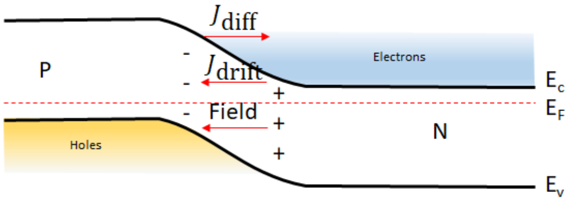

Figure 1: Potential band diagram of a PN junction under zero bias.

consequence worth noting is that the Fermi level will be the same on both sides of the junction. This is a necessary condition for equilibrium. It represents the condition that the average electron energy is the same everywhere in the structure to represent no net movement of energy. Therefore, we could have also derived the built-in voltage expression by simply taking the difference between the Fermi levels, as

$$qV_{bi} = E_{FN} - E_{FP} \tag{11}$$

where E_{FN} and E_{FP} are the n-type and p-type Fermi levels *prior to the formation of the junction*. These can be calculated from the doping levels using equations (26) and (27) from the previous chapter.

Another point worth noting is that the Fermi level passes through the intrinsic point (midway inside the bandgap) at the junction. This is known as the metallurgical junction. The carrier concentrations will be at the intrinsic semiconductor value n_i at this point.

We can sketch the carrier profiles for both electrons and holes from one side of the junction to the other side. Clearly, electron concentration would start at a value of N_D on the n-side and decline to a value of n_i^2/N_A on the p-side. The hole concentration would start at a value of N_A on the p-side and decline to n_i^2/N_D on the n-side. This is shown in Figure 2.

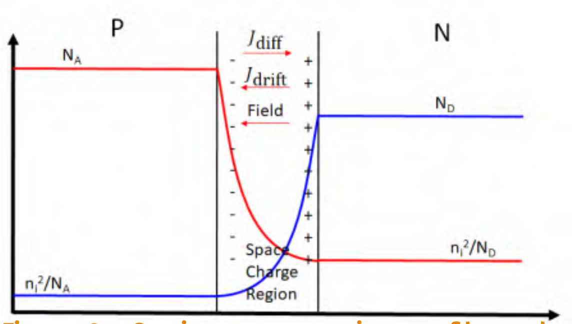

Figure 2: Carrier concentration profile under zero bias.

Forward Bias

When no external voltage is applied, the built-in potential barrier will prevent the diffusion current from flowing across the junction. However, we can apply a voltage to reduce the barrier voltage, which will then cause a diffusion current to flow. Since the built-in voltage is a result of the n-side of the junction being positively charged and the p-side of the junction being negatively charged, the built-in voltage can be reduced by applying an external positive voltage to the p-side and a negative voltage to the n-side. This is known as the forward bias condition of the diode. This will cause an imbalance in the drift and diffusion currents, and will cause a net current to flow in the direction of the diffusion current.

This condition can be represented as in Fig 3, where the potential barrier is now insufficient to prevent the carriers from spilling over to the other side. Using the analogy of a levee, this is like reducing the height of the levee and allowing some of the flood water to flow. There will be a net movement of energy from the n-side to the p-side, and the Fermi levels will also be different on either side of the PN junction. E_{FN} will be higher than E_{FP}, because the average energy of electrons on the

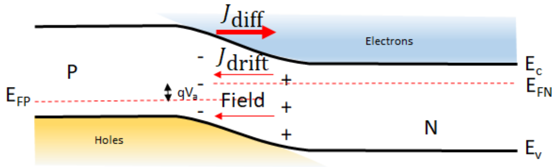

Figure 3: Potential band diagram of a PN junction under forward bias.

n-side will be higher than the p-side. The difference in the Fermi levels is equal to applied voltage:

$$E_{FN} - E_{FP} = qV_a. \tag{12}$$

We can sketch the carrier profiles for both electrons and holes from one side of the junction to the other side. Similar to the zero bias, the electron concentration would start at a value of N_D on the n-side and decline to a value of n_i^2/N_A on the p-side. However, at the edge of the space charge region, the electron concentration will be elevated as illustrated in Figure 4. Similarly, the hole concentration would start at a value of N_A on the p-side and decline to n_i^2/N_D on the n-side, except it will be elevated at the edge of the other side of the space charge region.

Looking at equation (10), we can also write it as

$$\frac{n_i^2}{N_A} = N_D e^{-V_{bi}/V_t}. \tag{13}$$

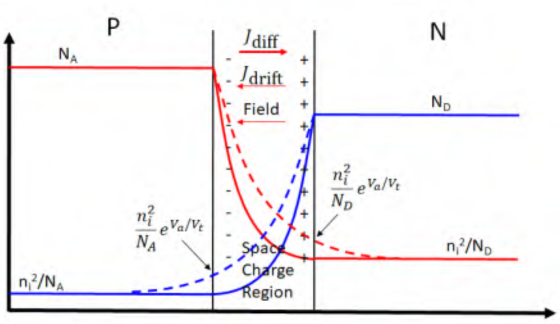

Figure 4: Potential band diagram of a PN junction under forward bias. The solid lines represent zero bias, and the dashed lines represent forward bias.

The left side of this equation is the minority carrier concentration in the p-side. We will represent it more generically as n_p (where the letters stand for electron concentration in the p-side). At zero bias, n_p will be equal to $\frac{n_i^2}{N_A}$. Under forward bias, n_p at the left edge of the space charge region will be higher than $\frac{n_i^2}{N_A}$. We will make the assumption that equation (13) still holds even at forward bias (hence the quasi-steady-state assumption). Since the net voltage across the junction is now $V_{bi} - V_a$, the minority carrier concentration at the edge of the space charge region on the p-side becomes

$$n_p = N_D e^{-(V_{bi}-V_a)/V_t}. \tag{14}$$

Combining equation (14) with (13) results in:

$$n_p = \frac{n_i^2}{N_A} e^{V_a/V_t}. \tag{15}$$

In other words, the minority electron concentration at the p-side of the junction will rise by a factor of e^{V_a/V_t}. This is an excess electron concentration above the steady state value of $\frac{n_i^2}{N_A}$. Similarly, on the n-side, there will be an excess of hole concentration:

$$p_n = \frac{n_i^2}{N_D} e^{V_a/V_t}. \tag{16}$$

These excess carriers will result in an excess recombination and excess diffusion current on both sides of the junction. The derivation is given in the next section, where we have shown that the diffusion length can be expressed as

$$L_n = \sqrt{D_n \tau} \tag{17}$$
$$L_p = \sqrt{D_p \tau}, \tag{18}$$

where τ is the electron-hole recombination lifetime and D_n and D_p are the electron and hole diffusion coefficients, respectively. The diffusion length can be interpreted as the average distance a carrier moves before it recombines and annihilates itself. The carrier profiles will exhibit a decay such as e^{-x/L_p} on the N-side and e^{x/L_n} on the P-side. This is illustrated in Figure 5.

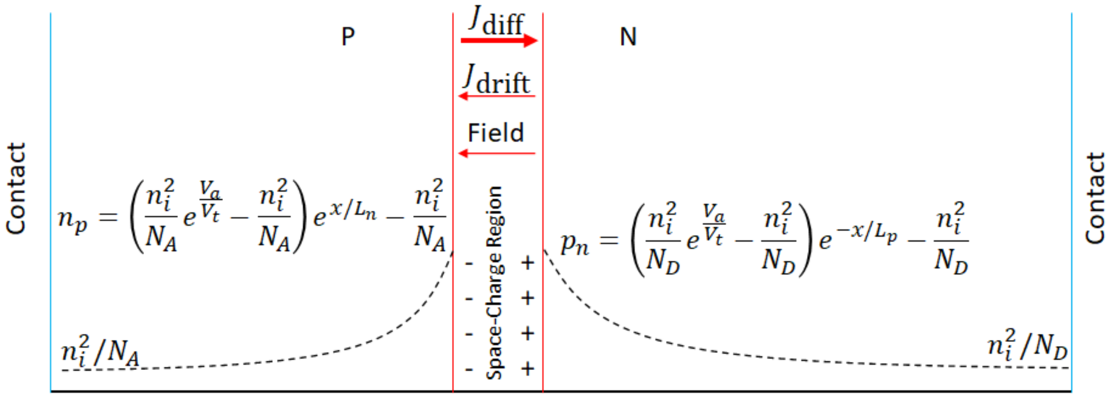
P
N
J_{diff}
J_{drift}
Field
Contact
Contact
$n_p = \left(\frac{n_i^2}{N_A} e^{\frac{V_a}{V_t}} - \frac{n_i^2}{N_A}\right) e^{x/L_n} - \frac{n_i^2}{N_A}$
$p_n = \left(\frac{n_i^2}{N_D} e^{\frac{V_a}{V_t}} - \frac{n_i^2}{N_D}\right) e^{-x/L_p} - \frac{n_i^2}{N_D}$
Space-Charge Region
- - - -
+ + + +
n_i^2/N_A
n_i^2/N_D

Figure 5: Minority carrier profiles under forward bias when the contacts are far from the junction (long diode model)

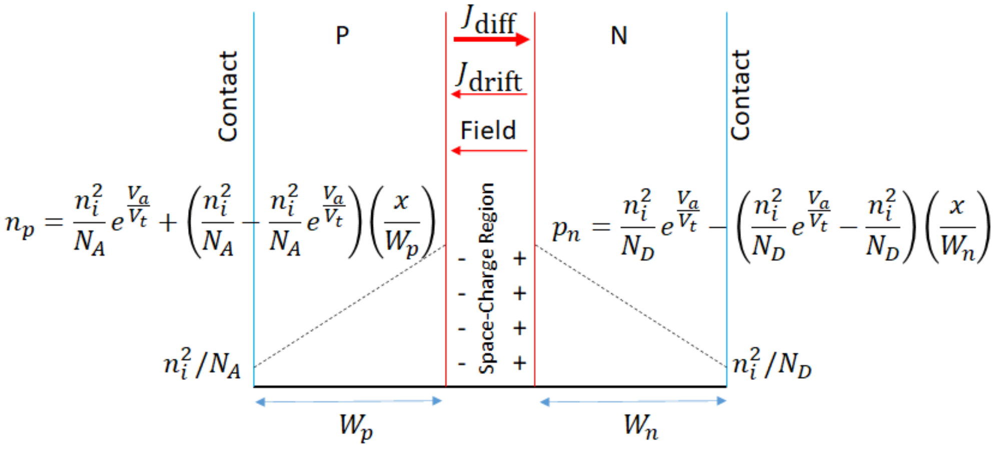
P
N
J_{diff}
J_{drift}
Field
Contact
Contact
$n_p = \frac{n_i^2}{N_A} e^{\frac{V_a}{V_t}} + \left(\frac{n_i^2}{N_A} - \frac{n_i^2}{N_A} e^{\frac{V_a}{V_t}}\right)\left(\frac{x}{W_p}\right)$
$p_n = \frac{n_i^2}{N_D} e^{\frac{V_a}{V_t}} - \left(\frac{n_i^2}{N_D} e^{\frac{V_a}{V_t}} - \frac{n_i^2}{N_D}\right)\left(\frac{x}{W_n}\right)$
Space-Charge Region
- - - -
+ + + +
n_i^2/N_A
n_i^2/N_D
W_p
W_n

Figure 6: Minority carrier profiles under forward bias when the contacts are close to the junction (short diode approximation)

If the distance between the edge of the space charge region and the electrical contacts is much smaller than the diffusion length (known as the short-diode approximation), the carrier profile can be assumed to be linear, declining from n_p at the edge of the space charge region to $\frac{n_i^2}{N_A}$ at the contact (for minority carrier electrons in the p-side). This is shown in Fig 6.

In both cases, the sum of the diffusion currents on both sides of the junction can be expressed as (derivation left as an exercise)

$$\begin{aligned} J &= J_n + J_p && (19) \\ &= q\frac{D_n}{L_n}\left(\frac{n_i^2}{N_A}e^{V_a/V_t} - \frac{n_i^2}{N_A}\right) + q\frac{D_p}{L_p}\left(\frac{n_i^2}{N_D}e^{V_a/V_t} - \frac{n_i^2}{N_D}\right) && (20) \\ &= q\left(\frac{D_n}{L_n}\frac{n_i^2}{N_A} + \frac{D_p}{L_p}\frac{n_i^2}{N_D}\right)\left(e^{V_a/V_t} - 1\right). && (21) \end{aligned}$$

In the case of the short-diode, the equation becomes

$$J = q\left(\frac{D_n}{W_n}\frac{n_i^2}{N_A} + \frac{D_p}{W_p}\frac{n_i^2}{N_D}\right)\left(e^{V_a/V_t} - 1\right). \tag{22}$$

where W_n and W_p are the distance between the edge of the space charge region and the contact in the n-type and p-type regions, respectively.

It is also possible to have the short-diode on one side the junction and a long-diode on the other side. Similarly, it is also possible to have one of the terms much smaller than the other. For example, consider a diode where $N_A >> N_D$. In this case, the long-diode model becomes:

$$J = q\frac{D_p}{L_p}\frac{n_i^2}{N_D}\left(e^{V_a/V_t} - 1\right). \tag{23}$$

If the cross-sectional area of the diode is A, this becomes

$$I = Aq\frac{D_p}{L_p}\frac{n_i^2}{N_D}\left(e^{V_a/V_t} - 1\right) \tag{24}$$

$$I = I_s\left(e^{V_a/V_t} - 1\right), \tag{25}$$

which is the well-known I-V curve of a classical diode. The overall I-V curve of a diode is depicted in Fig 7.

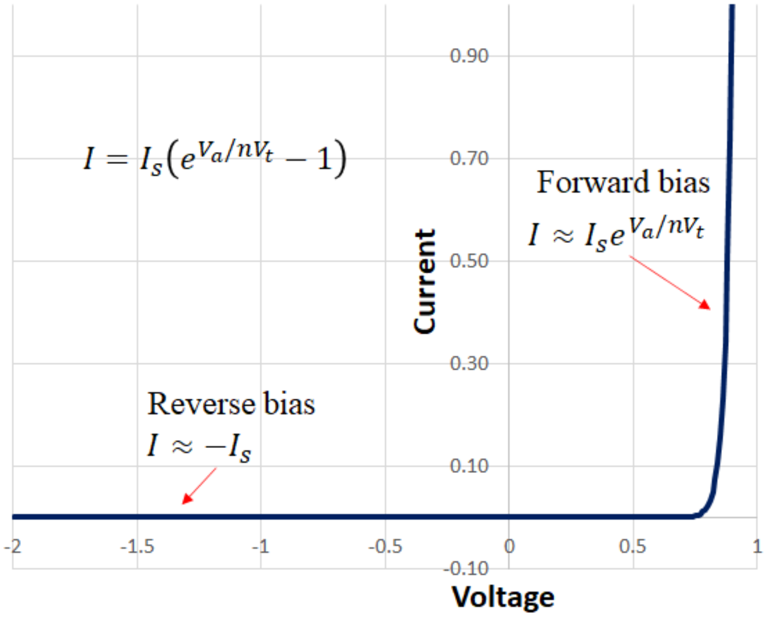

Figure 7: Current-Voltage (I-V) curve of a typical diode

We can see that the current on the reverse bias side is practically zero compared to the current on the forward bias side. Despite its small value, the reverse current still plays a major role in devices such as photodiodes and avalanche diodes. Furthermore, in practical diodes, the reverse current will not remain constant at $-I_s$, but will actually increase with increasing reverse bias voltage. These small but important effects cannot be discerned from the linear plot shown

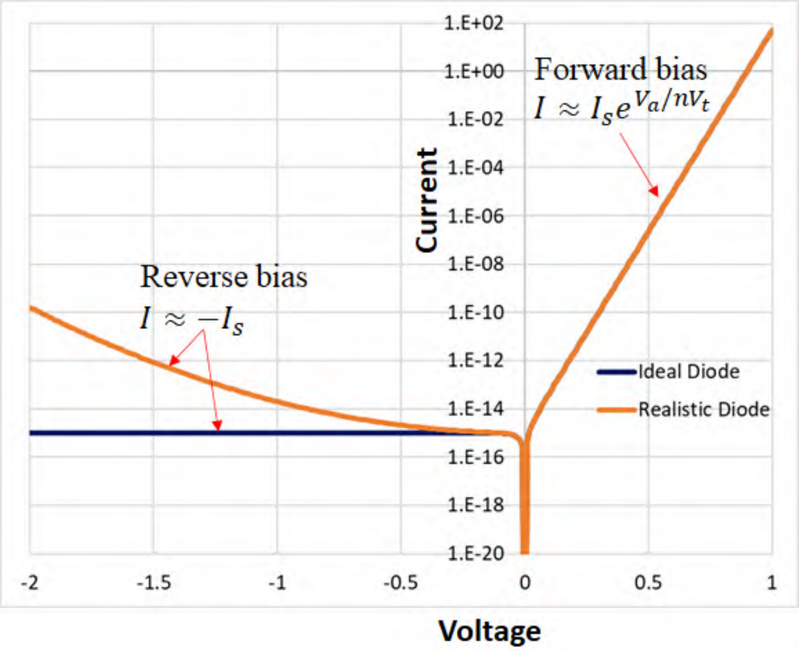

Figure 8: Current-Voltage (I-V) curve of an ideal diode and a real diode shown on a log scale

in Fig 7. A better way is to depict the current on a log scale (after taking its magnitude, to account for the negative current in the reverse direction). This is shown in Fig 8.

As an important side note, we can state that majority carriers move primarily by drift rather than diffusion. This is because even a tiny electric field can cause an appreciable flow of majority carriers. On the other hand, the transport of minority carriers is by diffusion. On Fig 5, we can see that the current (which is the sloped of the carrier profile) will decline from the edge of the space charge region. This occurs due to recombination. As minority carriers recombine, they will cause the majority carriers to move in the opposite direction to preserve the total current. In the long-diode model, the minority carrier profile by the time it reaches the contact is nearly flat. Therefore, the diffusion current will be zero, and all of the current will be carried by drift. This means, the electric field cannot be exactly zero outside the space charge region. In the integration of equation (8) we assumed the field was zero outside the space charge region. While this is strictly true under zero-bias condition, under forward bias, a small electric field has to exist outside the space charge region to support the drift current. Similarly, the injection of majority carriers from the incident side of the junction also occurs by drift transport. The presence of this field can be ignored in most cases, but it will become important in some cases, such as in photodiodes as we will see later.

Reverse bias

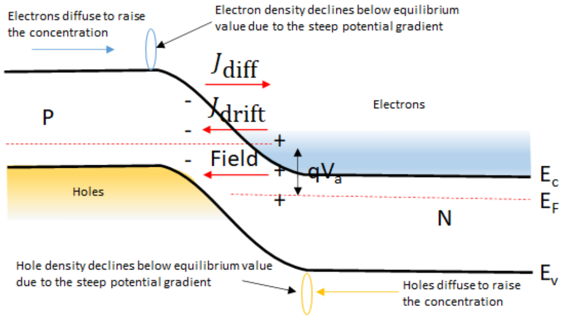
Electrons diffuse to raise the concentration
Electron density declines below equilibrium value due to the steep potential gradient
J_{diff}
J_{drift}
Field
P
Electrons
Holes
qV$_a$
E$_C$
E$_F$
N
E$_V$
Hole density declines below equilibrium value due to the steep potential gradient
Holes diffuse to raise the concentration
- - -
+ + + +

Figure 9: Potential band diagram of a PN junction under reverse bias.

Whereas a forward bias caused a reduction in the potential barrier, a reverse bias will increase the potential barrier as depicted in Fig 9. From equation (21) or (22), we can see that when V_a is negative, the current will reverse direction with a much smaller magnitude. This effect can be explained as follows:

Using the analogy of a levee, applying a reverse bias is like raising the height of the barrier. Clearly, raising the barrier should only increase the blocking effect of the barrier, and should not cause a reverse flow. However, this would be true only if the top of the barrier were absent of any carriers. There is , in fact, small number of carriers at the top of the barrier (minority carriers). When the barrier is raised, these carriers will roll down the barrier to produce a reverse flow. This will result in a depression in the minority carrier concentration at the top of the barrier. The minority carriers further away from this depression will diffuse to fill in the depression. This is how reverse current flows in a diode.

If we return to equations (15) and (16), as V_a becomes negative, the minority carrier densities at the edge of the space charge region will decline below their thermal equilibrium values (whereas in forward bias they increased above their thermal equilibrium values). In response to this depression of carrier density, minority carriers from further away will diffuse in to compensate for this depression. Given the small carrier densities, this carrier gradient will be small, and the resulting flow will also be small. The carrier profile corresponding to the reverse bias is illustrated in Figure 10. Since the slope of the carrier density at the edge of the space charge region is opposite that of the forward bias case, the current will also reverse direction.

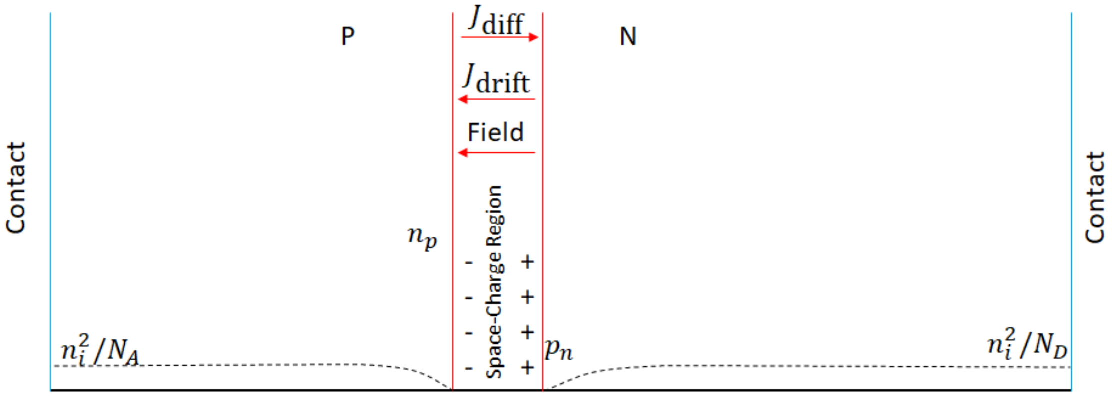

Figure 10: Minority carrier profiles under reverse bias in a long diode approximation

Additionally, the depression in carrier density below the thermal equilibrium value will cause the thermally generated carriers to become larger than the recombination rate, resulting in a net production of carriers. Whereas in a forward bias the injected excess minority carriers recombined to become majority carriers, under reverse bias the drawn minority carriers will be supplied by thermal generation. Since thermal generation creates an electron and a hole simultaneously, the effect is the same as before - the minority carriers become majority carriers, except in this case it is due to thermal generation.

Space Charge Width

The width of the space charge region extends on both sides of the metallurgical junction. There is equal number of charges in the space charge region on either side of the junction. Since the doping density is N_A and N_D for the p- and n-side, we can get an expression for the width of this region on either side of the junction. These are:

$$x_n = \sqrt{\frac{2\epsilon_s}{q}\frac{N_A}{N_D}\frac{1}{N_A+N_D}(V_{bi}-V_a)} \tag{26}$$

$$x_p = \sqrt{\frac{2\epsilon_s}{q}\frac{N_D}{N_A}\frac{1}{N_A+N_D}(V_{bi}-V_a)}, \tag{27}$$

where ϵ_s is the static permittivity of the semiconductor material. In these expressions, we have assumed V_a to be forward bias voltage. Therefore, under forward bias the space charge width will get smaller. Under reverse bias the space charge width will get larger.

Carrier Diffusion and Recombination

Equations (17) and (18) actually comes from the carrier diffusion equation, which is similar to the heat diffusion equation. This can be written as:

$$D_n\frac{\partial^2 n}{\partial x^2} = -\frac{n}{\tau}. \tag{28}$$

This equation basically states that any discontinuity in the diffusion current is due to the carriers lost due to recombination. A similar equation can be written for the hole diffusion:

$$D_p\frac{\partial^2 p}{\partial x^2} = -\frac{p}{\tau}. \tag{29}$$

The solution of these equations, assuming the boundary (in this case, the boundary is the electrical contacts) is at an infinite distance away, becomes:

$$n = Ae^{-x/\sqrt{D_n\tau}} + C \tag{30}$$

$$p = Be^{-x/\sqrt{D_p\tau}} + D \tag{31}$$

Since we know that the minority carrier concentrations at the edge of the junction was n_p on the p-side and p_n on the n-side, and the minority carrier concentrations far from the junction are n_i^2/N_A and n_i^2/N_D we can write these as

$$n = n_pe^{-x/\sqrt{D_n\tau}} + \frac{n_i^2}{N_A} \tag{32}$$

$$p = p_ne^{-x/\sqrt{D_p\tau}} + \frac{n_i^2}{N_D}, \tag{33}$$

which are the same carrier profiles assumed in the long-diode approximation. For the case where the contacts are much closer to the junction, we can also show that the carrier profiles will have a linear profile (short diode approximation).

In equations (32) and (33), the minority carrier profiles decay to from n_p to n_i^2/N_A. As a result, the diffusion current, which is the derivative of the carrier profile, also decays from its highest value at the edge of the junction (which will be $q\frac{D_n}{L_n}n_p$ on the p-side and $q\frac{D_p}{L_p}p_n$ on the n-side) to zero. This decay is due to carrier recombination. This recombination, which occurs on both sides of the junction can emit a photon or phonon. A photon will emit electromagnetic radiation, while a phonon will increase the lattice vibrations, and hence the temperature. The recombination lifetime due to photon emission is called the radiative lifetime, and the recombination lifetime due to phonon emission is called the non-radiative lifetime. We can write this as

$$\begin{aligned} R &= R_{nr} + R_r && (34) \\ \frac{n_p}{\tau} &= \frac{n_p}{\tau_{nr}} + \frac{n_p}{\tau_r} && (35) \end{aligned}$$

where τ_r and τ_{nr} are the radiative and non-radiative lifetimes, respectively.

In many applications, we are interested in radiative processes more than non-radiative processes. This can be quantified by taking the ratio between the two rates. This is often expressed as:

$$\eta_i = \frac{R_r}{R} = \frac{\tau_{nr}}{\tau_{nr} + \tau_r} \tag{36}$$

where η_i is known as the internal quantum efficiency, and is an intrinsic property of a material. Its value depends on the type of bandgap (direct or indirect), as well as the quality of the crystal and doping etc..

In silicon, $\tau_{nr} \approx 100\mu$s, and $\tau_r \approx 1$s, resulting in an internal quantum efficiency of $\eta_i = 10^{-4}$. Silicon is an indirect bandgap material, therefore, it is not very likely to emit a photon during a recombination event. In GaAs, $\tau_{nr} = 10$ns and $\tau_r = 5$ns, which results in an internal quantum efficiency of $\eta_i = 0.66$.

As depicted in Fig 5, the recombination takes place on either side of the junction. Additionally, from 4 we can see that both the electron and hole concentration profiles inside the space charge region are higher than their thermal equilibrium values. This will result in recombination inside the space charge region as well. As a result, we can say that recombination occurs within a two-dimensional plane whose thickness is $x_n + x_p + L_n + L_p$.

Homework 3

1. This question pertains to an InSb diode.

   - Using equation (13) from the previous chapter calculate the intrinsic carrier concentration of InSb at 77K and 300K.
   - Using the above values, calculate the built-in voltage of an InSb diode at 77K and 300K, using a p-side doping of 10^{16}cm^{-3} and n-side doping of 10^{17}cm^{-3}.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
  val NA = 1.0e16
  val ND = 1.0e17
  val mc = 0.014
  val mv = 0.48
  val m0 = 9.1e-31
  val Eg = 0.17
  val k = 1.38e-23
  val h = 6.602e-34
  val q = 1.602e-19

    fun niVbi(T:Double):Pair<Double,Double>{
        val ni = 4.0*2.0.pow(0.5)/h.pow(3)*(PI*k*T).pow(1.5)*(mc*mv*m0*m0).
            pow(0.75)*exp(-Eg/(2.0*k*T/q)) / 1.0e6
        val Vt = k*T/q
        val Vbi = Vt*ln(NA*ND/ni.pow(2))
        return Pair<Double,Double>(ni,Vbi)
    }

    niVbi(300.0).let{
        println(”ni at 300 K = ${”%.3e”.format(it.first)}”)
        println(”Vbi at 300 K = ${”%.4f”.format(it.second)}”)
    }

    niVbi(77.0).let{
        println(”ni at 77 K = ${”%.3e”.format(it.first)}”)
        println(”Vbi at 77 K = ${”%.4f”.format(it.second)}”)
    }
}

>>ni at 300 K = 2.215e+16
>>Vbi at 300 K = 0.0184
>>ni at 77 K = 2.101e+11
>>Vbi at 77 K = 0.1582
```

2. Consider a silicon photodiode as shown in Figure 11. The diode is fabricated on an n-type substrate whose resistivity is 1kΩ·cm. The p-side is doped to a concentration of 10^{17}cm^{-3} to a depth of 3μm. The substrate thickness is 300μm. The diameter of the diode is 250μm.

   - Calculate the substrate doping.
   - Considering only the ideal effects described in this chapter, calculate the reverse saturation current I_s of this photodiode (which is also known as the dark current).
   - Explain why the top contact is shaped as a ring while the bottom contact is shaped as a disc.

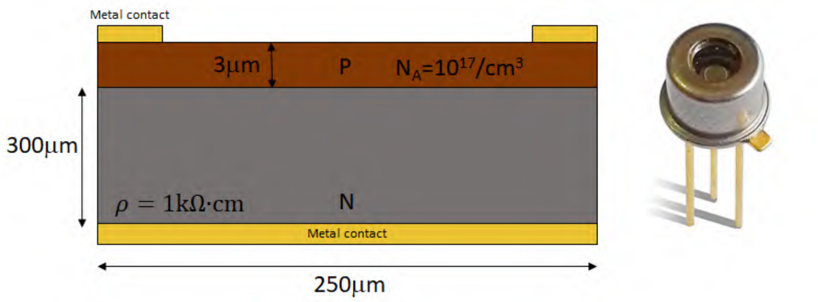

Figure 11: Silicon photodiode

Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val muN = 1400.0
    val q = 1.602e—19
    val rho = 1.0e3
    val ND = 1.0/(rho*q*muN)
    println(”N—side doping = ${”%.3e”.format(ND)}”)

    val Lp = 0.7
    val Dp = 12.0
    val A = PI*(250.0e—4/2).pow(2)
    val ni = 1.0e10
    val Is = q*A*Dp/Lp*ni.pow(2)/ND
    println(”Is = ${”%.3e”.format(Is)} Amps”)
}

>>N—side doping = 4.459e+12
>>Is = 3.023e—14 Amps
```

Light Emitting Diodes

Rectifying Diodes vs Light Emitting Diodes

The most widely used property of diodes is its asymmetric electrical conductance. Rectifying diodes are used to convert AC voltages to DC, Zener diodes used to regulate voltages etc.. Step-recovery diodes (SRD) are used to generate very short pulses due to the reversal from forward bias to reverse bias. Schottky barrier diodes (SBD) are formed between a semiconductor (typically silicon) and a metal, which can allow for a lower turn-on voltage compared to a p-n junction silicon diode.

A light emitting diode, on the other hand, is specifically designed to emit photons. That means, it has to be made from a direct bandgap material. Additionally, it is designed to maximize the recombination current component. It is also designed to extract and transmit the maximum number of photons from the semiconductor to the outlying regions, and reduce the number of photons that are re-absorbed or trapped inside the semiconductor chip.

Internal Quantum Efficiency

The internal quantum efficiency η_i is the fraction of recombinations that produce photons:

$$\eta_i = \frac{R_r}{R_{nr} + R_r}. \tag{1}$$

In terms of lifetimes, we can also express it as

$$\eta_i = \frac{\tau_{nr}}{\tau_{nr} + \tau_r}. \tag{2}$$

The primary factor that determines the internal quantum efficiency is the alignment of the conduction and valence band edges. Direct bandgap materials such as GaAs and InP exhibit a large internal quantum efficiency, in the range of 10%- 90%, while indirect bandgap materials have extremely small internal quantum efficiency.

Extraction Efficiency

Extraction efficiency η_{ext} is the fraction of emitted photons that emerge out of the LED structure. The characteristic of spontaneous emission is such that photons are randomly emitted in all directions. Most semiconductor materials have a high refractive index. As a result, only a small fraction of these photons will fall within the cone of angles will be able to escape the semiconductor. Most of the other photons will be trapped inside the semiconductor material due to total internal reflection, and be eventually absorbed.

Assuming a refractive index of 3.5 (which is typical for GaAs near the band edge), the critical angle for total internal reflection is $\theta_{cr} = \sin^{-1}\left(\frac{1}{3.5}\right) = 16.6°$. Therefore, only those photons that fall within this angle (referred to as the escape cone) will emerge from GaAs. All others angles will be internally reflected.

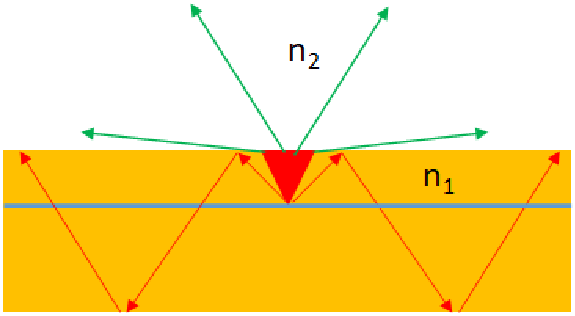

Figure 1: Fraction of photons that escape the semiconductor.

We can estimate the fraction of spontaneously emitted photons that fall within this escape cone. The emitting surface in an LED is essentially a two-dimensional plane. Even though each point on the surface emits isotropically, when these points are evenly distributed on a two-dimensional plane, the overall intensity will have an angular dependence of $\cos\theta$, where θ is measured from the surface-normal. This is known as Lambert's cosine law, and it arises due to the *radiance* (or brightness) being equal in all direction. Therefore, the intensity can be written as

$$I = I_o \cos\theta. \tag{3}$$

Using this assumption, we can calculate the fraction of photons that fall within the escape cone. Referring to Fig 2, the integrated power that fall between the normal direction to the emitting surface and θ_{cr} can be written as an integral in spherical co-ordinates:

$$P_{ext} = \int_0^{\theta_{cr}} (2\pi r \sin\theta)(rd\theta)(I_o \cos\theta)\, T(\theta), \tag{4}$$

where $T(\theta)$ is the transmission from the semiconductor to the air, and r is an arbitrary radius. The total emitted power in all directions can be obtained by taking the same integral over all angles:

$$P_{tot} = 2\int_0^{\pi/2} (2\pi r \sin\theta)(rd\theta)(I_o \cos\theta), \tag{5}$$

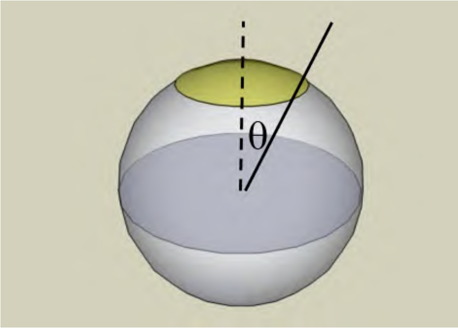

Figure 2: Depiction of the integral in spherical coordinates of the escape cone from a semiconductor

where the factor of 2 is to account for the top and bottom hemispheres. The extraction efficiency can be obtained by taking the ratio between these two powers. The transmission $T(\theta)$ falls from $1 - \left|\frac{n_s - n_a}{n_s + n_a}\right|^2$ at normal incidence ($\theta = 0$) to zero at the critical angle. Given that θ_{cr} is a relatively small angle, we can approximate the transmission function with an average value T_{av}.

Therefore, we can carry out this integral as follows:

$$\begin{aligned}
\eta_{ext} &= \frac{P_{ext}}{P_{tot}} && (6)\\
&= \frac{\int_0^{\theta_{cr}} \left(2\pi r \sin\theta\right)\left(r d\theta\right)\left(I_o \cos\theta\right) T\left(\theta\right)}{2\int_0^{\pi/2} \left(2\pi r \sin\theta\right)\left(r d\theta\right)\left(I_o \cos\theta\right)} && (7)\\
&= \frac{T_{av}\int_0^{\theta_{cr}} \sin\theta \cos\theta \, d\theta}{2\int_0^{\pi/2} \sin\theta\cos\theta \, d\theta} && (8)\\
&= \frac{\frac{1}{2}\sin^2\theta\Big|_0^{\theta_{cr}}}{\sin^2\theta\Big|_0^{\pi/2}} T_{av} && (9)\\
&= \underbrace{\frac{1}{2}\left(\frac{n_2}{n_1}\right)^2}_{\text{Escape Cone}} T_{av} && (10)
\end{aligned}$$

where n_1 is the refractive index of the LED, and n_2 is the refractive index of the exit medium (air). In this calculation, we have ignored the contribution of absorption in the semiconductor. In practice, photons have to travel a finite distance from the emission surface to the air interface. A semiconductor that emits photons will also absorb that same photon. Therefore, the absorption coefficient must also be considered in the calculation of extraction efficiency. Unfortunately, this factor cannot be neatly integrated as shown above. It can only be done numerically. For that reason, we will ignore this absorption factor.

Additionally, we have assumed that all of the photon emitted into the lower hemisphere are completely lost. This does not always have to be the case. Those photons can be redirected to the top surface with reflectors. However, the limitation of the escape cone is still valid. Therefore, at best, we can improve the extraction by a factor of two by redirecting all of the photons in the lower hemisphere.

T_{av} is the average transmission coefficient across the semiconductor/air interface within the escape cone. This can be approximated as one half of the normal incidence transmission:

$$T_{av} = \frac{1}{2}\left[1 - \left|\frac{n_1 - n_2}{n_1 + n_2}\right|^2\right]. \qquad (11)$$

External Quantum Efficiency

The external quantum efficiency can now be calculated by taking the product of the internal quantum efficiency and the extraction efficiency:

$$\eta_e = \eta_i \eta_{ext}. \qquad (12)$$

For example, if the refractive index of the LED is 3.5, we can calculate $\eta_{ext} = 0.0141$. If the internal quantum efficiency is $\eta_i = 0.5$, we can get an external quantum efficiency of 0.7%. In other words, only 0.7% of the electrons injected into the diode emerge as photons out of the LED.

Once the external quantum efficiency is known, we can calculate different parameters, such as responsivity and wall plug efficiency.

Responsivity

The responsivity of an LED is defined as the optical output power for input current. The units are in Watts/Amps. We can write the output power as

$$P_o = \frac{I}{q} h\nu \eta_i \eta_{ext} \tag{13}$$

where $h\nu$ is the photon energy and I is the diode current. Since the photon energy is equal to the bandgap energy, E_g, and $\eta_e = \eta_i \eta_{ext}$, we can also write this as

$$P_o = I \frac{E_g}{q} \eta_e. \tag{14}$$

Furthermore, if we define $\frac{E_g}{q}$ as an equivalent voltage V_g, we can

$$P_o = IV_g\eta_e = I\mathcal{R} \tag{15}$$

where the responsivity R is

$$\mathcal{R} = V_g\eta_e. \tag{16}$$

Using the previous example, assuming GaAs as the LED material whose bandgap is 1.42eV, we can get

$$\mathcal{R} = 1.42{\times}0.007 = 0.01\ \text{W/A or } 10\ \text{mW/A}. \tag{17}$$

The Light-Current curve of an LED is shown in Fig 3. We can see that the slope of the curve is the responsivity. However, this responsivity will generally decline at higher current levels due to heating, which arises from a reduction in the internal quantum efficiency.

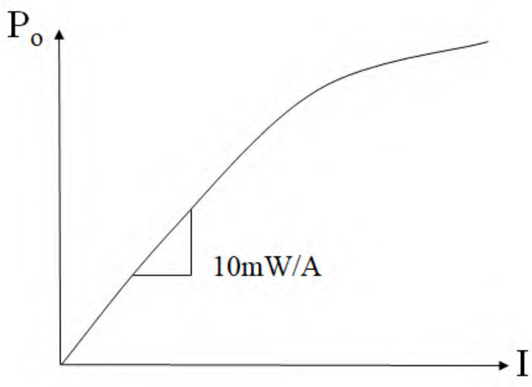

Figure 3: Light-Current curve of a typical LED

Wall Plug Efficiency

The wall plug efficiency is the total electrical power-in vs optical power-out efficiency. The electrical input power is

$$P_{in} = IV_a \tag{18}$$

where V_a is the applied voltage to the diode. The output power is

$$P_o = IV_g\eta_e. \tag{19}$$

The wall plug efficiency is, therefore

$$\eta_{wp} = \eta_e \frac{V_g}{V}. \tag{20}$$

Assuming a forward applied voltage of 1.8V, we can calculate the wall plug efficiency for the previous example as 0.5%.

Improving the Extraction Efficiency

Since the extraction efficiency is the primary limiting factor of the efficiency of LEDs, much work has been done in improving this factor. We noted that the limitation comes mostly from the escape cone. One method to increasing the escape cone is by increasing the refractive index of the surrounding medium.

However, simply depositing a film of a higher refractive index material will accomplish nothing to improve the escape cone. While the critical angle can be increased at the semiconductor/film interface by using a higher index film, the critical angle at the film/air interface will reduce the angle such that the overall escape cone will remain unaltered. The mathematical proof of this is left as an exercise. On the other hand, a material that is shaped like a dome, as shown in Fig 4, can significantly improve the extraction efficiency. In this case, the critical angle at the semiconductor/dome interface will increase, but the angle of incidence at the dome/air interface will always be normal. Assuming a small emission area and a spherical dome, we can write the extraction efficiency as

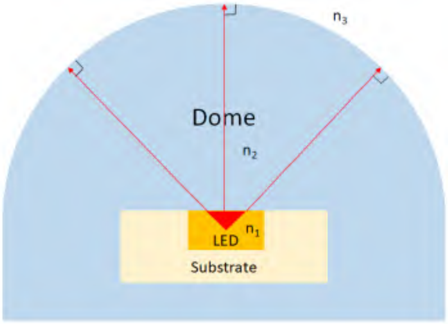

Figure 4: Escape cone improvement by dome encapsulation

$$\eta_{ext} = \underbrace{\frac{1}{2}\left(\frac{n_2}{n_1}\right)^2}_{\text{Escape Cone}} T_1 \ \ T_2 \tag{21}$$

where T_1 is the transmission across the LED/dome interface, and T_2 is across the dome/air interface (at normal incidence). Using a refractive index of $n_1 = 3.5$ for the semiconductor and a dome index of $n_2 = 1.6$, we can get an extraction efficiency of 4.3%, which is three times larger than without the dome.

In most inexpensive LEDs, the dome material is made of a transparent plastic. Although the refractive index is only around 1.6, this is simple to manufacture and yields a 300% improvement in efficiency. Another common technique is to roughen the LED surface. This will create thousands of miniature facets on the LED surface at random angles and allow more light to escape without being reflected. Additionally, the LED may also be mounted on a parabolic cup reflector or coated on the backside to help redirect the photons emerging from the backside of the diode.

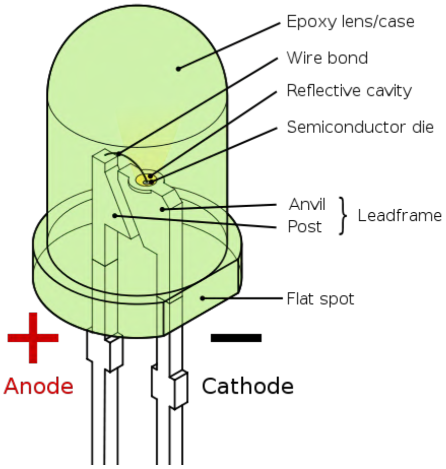

Figure 5: Typical construction of an LED

A dome with an even higher refractive index would be beneficial. High index plastics are very rare, but there are high index dielectrics such as titania or silicon carbide which have refractive indices around 2.5. The refractive index of GaN (which is the substrate used in white LEDs) is also 2.5. Therefore, by using a SiC dome with a reflector and an antireflection coating on the exterior dome surface, in theory at least, we should be able to achieve nearly 100% extraction efficiency from a GaN LED, but may be difficult to manufacture economically.

The substrate thickness is also an important limiting factor. Thicker substrates will re-absorb the emitted photons before they reach the surface. Therefore, high efficiency LEDs typically thin down the substrates in order to reduce this absorption.

Thermal Resistance

Besides the extraction efficiency, the next largest performance-limiting factor is heat extraction. Since the wall plug efficiency of LED is on the order of 1%, the remaining 99% of power is dissipated as heat in the semiconductor material. The internal quantum efficiency, η_i, is a strong function of temperature. Elevated temperatures lead to degradation of η_i due to increased non-radiative recombination processes. This is the primary reason for the reduced responsivity at higher currents in Fig 3.

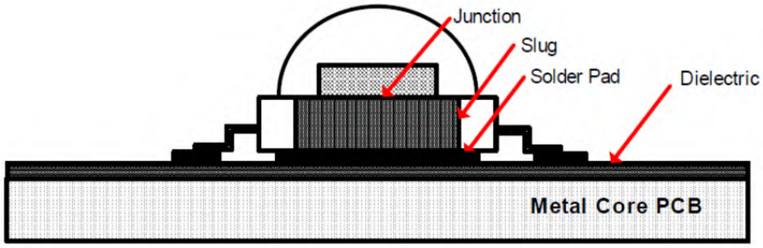

Figure 6: Cross-section of an LED designed for high power. Source: Cypress Semiconductors

While the construction shown in Fig 5 does not employ any special considerations to reduce the junction temperature, Fig 6 shows the cross section of an LED designed for higher power operation. In this case, the thermal resistance between the junction and the ambient becomes the most important factor. The semiconductor is typically mounted such that the junction is closer to the slug (known as flip-chip mounting). The slug has minimal thermal resistance, and is soldered to a metal-core printed circuit board (MC-PCB). The metal-core increases the thermal conductivity of the PCB, vertically as well as horizontally. Designing effective heat dissipation mechanisms is one of the areas of innovations in LED designs.

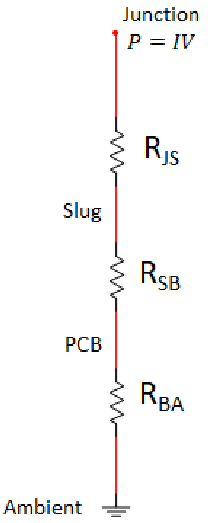

Figure 7: Thermal resistance model of an LED

The temperature of the junction will be determined by the sum of all the thermal resistances between the junction and the ambient (or heat sink). For example, we can represent the thermal resistances as R_{JS} (junction-to-slug), R_{SB} (slug-to-board) and R_{BA} (board-to-ambient), as shown in Fig 7. Additional resistances may be present depending on the exact configuration of the assembly. If we assume nearly all of the power in the LED is dissipated as heat (neglecting the very small fraction that is emitted as photons), then we can calculate the junction temperature as

$$T_J = P \times (R_{JS} + R_{SB} + R_{BA})\,. \tag{22}$$

For example, if $R_{JS} = 5$K/W, $R_{SB} = 5$K/W, $R_{BA} = 10$K/W, the total thermal resistance to ambient is 20K/W. If the LED electrical power is 5W, the junction will be 5×20K warmer than the ambient. If the ambient is $20°$C, the junction temperature will be $120°$C. As noted earlier, the internal quantum efficiency will be lower at this elevated temperature, which is the main reason why the responsivity declines with increasing current. Additionally, the mean time between failure (MTBF) will also increase when the operating temperature is high.

LED Materials

LEDs require photon emission, which implies they can only be made using direct bandgap materials. Most of the LED materials are III-V or II-V semiconductors. The choice largely depend on the emission wavelength and cost. The following a some of the most widely used semiconductors for LEDs:

- **Al$_x$Ga$_{1-x}$As:** GaAs is the most mature III-V semiconductor and is widely used in optoelectronics. Its bandgap is 1.42eV, which corresponds to a wavelength of 870nm. By alloying GaAs with AlAs, it is possible to create a range of Al$_x$Ga$_{1-x}$As compositions. Most importantly, AlAs is lattice matched to GaAs, so very defect-free materials can be created this way, resulting in very high quality (high internal quantum efficiency). The bandgap of Al$_x$Ga$_{1-x}$As ranges from 1.42eV (when $x = 0$) to 2.17eV (when $x = 1.0$), however, the material has a direct bandgap only for $x < 0.45$, which corresponds to a bandgap of 2.02eV, or a wavelength of 615nm. Therefore, this material system is used for producing near-infrared LEDs and red LEDs.

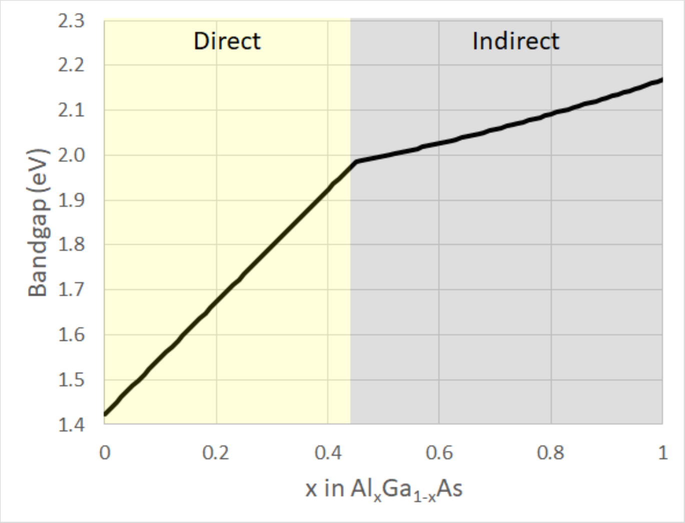

Figure 8: Bandgap of Al$_x$Ga$_{1-x}$As as a function of x

- **GaAs$_{1-x}$P$_x$:** GaAs and GaP can be mixed to create GaAs$_{1-x}$P$_x$ alloys. However, they do not share the same lattice constant, so adding phosphorous to GaAs will generally result in strain and dislocations. Direct bandgap can be maintained only up to $x = 0.5$. The bandgap of GaAs$_{1-x}$P$_x$ with $x = 0.5$ is very similar to Al$_x$Ga$_{1-x}$As with $x = 0.45$. However, adding nitrogen to the mixture allows higher values of x to be used, resulting in a direct bandgap as large 2.25eV. As a result, GaAs$_{1-x}$P$_x$:N can be used to produce emission wavelengths between 870nm ($x = 0$) and 550nm ($x = 1.0$). Therefore, GaAs$_{1-x}$P$_x$:N can access a larger range of wavelengths from near infrared to the yellow.

- **AlGa$_{1-x}$P$_x$:** This is an indirect bandgap material, but doping with nitrogen makes it direct. The range of bandgaps is fairly small, but it is one of the few materials that can emit in the green (~ 530nm).
- **GaIn$_{1-x}$N$_x$:** This is an alloy between GaN and InN. Blue and UV LEDs are produced using this material system. Most white LEDs are actually constructed from a UV LED, whose emission is absorbed and re-emitted using a phosphorescence coating. There are no substrates lattice matched to this material system. Al$_2$O$_3$ (sapphire) substrates are widely used even though it is not perfectly lattice matched.

Fiber Coupling

How effectively light can be coupled into an optical fiber is largely determined by the numerical aperture of the fiber and the radiance (brightness) of the source. High brightness and high numerical aperture is the best case scenario. We can derive the coupling efficiency as follows.

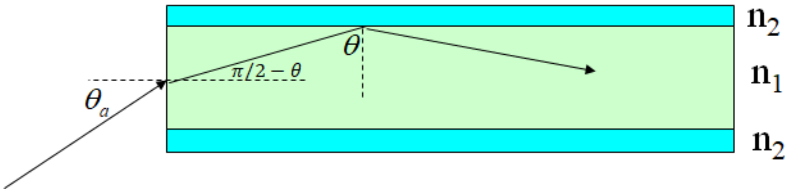

Figure 9: Coupling light into an optical fiber

Consider a fiber as shown in Fig 9, with a core index of n_1 and a clad index of n_2. Assuming a large core multimode fiber, in order to support a guided mode, the angle of incidence at the core/clad interface has to be larger than the critical angle, i.e.,

$$\sin\theta > \sin\theta_{cr} = \frac{n_2}{n_1}. \tag{23}$$

At $\theta = \theta_{cr}$, the angle at the core/air interface will be

$$\frac{\pi}{2} - \theta_{cr} = \frac{\pi}{2} - \sin^{-1}\left(\frac{n_2}{n_1}\right). \tag{24}$$

Applying Snell's law at this interface, we can get the angle on the air side of the interface, θ_a.

$$\begin{aligned}
\sin\theta_a &= n_1 \sin\left(\frac{\pi}{2} - \theta_{cr}\right) && (25)\\
&= n_1 \cos\theta_{cr} && (26)\\
&= n_1\sqrt{1 - \sin^2\theta_{cr}} && (27)\\
&= n_1\sqrt{1 - \left(\frac{n_2}{n_1}\right)^2} && (28)\\
&= \sqrt{n_1^2 - n_2^2} && (29)
\end{aligned}$$

In other words, the maximum incident angle that will be guided by the optical fiber is $\sin^{-1}\left(\sqrt{n_1^2 - n_2^2}\right)$. This angle, $\sin\theta_a$, is also known as the *Numerical Aperture (NA)* of the fiber.

Earlier we assumed that the light emission from the LED junction was Lambertian. The light emission from the LED's escape cone can become modified by the geometry of the dome and additional reflectors. In general, LED manufacturers fit the emission profile to an empirical equation

$$I = I_o \cos^n \theta. \tag{30}$$

We can calculate the fraction of the power coupled into the fiber by representing the LED as a collection of point sources. For each point source, we can calculate the coupling efficiency as:

$$\begin{aligned}
\eta_c &= \frac{\int_0^{\theta_a} (2\pi r \sin\theta)(r d\theta)(I_o \cos^n\theta)}{\int_0^{\pi/2} (2\pi r \sin\theta)(r d\theta)(I_o \cos^n\theta)} && (31)\\
&= \frac{\int_0^{\theta_a} \sin\theta \cos^n\theta \, d\theta}{\int_0^{\pi/2} \sin\theta \cos^n\theta \, d\theta} && (32)\\
&= \frac{\left(\frac{1}{n+1}\right) \cos^{n+1}\theta\big|_0^{\theta_a}}{\left(\frac{1}{n+1}\right) \cos^{n+1}\theta\big|_0^{\pi/2}} && (33)\\
&= \frac{\cos^{n+1}\theta\big|_0^{\theta_a}}{\cos^{n+1}\theta\big|_0^{\pi/2}} && (34)\\
&= 1 - \cos^{n+1}\theta_a. && (35)
\end{aligned}$$

In the case of $n = 1$ (Lambertian), the expression simplifies even further. It becomes

$$\begin{aligned}
\eta_c &= \sin^2\theta_a && (36)\\
&= (\mathrm{NA})^2. && (37)
\end{aligned}$$

The numerical aperture of multimode fibers can range between 0.1 and 0.5, corresponding to an acceptance angle between $6°$ and $30°$. For $n = 1$ this results in a coupling efficiency between 1% and 25%. For $n = 3$, this results in coupling between 2% and 44%. The higher value of n results in a greater coupling because the beam is narrower and more focused, resulting in more energy being contained within the acceptance angle of the fiber. This is shown in Fig 10.

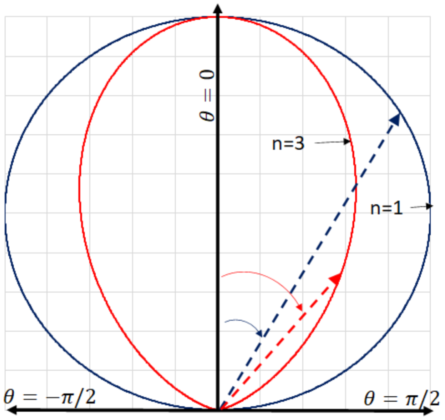

Figure 10: Polar intensity plot of $\cos^3\theta$ and $\cos\theta$.

It should be noted that this result was derived by assuming the emission area of the LED to be small. and the LED to be placed directly against the face of the fiber. If the LED emission area is larger than the fiber core, the overall coupling will be the fractional overlap area between the LED and the fiber core multiplied by the same coupling expression (35). For example, if the LED diameter is D_L and the fiber core diameter is D_f, expression (35) would become

$$\eta_c = \left(1 - \cos^{n+1}\theta_a\right) \quad \text{for } D_L < D_f \tag{38}$$

$$\eta_c = \left(\frac{D_f}{D_L}\right)^2 \left(1 - \cos^{n+1}\theta_a\right) \quad \text{for } D_L > D_f \tag{39}$$

At this point, it is important to review a fundamental principle in electromagnetics known as the Brightness Theorem. Brightness is the intensity contained within a solid angle. The theorem states that the light intensity (power per unit area) contained within a solid angle remains constant. The product of the solid angle and area is known as etendu or optical throughput. The brightness theorem can also described in terms of Lagrange invariant in geometrical optics.

Returning the LED coupling discussion, it is possible, in some cases, to increase the coupling between an LED and a fiber by using a lens. *This works only if the emitting surface is smaller than the fiber core.* A lens can be used to magnify and match the emitting surface onto the fiber core, which effectively modifies the emission profile of the source. Remember that radiance (brightness) is a conserved quantity in an optical system. That means, if the size of the source is magnified by a factor M, then the emission angle from the image will get smaller by the same factor to preserve the brightness. In other words, the emission beam angle will become narrower when a source is magnified. This is illustrated in Fig 11. Assuming an emission profile from the LED of $I = I_o \cos^n \theta$, this will result in a coupling efficiency of

$$\eta_c = \left(1 - \cos^{n+1}(M\theta_a)\right) \quad \text{for } MD_L < D_f \tag{40}$$

$$\eta_c = \left(\frac{D_f}{MD_L}\right)^2 \left(1 - \cos^{n+1}(M\theta_a)\right) \quad \text{for } MD_L > D_f \tag{41}$$

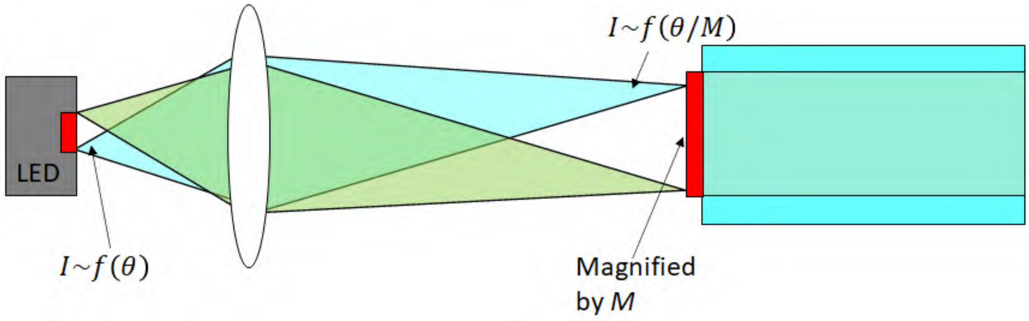

Figure 11: Improving the coupling from an LED to a fiber by magnifying the emission area. This figure illustrates the situation where the magnification is such that $\frac{D_f}{MD_L} = 1$ with $\frac{D_f}{D_L} > 1$.

For example, if the emission area of an LED has a diameter of 25μm with an emission profile $I = I_o \cos^2 \theta$, and fiber has a core diameter of 100μm and a numerical aperture of 0.25 ($\theta_a = \sin^{-1} 0.25 = 14.4°$), the coupling efficiency of $1 - \cos^3 \theta_a = 9.2\%$ can be obtained by simply butt-coupling the LED on the face of the fiber. However, it is possible to use a lens to magnify the emission area by a factor of 4 to match the fiber core diameter. This will result in a reduction in the emission angle by a factor of 4. The emission profile will become $I = I_0 \cos^2(\theta/4)$, where θ corresponds to the emission angle from the unmagnified LED. The coupling efficiency will be $\eta_c = 1 - \cos^3(4\theta_a) = 0.85$, or 85%. Magnifying beyond the size of the fiber will, of course, not result in any improvement. Therefore, this technique only works when the LED emission area is smaller than the fiber core.

If the LED emission area is larger than the fiber, whether or not using a lens will result in any improvement depends on the LED emission profile and the size mismatch. In general, this will not result in a noticeable improvement because any gain in the de-magnification of the area will be offset by the increase in the angular divergence of the image. For example, using the same fiber as above, consider a 200μm LED emission area with a profile $I = I_o \cos \theta$.

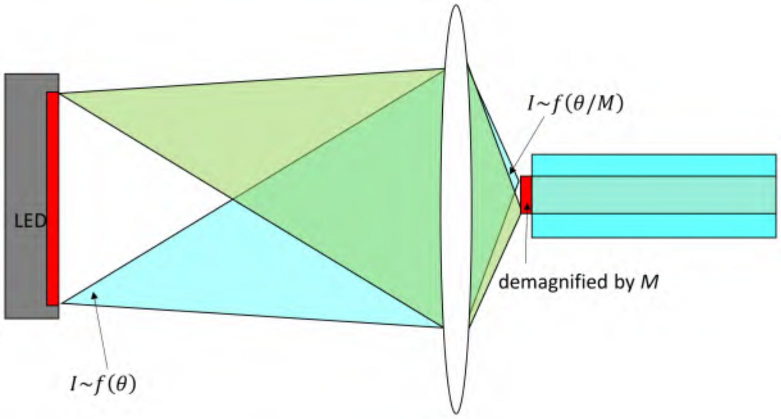

Figure 12: Imaging an LED to a fiber by demagnifying the emission area does not increase coupling. This figure illustrates the situation where the de-magnification is such that $\frac{D_f}{MD_L} = 1$ with $\frac{D_f}{D_L} < 1$

The magnification in this case is $M = 0.5$. Without a lens, the butt-coupled efficiency will be $\eta_c = 0.25\left[1 - \cos^2(\theta_a)\right] = 0.0156$. With the lens, $\eta_c = 1 - \cos^2(0.5\theta_a) = 0.0159$, which is nearly the same as before. Therefore, the added complexity of using a lens is not justified.

Incidentally, we should be able to verify demagnifying a large source can improve the coupling if n is very large. For example, if $n = 50$ in the above example, the butt-coupled efficiency will be $\eta_c = 0.25\left[1 - \cos^{51}(\theta_a)\right] = 0.20$. With demagnification, $\eta_c = 1 - \cos^{51}(0.5\theta_a) = 0.33$, which is significantly higher. This is the main reason why laser beams (which have large values of n) can be focused with a lens to improve coupling into a fiber.

Biasing and Modulating an LED

LEDs are diodes, and just like any diode, they cannot be directly connected to a constant-voltage source. A limiting resistor must be used, or a constant-current source must be used. We know that silicon diodes have a built-in (or barrier) voltage of about 0.7V, but LEDs have a wide range of built-in voltages because they are made from a variety of different semiconducting materials. Wide bandgap materials (shorter emission wavelength) generally have a larger built-in voltage, and narrower bandgap materials (longer wavelength) will have a smaller built-in voltage. This is related to the intrinsic carrier concentration n_i. In wider bandgap materials n_i will be lower, which leads to a higher barrier voltage from equation (10). For example, GaN LEDs (blue) need to be operated near 3.0V while GaAs LEDs (near infrared) need to be operated at 1.6V.

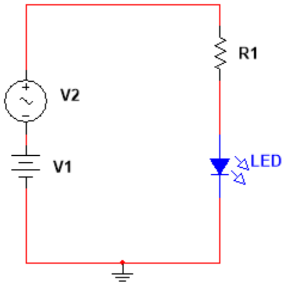

Figure 13: An LED biasing circuit with a current-limiting resistor

A typical circuit for biasing an LED is shown in Fig 13. For example, if the LED has a forward voltage of $Vf = 1.6$V and the desired operating current is $I = 100$mA, and the supply voltage source is $V1 = 5$V, the required series resistance can be calculated as

$$R1 = \frac{V1 - Vf}{I} = 34\Omega. \tag{42}$$

$V2$ is the small signal modulation voltage that is superimposed on the DC bias. To reduce nonlinearity (P_o vs I), the amplitude of the modulation voltage has to be kept fairly small.

Small Signal Current Modulation of LEDs

One of the main difference between the LEDs and other conventional light sources is that LEDs have the capability to be modulated at a reasonably fast rate. LEDs are used in short-range fiber communication systems that do not require very high speeds, and in free-space remote control units. LED lighting has also been combined with modulation to simultaneously provide illumination and communication capability, known as *Lifi* systems.

As noted before, the spontaneous emission in LEDs occur on either side of the junction due to radiative recombinations. But the net carrier decay is due to both radiative and non-radiative recombination. Assuming the carrier profile is confined to a small width on either side of the junction, we can write

$$\frac{dn}{dt} = \frac{I}{qV} - \frac{n - n_s}{\tau} \tag{43}$$

where I is the forward current. V is the approximate volume of the region where the carriers recombine and n_s is the average steady-state carrier concentration. Therefore, n is the *average* carrier concentration in this volume. The subtraction of n_i is not critical because the excess carrier concentrations are typically several orders of magnitude larger than the thermal equilibrium values.

Next, we will assume a small signal modulation on the current. Since we are assuming a linear system, this will produce a corresponding small signal modulation in the carrier density, which will then produce a small signal modulation in the optical power output. This is illustrated in Fig 14. Therefore:

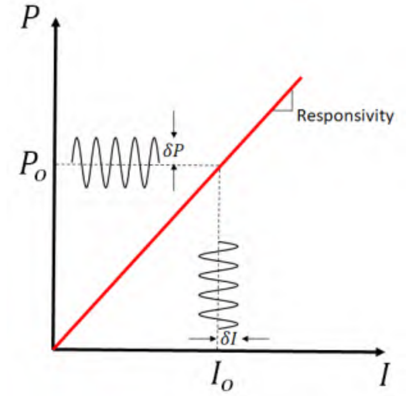

Figure 14: Illustration of a small AC signal impressed on the DC signal.

$$\begin{aligned} I &= I_o + \delta I e^{j\omega t} && (44)\\ P &= P_o + \delta P e^{j\omega t} && (45)\\ n &= n_o + \delta n e^{j\omega t} && (46) \end{aligned}$$

where I_o is the dc bias current of the LED. In other words,

$$0 = \frac{I_o}{qV} - \frac{n_o - n_s}{\tau}. \tag{47}$$

Substituting equations (44) and (46) into equation (43) results in

$$\frac{dn}{dt} = \frac{I_o + \delta I e^{j\omega t}}{qV} - \frac{n_o + \delta n e^{j\omega t} - n_s}{\tau}. \tag{48}$$

Substituting the dc bias condition from equation (47) results in

$$\begin{aligned} j\omega\,\delta n\,e^{j\omega t} &= \frac{\delta I e^{j\omega t}}{qV} - \frac{\delta n e^{j\omega t}}{\tau} && (49) \\ j\omega\,\delta n &= \frac{\delta I}{qV} - \frac{\delta n}{\tau} && (50) \\ \delta n\left(j\omega + \frac{1}{\tau}\right) &= \frac{\delta I}{qV} && (51) \\ \frac{\delta n}{\delta I} &= \frac{1}{qV\left(j\omega + \frac{1}{\tau}\right)}. && (52) \end{aligned}$$

We can manipulate this expression further:

$$\begin{aligned} \frac{\delta n\, qV}{\delta I} &= \frac{1}{\left(j\omega + \frac{1}{\tau}\right)} && (53) \\ &= \frac{\tau}{(1 + j\omega\tau)}. && (54) \end{aligned}$$

Now, multiply both sides by $\frac{E_g}{q}\eta_e$, which is the definition of Resonsivity, as per equation (16), we can get:

$$\begin{aligned} \frac{(\delta n/\tau)\, qV}{\delta I}\frac{E_g}{q}\eta_e &= \frac{1}{(1 + j\omega\tau)}\frac{E_g}{q}\eta_e && (55) \\ &= \frac{R}{(1 + j\omega\tau)}. && (56) \end{aligned}$$

We can interpret the numerator on the left side, $\frac{\delta n V E_g \eta_e}{\tau}$ as the modulated optical power output from the LED. Therefore, we can represent it with the symbol δP. This leads to:

$$\begin{aligned} \frac{\delta P}{\delta I} &= \frac{R}{(1 + j\omega\tau)} && (57) \\ r &= \frac{R}{(1 + j\omega\tau)} && (58) \end{aligned}$$

where we have represented $\frac{\delta P}{\delta I}$ as the AC responsivity, r. As we can see, the modulated power declines from the DC responsivity value as the modulation frequency increases.

In magnitude, we can write equation (58) as

$$|r| = \frac{|R|}{\sqrt{(\omega\tau)^2 + 1}} \tag{59}$$

From this, we can note that the magnitude of the AC responsivity drops to half the value of the DC responsivity at a frequency of

$$f_{3dB} = \frac{\sqrt{3}}{2\pi\tau}. \tag{60}$$

This is designated as the 3dB modulation bandwidth of the LED, and is typically the useful modulation bandwidth of the device. For example, if $\tau = 10$ns, the 3dB modulation bandwidth of the LED will be 27MHz.

The responsivity can also be plotted as a function of frequency, typically on a log-log scale. This allows the knee of the curve to be identified as the 3dB point. An example is shown in Fig 15.

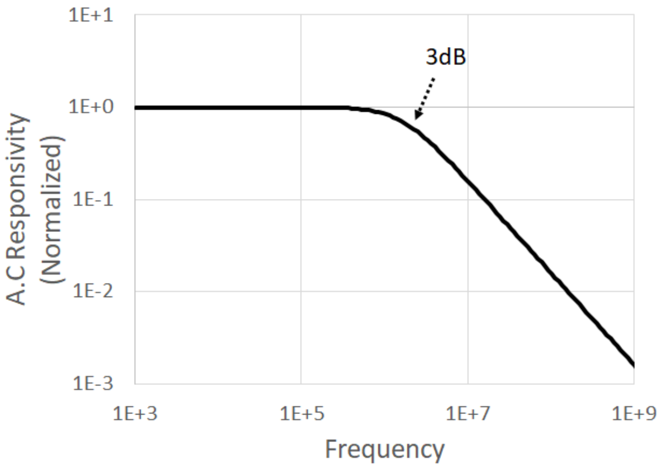

Figure 15: Responsivity vs frequency

Large Signal Current Modulation

The large signal modulation response can be determined by directly solving the rate equation (43). For example, the step-up response of the LED can be determined by integrating the first order differential equation with the appropriate initial conditions. The current starts from zero at $t = 0$ and steps up to a value of I_o and remains there until time t. The carriers will respond according to the following equation:

$$\int_{n_s}^{n} \frac{dn}{I_o\tau - qV\left(n - n_s\right)} = \frac{1}{qV\tau}\int_0^t dt, \tag{61}$$

where V is the volume. This results in

$$n = n_s + \frac{I_o\tau}{qV}\left(1 - e^{-t/\tau}\right). \tag{62}$$

From this, the optical output power can be calculated using

$$P = V\frac{n - n_s}{\tau_r}E_g\eta_{ext}. \tag{63}$$

Substituting for $n - n_s$ from equation (62), we can get

$$\begin{aligned} P &= I_o\left(\frac{\tau}{\tau_r}\right)\left(\frac{E_g}{q}\right)\eta_{ext}\left(1 - e^{-t/\tau}\right) && (64)\\ &= I_o\eta_i V_g\eta_{ext}\left(1 - e^{-t/\tau}\right) && (65)\\ &= I_o\,\mathcal{R}\left(1 - e^{-t/\tau}\right). && (66) \end{aligned}$$

Similarly, the step-down response can be solved by replacing the initial conditions on the integral, resulting in

$$P = I_o\mathcal{R}\,e^{-t/\tau}. \tag{67}$$

To make the numerical solution easier, we can also express the step-up response as

$$\frac{dP}{dt}=\frac{I_o\mathcal{R}-P}{\tau}, \tag{68}$$

and the step-down response as

$$\frac{dP}{dt}=\frac{P}{\tau}. \tag{69}$$

Fig 16 shows the light output from an LED when it is modulated at 27Mb/s using RZ (Return to Zero) encoding. The orange line corresponds to an LED with a lifetime of 10ns (which makes the digital modulation frequency the same as the 3dB frequency, f_{3dB}). The magenta line corresponds to a lifetime of 1ns (which corresponds to a digital modulation frequency of $f_{3dB}/10$).

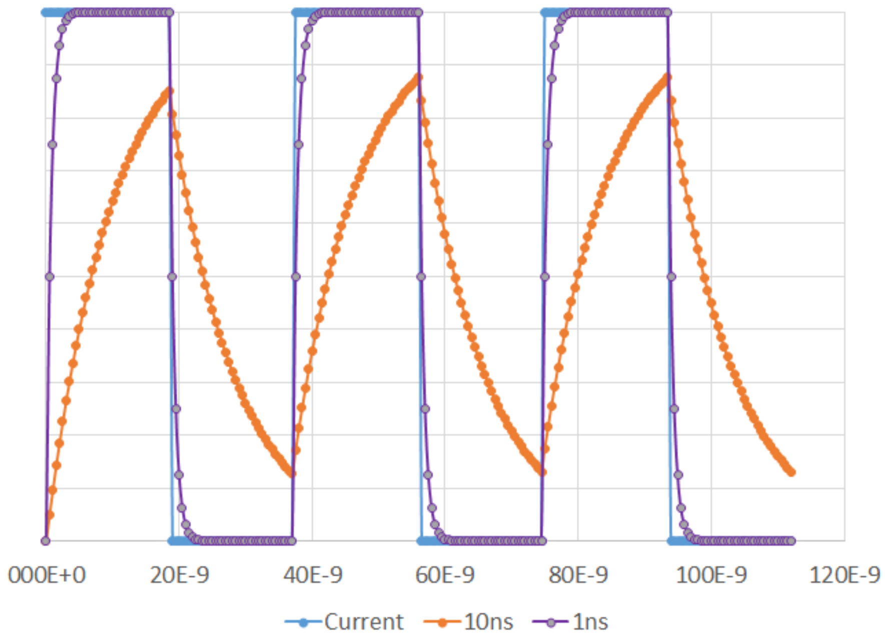

Figure 16: Optical waveform using RZ encoding at 27Mb/s on an LED with lifetimes of 10ns and 1ns.

Spontaneous Emission

The photons emitted in an LED occur as a result of random recombination between electrons and holes. Even though this process has a recombination lifetime τ_r, this lifetime is a statistical average, not a precise value. Furthermore, each recombination is uncorrelated to the previous recombination. As a result, these photons arrive at random times with random directions and orientations. This is known as spontaneous emission. In this section we will derive the spectral shape of an LED emission.

In terms of the band structure of the material, the spontaneous emission rate can be written as

$$\downarrow r_{21}\left(E\right)=\left[\rho_c^{-1}\left(E_2\right)+\rho_v^{-1}\left(E_1\right)\right]^{-1}f\left(E_2\right)\left[1-f\left(E_1\right)\right]\ A_{21}, \tag{70}$$

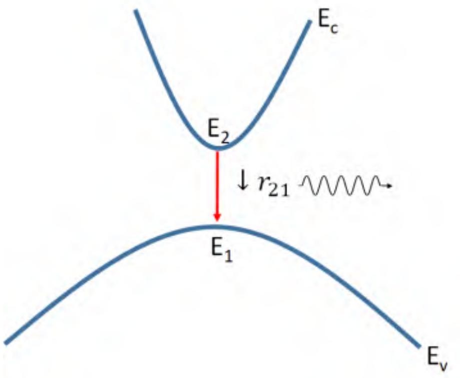

Figure 17: Spontaneous Emission

where

- $\downarrow r_{21}\left(E\right)$ is the number of photons emitted per unit time per unit volume per unit energy.
- $\left[\rho_c^{-1}\left(E_2\right)+\rho_v^{-1}\left(E_1\right)\right]^{-1}$ is the joint density of states, often written as $\rho_J\left(E\right)$, in the units of number of states per unit volume per unit energy.
- $f\left(E_2\right)$ is the Fermi distribution in the conduction band (i.e., the probability of a conduction band state at energy E_2 being occupied).
- $\left[1-f\left(E_1\right)\right]$ is the probability of a valence band state with energy E_1 being vacant.
- E_2-E_1 is the photon energy.
- A_{21} is the transition rate constant.

With no forward bias, the p-side and the n-side will be characterized by a single Fermi level E_F. Under forward bias, we demonstrated that the Fermi level will split into E_{FN} and E_{FP} near the junction, producing an non-equilibrium electron-hole concentration. These are known as quasi-Fermi levels. Because the Fermi levels are different for $f\left(E_1\right)$ and $f\left(E_2\right)$, we need to make a distinction between the two expressions. We will write these as

$$f\left(E_2\right)= \quad f_c\left(E_2\right) \quad = \frac{1}{e^{(E_2-E_{FN})/kT}+1} \tag{71}$$

$$f\left(E_1\right)= \quad f_v\left(E_1\right) \quad = \frac{1}{e^{(E_1-E_{FP})/kT}+1}. \tag{72}$$

Therefore, the emission rate can be written as

$$\downarrow r_{21}\left(E\right)=\left[\rho_c^{-1}\left(E_2\right)+\rho_v^{-1}\left(E_1\right)\right]^{-1}\left[\frac{1}{e^{(E_2-E_{FN})/kT}+1}\right]\left[1-\frac{1}{e^{(E_1-E_{FP})/kT}+1}\right]A_{21}. \tag{73}$$

One problem with equation (73) is that it is in terms of E_2 (energy of the electron in the conduction band) and E_1 (energy of the hole in the valence band), while the photon energy is in terms of $E=E_2-E_1$. Although they are related quantities, it would be simpler to have the entire expression in terms of E rather than E. This can be done by utilizing the fact that the parabolic band structure in the conduction and valence bands are characterized by

$$E_2 = E_c + \frac{\hbar^2 k^2}{2m_c} \tag{74}$$

$$E_1 = E_v - \frac{\hbar^2 k^2}{2m_v} \tag{75}$$

where E_c and E_v are the conduction and valence band edges, and m_c and m_v are the conduction band and valence band effective masses. Subtracting equation (75) from (74) results in

$$E_2 - E_1 = (E_c - E_v) + \frac{\hbar^2 k^2}{2}\left(\frac{1}{m_c} + \frac{1}{m_v}\right). \tag{76}$$

This can be written as

$$E = E_g + \frac{\hbar^2 k^2}{2m_r}, \tag{77}$$

where E_g is the material bandgap, E is the photon energy and m_r is a reduced effective mass defined as

$$\frac{1}{m_r} = \frac{1}{m_c} + \frac{1}{m_v}. \tag{78}$$

We can re-arrange equation (77), to get

$$\hbar^2 k^2 = 2m_r\left(E - E_g\right). \tag{79}$$

We can then substitute equation (79) into equations (74) and (75) allows us to get an expression for E_1 and E_2 in terms of the photon energy E:

$$E_2 = E_c + \left(E - E_g\right)\frac{m_r}{m_c} \tag{80}$$

$$E_1 = E_v - \left(E - E_g\right)\frac{m_r}{m_v}. \tag{81}$$

As a result, we can write the expressions for the Fermi distributions in terms of the photon energy E:

$$f_c\left(E_2\right) = \frac{1}{e^{\left(E_c + (E - E_g)\frac{m_r}{m_c} - E_{FN}\right)/kT} + 1} \tag{82}$$

$$f_v\left(E_1\right) = \frac{1}{e^{\left(E_v - (E - E_g)\frac{m_r}{m_v} - E_{FP}\right)/kT} + 1}. \tag{83}$$

Additionally, the joint density of states can also be simplified in terms of the reduced effective mass m_r. In fact, it has the exact form as the conduction or valence band expressions:

$$\rho_J\left(E\right) = \frac{\sqrt{2}m_r^{3/2}}{\pi^2 \hbar^3}\left(E - E_g\right)^{1/2}, \tag{84}$$

Now, we can express the spontaneous emission entirely in terms of E as the independent variable:

$$\downarrow r_{21}\left(E\right) = \underbrace{\left[\frac{\sqrt{2}m_r^{3/2}}{\pi^2\hbar^3}\left(E-E_g\right)^{1/2}\right]}_{\rho_J(E)} \underbrace{\left[\frac{1}{e^{\left(E_c+(E-E_g)\frac{m_r}{m_c}-E_{FN}\right)/kT}+1}\right]}_{f_c(E_2)} \underbrace{\left[1-\frac{1}{e^{\left(E_v-(E-E_g)\frac{m_r}{m_v}-E_{FP}\right)/kT}+1}\right]}_{1-f_v(E_1)} A_{21}. \tag{85}$$

E_{FN} is quasi-Fermi level on the n-side, and E_{FP} is the quasi-Fermi level on the p-side. At zero bias, the Fermi level will be at the intrinsic value at the metallurgical junction. Representing this Fermi level position as E_i, the quasi-Fermi levels can be written as

$$E_{FN} = E_i + \Delta E_{FN} \tag{86}$$

where ΔE_{FN} is the shift in Fermi level due to the applied bias

$$\Delta E_{FN} = E_{FN} - E_i. \tag{87}$$

Similarly, the p-side quasi-Fermi Level can be written as

$$E_{FP} = E_i - \Delta E_{FP} \tag{88}$$

where

$$\Delta E_{FP} = E_i - E_{FP}. \tag{89}$$

Using this, equation (85) can be written as

$$\downarrow r_{21}\left(E\right) = \underbrace{\left[\frac{\sqrt{2}m_r^{3/2}}{\pi^2\hbar^3}\left(E-E_g\right)^{1/2}\right]}_{\rho_J(E)} \underbrace{\left[\frac{1}{e^{\left(E_c-E_i-\Delta E_{FN}+(E-E_g)\frac{m_r}{m_c}\right)/kT}+1}\right]}_{f_c(E_2)} \underbrace{\left[1-\frac{1}{e^{\left(E_v-E_i+\Delta E_{FP}-(E-E_g)\frac{m_r}{m_v}\right)/kT}+1}\right]}_{1-f_v(E_1)} A_{21}. \tag{90}$$

Finally, we can also make the assumption that the applied voltage V_a contributes equally in moving the quasi-Fermi levels. That is,

$$\Delta E_{FN} = \Delta E_{FP} = \frac{qV_a}{2}, \tag{91}$$

which allows us to get the final expression for spontaneous emission:

$$\downarrow r_{21}\left(E\right) = \underbrace{\left[\frac{\sqrt{2}m_r^{3/2}}{\pi^2\hbar^3}\left(E-E_g\right)^{1/2}\right]}_{\rho_J(E)} \underbrace{\left[\frac{1}{e^{\left(E_c-E_i-qV_a/2+(E-E_g)\frac{m_r}{m_c}\right)/kT}+1}\right]}_{f_c(E_2)} \underbrace{\left[1-\frac{1}{e^{\left(E_v-E_i+qV_a/2-(E-E_g)\frac{m_r}{m_v}\right)/kT}+1}\right]}_{1-f_v(E_1)} A_{21}. \tag{92}$$

Since A_{21} is a constant, we can disregard it for now and plot the emission spectrum on a relative scale. Using GaAs as the LED material, with $E_g = 1.42$eV, $m_c = 0.067m_0$, $m_v = 0.5m_0$, $T = 300$K, we can get the emission spectrum as depicted in 18.

We can clearly see that the emission starts at the bandgap wavelength ($\lambda_g = 1.24/1.42 = 0.873\mu$m, and extends to shorter wavelengths. The bias voltage has a pronounced effect on the amplitude of the emission. As a matter of fact, even zero bias voltage will produce an emission profile. In Fig 18, this emission has been subtracted out, so what appears is actually the excess emission. This may seem a bit odd, but the presence of emission at zero bias is a result of thermal excitation. Even at zero bias, carriers are continuously being excited from valence band to the conduction band, resulting in a steady-state background recombination.

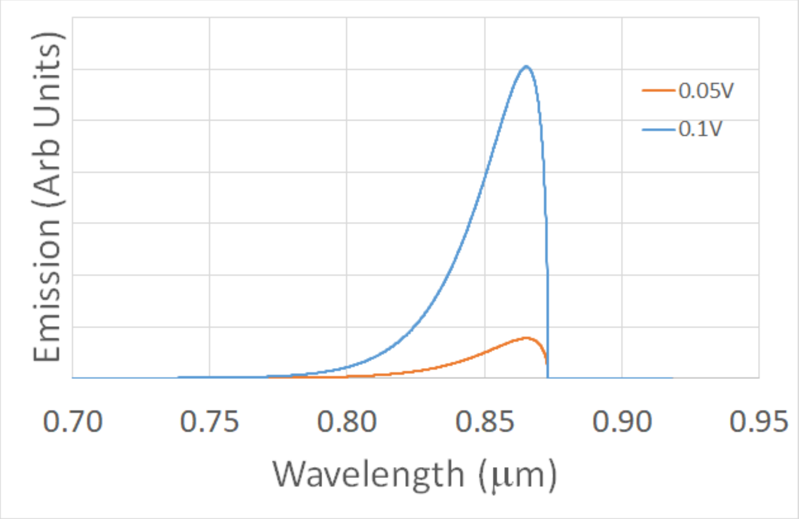

Figure 18: Calculated Spontaneous Emission Spectrum for a GaAs LED

Homework 4

1. Consider a GaAs LED with a dome-shaped epoxy encapsulation (refractive index=1.6). The internal quantum efficiency is 50%. Estimate the responsivity.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val etaInt = 0.5
    val nS = 3.5
    val nEpoxy = 1.6
    val nA = 1.0
    val Vg = 1.42
    val etaExt = (nEpoxy/nS).pow(2)*0.5*(1.0 −((nS−nEpoxy)/(nS+nEpoxy)).pow(2))*(1.0
    − ((nEpoxy−nA)/(nEpoxy+nA)).pow(2))
    println(”eta_ext = ${”%.3f”.format(etaExt)}”)
    val etaE = etaInt * etaExt
    val R = Vg*etaE
    println(”R = ${”%.3f”.format(R*1000.0)} mW/A”)
}

>>eta_ext = 0.085
>>R = 60.489 mW/A
```

2. In an attempt to increase the extraction efficiency from a GaAs LED, it was coated with a 1μm thick film of refractive index 2.2. Using ray angles, show that the extraction efficiency will not be improved by this coating.

   See Fig 19.

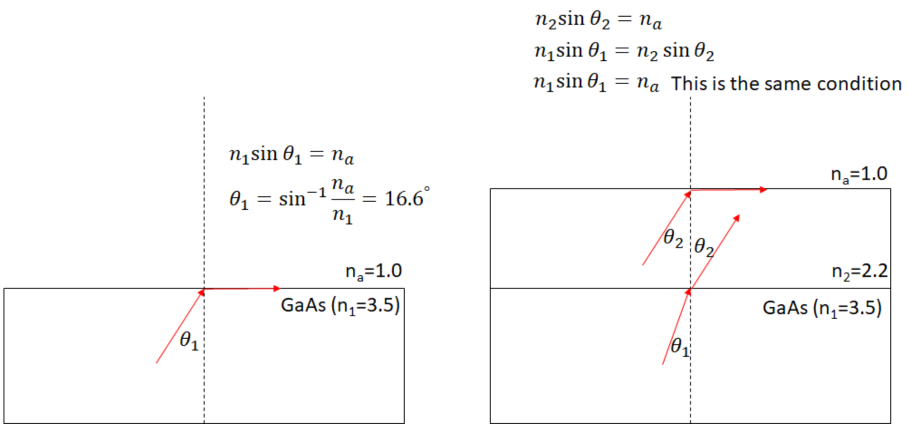

Figure 19: Comparison of escape cone with and without a film

3. Consider an LED with a 500μm-diameter emitting aperture and an angular dependence of intensity that is proportional to $\cos^2\theta$. A 50μm-diameter optical fiber with a core refractive index of 1.46 and a numerical aperture of 0.3 is used to collect the LED emission. Calculate the fraction of LED optical power that is coupled into the fiber. Can a lens be used to collimate this LED to improve coupling?

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val LED = 500.0
    val fiberD = 50.0
    val NA = 0.3
    val n = 2
    val thetaA = asin(NA)

    val etaC = (fiberD/LED).pow(2) * (1.0-cos(thetaA).pow(n+1))
    println(”Coupling without lens = ${”%.5f”.format(etaC)}”)
}

Coupling without lens = 0.00132
```

4. A miniature LED chip is aligned to a 0.25 NA optical fiber using a ball lens. The fiber core diameter is 50μm. The led emission aperture is 10μm. The intensity distribution of the LED chip is $\cos\theta$. What is the highest coupling that could be achieved with this setup?

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val LED = 10.0
    val fiberD = 50.0
    val NA = 0.25
    val n = 1
    val thetaA = asin(NA)
    val M = fiberD/LED
    val etaC = 1.0-cos(M*thetaA).pow(n+1)
    println(”Coupling with lens = ${”%.3f”.format(etaC)}”)
}

>>Coupling with lens = 0.908
```

5. The internal quantum efficiency of an LED is 0.5. The radiative recombination lifetime is 10ns. What is the 3dB modulation bandwidth of the LED?
   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val etaI = 0.5
    val tauR = 10.0
    val tauNR = tauR/(1.0/etaI -1.0)

    println(”tau_nr = ${”%.3f”.format(tauNR)}”)
    val tau = 1.0/(1.0/tauR + 1.0/tauNR)
    println(”tau = ${”%.3f”.format(tau)}”)
    val dB3 = 3.0.pow(0.5)/(2.0*PI*tau)
    println(”3dB = ${”%.3f”.format(dB3*1000.0)} MHz”)
}

>>tau_nr = 10.000
>>tau = 5.000
>>3dB = 55.133 MHz
```

Luminosity & Photometrics

When used for visible illumination applications, the relevant quantity is not radiative power, but *luminous* power. The spectral response of the human eye greatly influences the perception of light power. Human eye response peaks at a wavelength of 555nm, and falls off to nearly zero at 420nm in the blue, and at 700nm in the red. We can consider the human eye as an optical filter. The transmittance of this filter, normalized to a maximum value of 1.0 at the peak wavelength of 555nm, is referred to as the luminosity function, $V(\lambda)$.

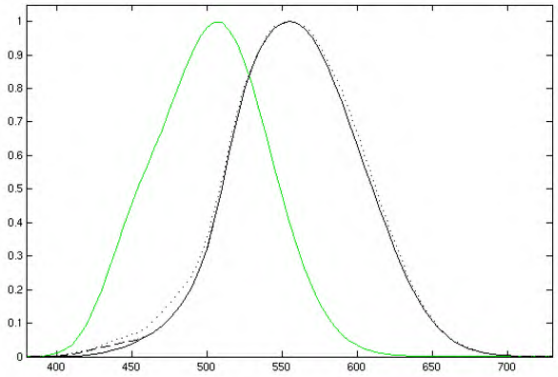

Figure 1: Luminosity function: black is for daytime vision, and green is for nighttime vision. Source: Wikipedia

An approximate expression for the luminosity function can be written empirically by

$$V(\lambda) = 1.019e^{-285.4(\lambda-0.559)^2}, \tag{1}$$

where λ is the wavelength in μm. The luminosity function varies from person to person, and also from daytime to nighttime. Under bright illumination, the peak response of the eye is at 555nm. Under dimly lit conditions, different parts of the eye (the rod cells) become more active, whose luminosity function is blue-shifted by about 50nm. These functions are shown in Fig 1. The peak value of the nighttime function is about three times larger (greater sensitivity) than the daytime function, but they are shown normalized in this plot.

All radiometric quantities have equivalent photometric counterparts. Before discussing these, it is useful to briefly review the radiometric quantities:

- Radiant power (usually in watts) is the total electromagnetic power emitted by a source in all directions.
- Irradiance (watts/m^2) is the electromagnetic power that is received on a unit surface area of the target.
- Radiant intensity (watts/sr) is the electromagnetic power emitted by the source within one unit of solid angle.
- Radiance (watts/sr/m^2) is the radiant intensity divided by the surface area of the source. This is what is generally known as *brightness*. Radiance is an invariant quantity in an optical system.
- Radiant efficiency (wall plug efficiency) is the ratio of radiant power emitted by the source divided by the electrical power supplied to that source.

When these radiometric quantities are filtered by the luminosity function, we get their equivalent photometric quantities.

- Photometric power (luminous power or luminous flux) is the radiometric power filtered by the luminosity function. It is usually measured in Lumens. This is the radiative power perceived by the human eye. One watt of optical power at a monochromatic wavelength of 555nm is defined as being equivalent to 638 Lumens. Therefore, if $\phi_e(\lambda)$ is the radiative spectral emission (in watts/nm), the number of Lumens in that emission can be calculated by integrating the product of spectral emission and the luminosity function normalized to a peak value of 1.0, and then multiplying by 638 Lumens/Watt:

$$\Phi_v = 638 \int_0^\infty V(\lambda)\,\phi_e(\lambda)\,d\lambda. \tag{2}$$

- The photometric equivalent of irradiance (watts received per unit area) is *illuminance*. It is measured in Lux, where one Lux is defined as one lumen per square meter.
- Photometric intensity (luminous intensity) is measured in number of lumens per unit solid angle. One lumen per steradian is defined as a candela.
- Photometric radiance (luminance or brightness) is measured in candelas per square meter.
- Luminous efficiency is the number of lumens emitted by a source divided by the electrical power supplied to that source, expressed in lumens per watt.
- There is also an alternative definition, known as *luminous efficacy*, K. This is the number of lumens in one unit of radiant power (instead of electrical power). This is calculated as

$$K = \frac{\Phi_v}{\Phi_e} = \frac{638\int_0^\infty V(\lambda)\,\phi_e(\lambda)\,d\lambda}{\int_0^\infty \phi_e(\lambda)\,d\lambda}. \tag{3}$$

A summary of these quantities is shown in table 1.

Radiometric	Units	**Photometric**	**Units**	**Conversion**
Radiant Power	W	Luminous Flux	Lumens	1W@555nm = 638 Lumens
Irradiance	W/m^2	Illuminance	Lux	Lux = Lumen/m^2
Radiant Intensity	W/sr	Luminous Intensity	Candela	Candela = Lumens/sr
Radiance (Brightness)	W/sr/m^2	Luminance	Candela/m^2	
Radiant Efficiency	W (radiation)/W (electrical)	Luminous Efficiency	Lumens/W (electrical)	
		Luminous Efficacy	Lumens/W (Radiant)	

Table 1: Summary radiometric and photometric quantities

Table 2 lists the luminous efficiency of common types of lamps. Incandescent lamps have the lowest number of lumens per electrical watts, while sodium vapor lamps have the highest. White LEDs fall somewhere between compact fluorescent lamps and halogen lamps.

For example, a standard 60W incandescent light bulb will produce about 800 lumens, whereas the same 800 lumens can be achieved with a 12W LED lamp, or a 14W compact fluorescent lamp. A T8 linear fluorescent lamp (16W) will produce 1600 lumens.

Luminous Efficiency of Thermal Radiation

Every object whose temperature is greater than 0 K will emit photons. The intensity and wavelength of these photons will be a strong function of temperature, as well as its surface properties. The latter is characterized by a dimensionless constant known as emissivity, ϵ. The

Lamp Type	Luminous Efficiency (Lumens/Watt)
Incandescent	8-18
Halogen	20-30
CFL	50-75
White LED	100-200
Linear Fluorescent	80-110
Sodium Vapor Discharge	100-200

Table 2: Luminous efficiency of common lamps

emission spectrum from a point source is described by Plank's law, which states

$$\phi_e^p(\lambda) = \epsilon \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1} \tag{4}$$

where $\phi_e^p(\lambda)$ is the spectral radiance (power per unit area per unit solid angle per unit wavelength). This expression does not have any angular dependence (i.e. it is independent of Ω and θ) because a point source will emit equally in all directions. A collection of such points on a surface, however, will have an angular distribution, which follows Lambert's cosine law. As discussed earlier, this arises directly as a result of the invariance of radiance. Consequently, the spectral radiance of a surface will have an additional $\cos\theta$ term:

$$\phi_e^s(\lambda, \theta) = \phi_e^p(\lambda) \cos\theta. \tag{5}$$

Following the same procedure as with the LED extraction efficiency, we can calculate the power emitted by one face of a surface per unit area per unit wavelength by performing the integral in spherical coordinates:

$$\begin{aligned} \phi_e^s(\lambda) &= \int_0^{2\pi} \int_0^{\pi/2} \phi_e^s(\lambda, \theta) \sin\theta \, d\theta \, d\phi && (6) \\ &= \int_0^{2\pi} \int_0^{\pi/2} \phi_e^p(\lambda) \cos\theta \sin\theta \, d\theta \, d\phi && (7) \\ &= 2\pi \phi_e^p(\lambda) \left. \frac{\sin^2\theta}{2} \right|_0^{\pi/2} && (8) \\ &= \pi \phi_e^p(\lambda) && (9) \end{aligned}$$

The total radiant power emitted by the surface over all wavelengths becomes:

$$\Phi_e = \int_0^\infty \phi_e^s(\lambda) \, d\lambda. \tag{10}$$

The photometric power can be obtained by applying the luminosity function to the radiant power and multiplying by 683 Lumens/Watt:

$$\Phi_v = 683 \int_0^\infty V(\lambda) \, \phi_e^s(\lambda) \, d\lambda. \tag{11}$$

The value of emissivity ϵ in equation (4) can vary between 0 and 1. Highly reflective surfaces have low values of ϵ, whereas low-reflective or textured surfaces have values closer to 1.0. For example, reflective aluminum has a value of 0.05, whereas paper, skin and other textured surfaces, including smooth surfaces such as glass have values around 0.95. An ideal blackbody source will have $\epsilon = 1.0$.

Luminous efficacy is the ratio between Φ_v and Φ_e. This is the visible fraction of the radiative power. Fig 2 shows the black body spectrum of the sun (5778K) assuming $\epsilon = 1.0$, using the luminosity function corresponding to day time vision. The integrated power Φ_e using equation (10) is 63.1MW/m^2. This is the power density at the surface of the sun. The luminous efficacy (the ratio between Φ_v and Φ_e) is 92 Lumens/Watt. The actual efficacy value on earth may vary somewhat due to solar activities, atmospheric absorption, Rayleigh scattering, etc.. but it is in the range of 80-100 Lumens/Watt.

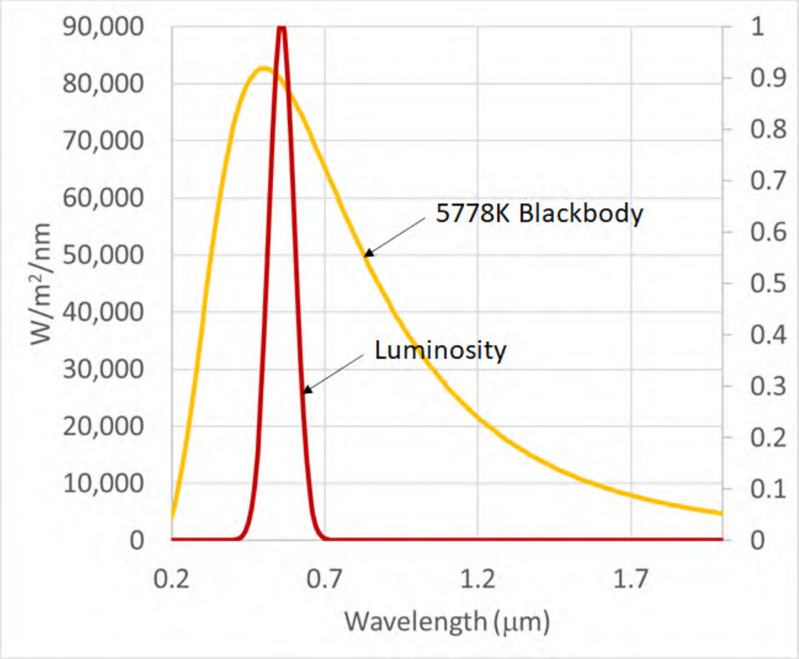

Figure 2: Black body spectrum of the sun and the luminosity function

Homework 5

1. Considering a tungsten halogen lamp as an ideal black body source with a filament temperature of 3000K, calculate the luminous efficacy (lumens per watt) of this light bulb. How does it compare with the efficacy of solar illumination?

   Run this code

```
import kotlin.math.*
//Andrew Sarangan
//Repeat the same calculation as in notes with T=3000K

fun main() {
    val c = 3.0e8
    val k = 1.38e-23
    val h = 6.62607e-34
    val epsilon = 1.0
    val start = 0.1e-6
    val end = 5.0e-6
    val dlambda = 1.0e-10
    val T = 3000.0

    val wavelengths = DoubleArray(((end-start)/dlambda).toInt()){start + it*dlambda}
    val phi_es = wavelengths.map{
                  epsilon*PI*2.0*h*c.pow(2)/it.pow(5)/(exp(h*c/(it*k*T))-1.0)
                  }.toDoubleArray()
    val PHIe = phi_es.sum()*dlambda
    val V = wavelengths.map{
                1.019*exp(-285.4*(it*1.0e6-0.559).pow(2))}.toDoubleArray()
    val PHIv = 683.0*phi_es.zip(V){a,b -> a*b}.sum()*dlambda
    println("${"%.3f".format(PHIv/PHIe)} Lumens/W")

//     wavelengths.forEachIndexed{ i,v ->
//         println("${v}\t${phi_es[i]}")
//     }
}

>>21.081 Lumens/W
```

Semiconductor Diode Lasers

Basic Description

A semiconductor diode laser is an oscillator, just like a microwave or RF oscillator. To build an oscillator, we need two components

- An amplifier;
- A feedback mechanism.

Schematically, we can represent an oscillator connecting the output of the amplifier back to the input side via a feedback filter, as depicted in Fig 1. The function of this circuit can be derived as follows. At the summing junction on the input side, we can write

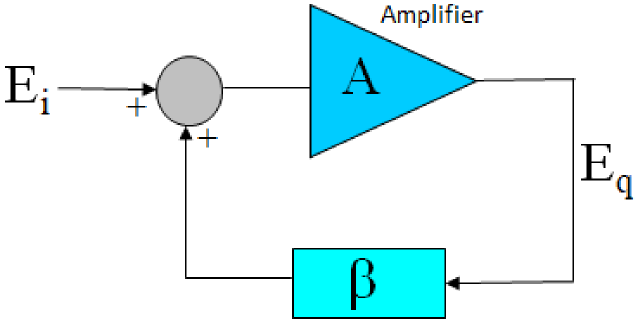

Figure 1: A generic oscillator configuration using an amplifier and a feedback.

$$E_i + E_q\beta = \frac{E_q}{A}. \tag{1}$$

where β is the feedback. From this, we can get

$$\begin{aligned} E_i &= \left(\frac{1}{A} - \beta\right) E_q, && (2)\\ \frac{E_q}{E_i} &= \frac{A}{1-\beta A}. && (3)\end{aligned}$$

The expression for $\frac{E_q}{E_i}$ represents the ratio between the circulating field in the circuit and the input field. We can see that this ratio becomes very large when $\beta A \to 1$. In other words, when the round trip factor (also known as the loop gain) becomes equal to one, we can get a very large circulating field for a very small input signal. However, in general, A and β will have some frequency dependence. This means $\beta A \to 1$ will be satisfied only at specific frequencies. These are the oscillation frequencies of this oscillator. Since the ratio is extremely large at these frequencies, the circulating field inside the cavity will be finite even with a nearly zero input (which actually arises from noise sources in the environment). Hence the circuit functions as a self-sustaining oscillator.

An optical oscillator (laser) is built using

- An optical amplifier with a gain of $A(\omega)$, which can be obtained from semiconductors as well as in certain gas plasmas and solids;
- A feedback mechanism, which is normally realized using mirrors or other types of reflectors.

Consider an optical plane wave $E_q e^{-jkz}$ where k is the wave vector given by $k = \frac{2\pi n_{\text{eff}}}{\lambda}$, n_{eff} is the effective index of the medium and λ is the wavelength. If this wave is traveling between two mirrors with reflectivity r and a distance L, as depicted in Fig 2, we can trace the plane wave from some point between the mirrors towards the right mirror, and then the left mirror, and returning to the original point. If the original field had an amplitude of E_q, by the time the field completes one round-trip, it will have an amplitude of $r^2 E_q$ and a phase of e^{-2jkL}. Therefore, we can say that the round trip factor $\beta(\omega)A(\omega)$ will be equal to $r^2 e^{-2jkL}$. Since $|r| < 1$, the round trip factor will be $|\beta(\omega)A(\omega)| < 1$. As a result, there will be no oscillations possible.

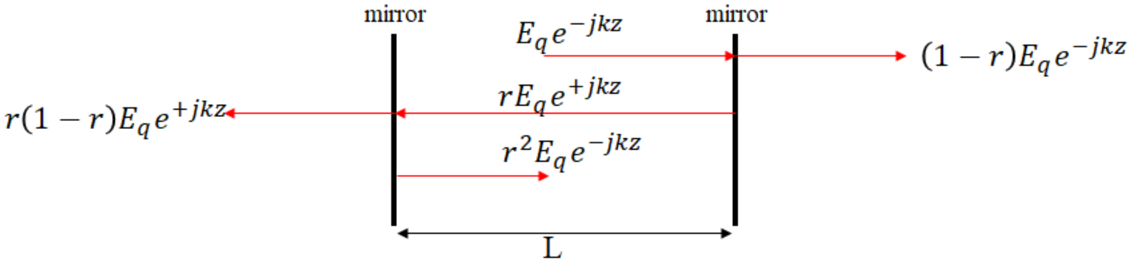

Figure 2: A simple optical cavity between two mirrors

Next, consider an amplifier placed between the mirrors such that the field grows exponentially as it travels through this amplifier. This is shown in Fig 3. We will assume that the amplifier has the same phase value of k except for the addition field gain coefficient γ. As a result, the round trip factor $\beta(\omega)A(\omega)$ will be $r^2 e^{-2jkL} e^{2\gamma L}$. We should be able to see that the round trip factor can now become equal to 1, and allow oscillation to take place.

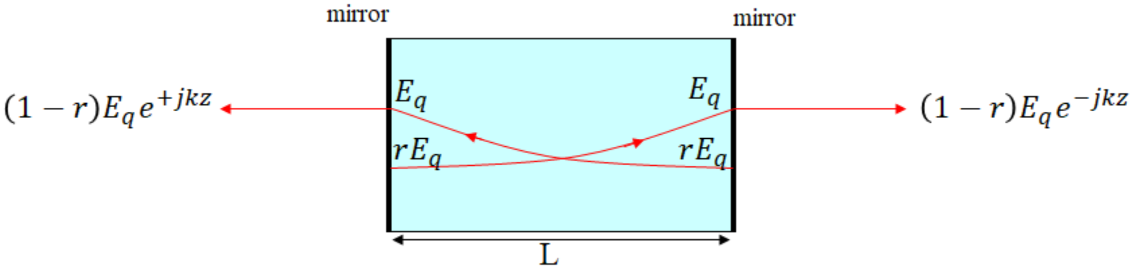

Figure 3: Same optical cavity as in Fig 2 with an amplifier.

In order for the round trip factor to be 1.0, two aspects must be satisfied:

- The amplitude of the round trip factor must be equal to 1.0;
- The phase of the round trip factor must be in multiples of 2π.

We will examine each of these separately.

Amplitude

The amplitude condition $|\beta A|$ can be expressed as

$$\left| r^2 e^{-2jkL} e^{2\gamma L} \right| = 1. \tag{4}$$

Assuming the reflectivity r is real, this becomes:

$$r^2 e^{2\gamma L} = 1, \tag{5}$$

from which we can get

$$\gamma = \frac{1}{2L} \ln\left(\frac{1}{r^2}\right). \tag{6}$$

This is known as the threshold gain condition of the laser, and if often represented by the symbol γ_{th}. We can see that the threshold gain is inversely related to the length of the amplifier (i.e., a higher gain is required from a shorter amplifier). The threshold gain is also inversely related to the reflectivity (i.e., a higher gain is required if the reflectivity of the mirrors are low).

The threshold gain is often expressed in terms of optical power (rather than field amplitudes). Since power scales as the square of the field amplitudes, the threshold power gain will be

$$2\gamma_{th} = \frac{1}{2L} \ln\left(\frac{1}{r^4}\right), \tag{7}$$

$$G_{th} = \frac{1}{2L} \ln\left(\frac{1}{R^2}\right) \tag{8}$$

where $R = r^2$ is the reflectivity of optical power, whereas r is reflectivity of optical fields.

Phase

The phase condition of the round trip factor can be expressed as

$$e^{-2jkL} = e^{-2jN\pi} \tag{9}$$

where N is an integer. This can be further manipulated as

$$2kL = 2N\pi \tag{10}$$

$$k = \frac{N\pi}{L} \tag{11}$$

$$\lambda = \frac{2n_{\text{eff}}L}{N}. \tag{12}$$

From this, we can conclude that the oscillation will be at discrete wavelengths corresponding to integer values of N.

Typically, a number of wavelengths can simultaneously oscillate, resulting in what is known as multi-longitudinal-mode lasing. This does not mean an infinite number of lines will oscillate. Only a handful of lines near the spectral gain peak of the semiconductor material will oscillate. Other lines farther from the gain peak will not reach the threshold gain, and will not oscillate. This is illustrated in Fig 4.

This means that the value of N cannot be any integer. For example, a 1.5μm diode laser will have its amplifier gain limited to a small range around 1.5μm. If $L = 1$mm and $n_{\text{eff}} = 3.5$, we can find that N will be around 4600.

Strictly speaking, only one or two laser lines will be at or near the peak of the gain spectrum. This is due to the shape of the gain spectrum. This situation is illustrated in Fig 5. In practice, however, the gain spectrum will never be perfectly static. Small fluctuations in the forward current will produce small fluctuations in the carrier density, which will shift the gain peak to the left and right. As a result, on average, there will be many more laser lines that will oscillate. The number of lasing lines will depend on the cavity length L, current and temperature stability of the system.

Fig 6 shows the actual emission spectrum from a Fabry-Perot laser, designed for operation near a wavelength of 1600nm. We can clearly identify the shape of the gain spectrum as well as the Fabry-Perot lasing peaks. As we will see in the discussion of gain in semiconductors, the location of the gain peak is a function of temperature as well as the current density. Therefore, the wavelength of Fabry-Perot lasers will never be stable. There are other types of laser cavities where the lasing wavelength can be designed to be relatively more stable (such as DFB, DBR and VCSELs).

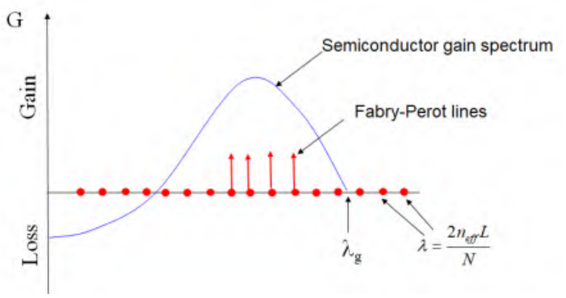

Figure 4: Fabry-Perot oscillations near the gain peak of the semiconductor

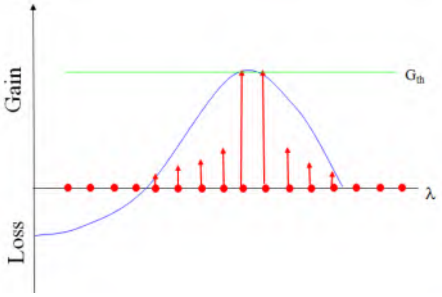

Figure 5: Only one or two Fabry-Perot lines will reach threshold

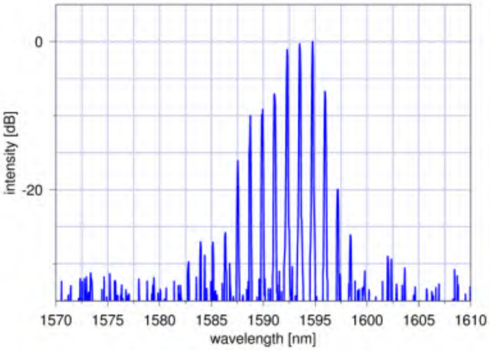

Figure 6: Measured spectrum of a Fabry-Perot laser. Source: Nanoplus, Germany

Fabry-Perot Line Spacing

The type of a laser cavity consisting of two mirrors enclosing a gain region is known as a Fabry-Perot Cavity. This is the simplest type of laser cavity. Other types of laser cavities will be explored later in this document.

We saw that the emission wavelengths of a Fabry-Perot cavity occur at discrete values of N. Although we can calculate the spacing between these wavelengths by considering adjacent integer values of N, the material dispersion also plays an important role in this calculation. We can derive an expression for the wavelength spacing $\Delta\lambda$ as follows:

We can re-write equation (12) as

$$N = \frac{2n_{\text{eff}}L}{\lambda}. \tag{13}$$

We can now find

$$\frac{\Delta N}{\Delta\lambda} = -\frac{2n_{\text{eff}}L}{\lambda^2} + \frac{2L}{\lambda}\frac{dn_{\text{eff}}}{d\lambda}. \tag{14}$$

From this, we can get

$$\Delta\lambda = \frac{\Delta N}{-\frac{2n_{\text{eff}}L}{\lambda^2} + \frac{2L}{\lambda}\frac{dn_{\text{eff}}}{d\lambda}}. \tag{15}$$

Since the smallest increment in wavelength $\Delta\lambda$ will be due to $\Delta N = -1$, the above equation becomes:

$$\begin{aligned} \Delta\lambda &= \frac{-1}{-\frac{2n_{\text{eff}}L}{\lambda^2} + \frac{2L}{\lambda}\frac{dn_{\text{eff}}}{d\lambda}} && (16) \\ &= \frac{\lambda^2}{2n_{\text{eff}}L}\left(1 - \frac{\lambda}{n_{\text{eff}}}\frac{dn_{\text{eff}}}{d\lambda}.\right). && (17) \end{aligned}$$

This can also be expressed as

$$\Delta\lambda = \frac{\lambda^2}{2n_gL} \tag{18}$$

where n_g is the group index of refraction defined as

$$n_g = n_{\text{eff}}\left(1 - \frac{\lambda}{n_{\text{eff}}}\frac{dn_{\text{eff}}}{d\lambda}.\right)^{-1}. \tag{19}$$

In terms of frequency, this can also be written as

$$\Delta\nu = \frac{c}{n_gL}. \tag{20}$$

The group index n_g can be very different than n_{eff}. The wavelength spacing between emission lines is, therefore, determined by the laser cavity length and the group refractive index. This is also known as the free spectral range (FSR). For example, the value of n_{eff} in most semiconductor materials is about 3.5, whereas n_g can be as large as 4.5. The resulting value for $\Delta\lambda$ would be significantly different had we neglected the effects of dispersion, viz., $\frac{dn_{\text{eff}}}{d\lambda}$.

Cavity Losses & Efficiency

If the mirror reflectivities are different, the expression for threshold gain (power) will become

$$G_{th} = \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right), \tag{21}$$

where R_1 and R_2 are the reflection from the two mirrors.

We normally write

$$\alpha_m = \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right), \tag{22}$$

where α_m represents the ”losses” associated with the mirrors, and is referred to as the mirror loss term. This is the equivalent attenuation per unit length along the laser cavity due to the mirror reflectivities being less than unity. One can also view this as arising due to the transmission through the mirrors (considering the transmitted power as being ”lost”).

We can also separate α_m into α_{m1} and α_{m2} associated with each mirror. α_{m1} is the mirror loss due to the first mirror, and α_{m2} is the mirror loss due to the second mirror. Therefore,

$$\alpha_m = \alpha_{m1} + \alpha_{m2}, \tag{23}$$

where

$$\alpha_{m1} = \frac{1}{2L} \ln\left(\frac{1}{R_1}\right) \tag{24}$$

$$\alpha_{m2} = \frac{1}{2L} \ln\left(\frac{1}{R_2}\right). \tag{25}$$

All semiconductor lasers have other losses in addition to mirror losses, such as scattering losses α_s and absorption losses α_a. These are treated exactly like the mirror losses. We can lump all of these losses into a single parameter known as the cavity loss α_c, which becomes:

$$\alpha_c = \alpha_{m1} + \alpha_{m2} + \alpha_s + \alpha_a. \tag{26}$$

For laser oscillation to occur, threshold gain has to overcome all of these losses, not just the mirror losses. Here we are assuming that α_s and α_a exist only within the semiconductor length L, and not over the entire cavity length L. In practice, however, this is a moot point because in nearly all semiconductor lasers the cavity length L is exactly equal to the gain length L. Now, equation (21) can be modified to include these additional losses, such as

$$G_{th} = \alpha_{m1} + \alpha_{m2} + \alpha_s + \alpha_a \tag{27}$$

$$= \alpha_c. \tag{28}$$

Normally, the transmission through one mirror is considered the useful output from the laser, and the transmission through the other mirror is discarded. Naturally, to avoid unnecessary losses, the output mirror is carefully chosen to optimize the laser performance, whereas the other mirror (backside mirror) is chosen to have a very high reflectivity to prevent unnecessary transmission losses.

The efficiency of a laser is the ratio between the ”useful” losses (extracted photons) and the total losses in the laser cavity. Considering mirror #1 as the output mirror (useful loss), we can write the extraction efficiency as

$$\eta_{ext} = \frac{\alpha_{m1}}{\alpha_c}. \tag{29}$$

The interpretation of the extraction efficiency η_{ext} is identical to light emitting diodes. If we express equations (28) and (29) as

$$G_{th} = \frac{1}{2L}\ln\left(\frac{1}{R_1R_2}\right) + \alpha_s + \alpha_a, \tag{30}$$

$$\eta_{ext} = \frac{\frac{1}{2L}\ln\left(\frac{1}{R_1}\right)}{\frac{1}{2L}\ln\left(\frac{1}{R_1R_1}\right) + \alpha_s + \alpha_a}, \tag{31}$$

$$= \frac{\frac{1}{2}\ln\left(\frac{1}{R_1}\right)}{\frac{1}{2}\ln\left(\frac{1}{R_1R_1}\right) + \alpha_s L + \alpha_a L}, \tag{32}$$

we can come to some interesting conclusions. From the above expressions, we can see that threshold gain will decrease as the gain length L increases, but the laser extraction efficiency will *decrease* as the gain length increases. In other words, unless we can eliminate α_s and α_a entirely, we cannot simultaneously achieve a low threshold gain and high efficiency.

Example

Consider a semiconductor laser material with $L = 300\mu$m, $\lambda = 0.85\mu$m, $n_{\text{eff}} = 3.5$, $\alpha_s + \alpha_a = 5/$cm. The output facet is as-cleaved, and the other facet is coated for 95% reflectivity.

Using the given values, we can calculate the reflectivity at the as-cleaved semiconductor/air interface as:

$$R_1 = \left|\frac{3.5 - 1.0}{3.5 + 1.0}\right|^2 = 0.31. \tag{33}$$

The mirror loss α_{m1} can be calculated as:

$$\alpha_{m1} = \frac{1}{2L}\ln\left(\frac{1}{R_1}\right) = 19.6/\text{cm}. \tag{34}$$

The second (coated) mirror loss is:

$$\alpha_{m2} = \frac{1}{2L}\ln\left(\frac{1}{R_2}\right) = 0.85/\text{cm}. \tag{35}$$

The total cavity loss can then be calculated:

$$\alpha_c = \alpha_{m1} + \alpha_{m2} + \alpha_s + \alpha_a = 25.4/\text{cm}. \tag{36}$$

Therefore, the threshold gain of this laser cavity will be

$$G_{th} = 25.4/\text{cm}. \tag{37}$$

The extraction efficiency can be calculated as

$$\eta_{ext} = \frac{\alpha_{m1}}{\alpha_c} = \frac{19.6}{25.4} = 77.1\%. \tag{38}$$

Compared to LED's, this value is significantly higher.

Photon Lifetimes

It is also worthwhile introducing the concept of a photon lifetime, τ_p. This is the temporal equivalent of the cavity loss α_c. Whereas $\frac{1}{\alpha_c}$ is the average distance a photon travels before it is lost (due to transmission, absorption or scattering losses), τ_p is the average time a photon survives before it is lost to these processes. Therefore, we can define

$$\tau_p = \left(\frac{1}{\alpha_c}\right)\left(\frac{n_{\text{eff}}}{c}\right) \tag{39}$$

where c is the speed of light in free space.

Similarly, we can also define a photon lifetime due to a specific loss mechanism. For example, the photon lifetime due to the transmission through the output mirror can be defined as

$$\tau_{pm1} = \left(\frac{1}{\alpha_{m1}}\right)\left(\frac{n_{\text{eff}}}{c}\right), \tag{40}$$

where α_{m1} is the mirror loss as defined in equation (24).

We can also write the extraction efficiency, equation (29), in terms of lifetimes:

$$\eta_{ext} = \frac{\alpha_{m1}}{\alpha_c} = \frac{\tau_p}{\tau_{pm1}}. \tag{41}$$

Quality Factor

Optical cavities are more general than Fabry-Perot cavities consisting of two parallel mirrors. The parameter used for characterizing resonators is the quality factor (Q-factor). This is the ratio between the energy stored in the cavity and the energy lost during one cycle of oscillation expressed in radians. That is,

$$Q = 2\pi\frac{\text{Energy stored in the cavity}}{\text{Energy lost per cycle}}. \tag{42}$$

If we represent the initial energy in the cavity as W_o, using the photon lifetime τ_p, we can write the energy decay as a function of time as

$$W(t) = W_o e^{-\frac{t}{\tau_p}}. \tag{43}$$

Since the duration of one oscillation is $\frac{1}{\nu}$, where ν is the frequency of light, we can write

$$\text{Energy lost per cycle} = W\left(1 - e^{-\frac{1}{\nu\tau_p}}\right). \tag{44}$$

Assuming that this energy loss is not too excessive (i.e., the cavity has a reasonably high Q), then $\nu\tau_p \gg 1$. This results in

$$1 - e^{-\frac{1}{\nu\tau_p}} \approx \frac{1}{\nu\tau_p}. \tag{45}$$

Therefore, the expression for the Q-factor becomes

$$Q = 2\pi\left(\frac{W}{W/(\nu\tau_p)}\right) = \omega\tau_p \tag{46}$$

where ω is the angular frequency of light. In other words, the quality factor is the number of radians of oscillations that light undergoes in one photon lifetime.

Example (Cont'd)

Continuing the previous example, if $n_g = 4.5$, we can calculate the spacing between lasing lines. This becomes:

$$\Delta\lambda = \frac{\lambda^2}{2n_g L} = 0.27 \text{ nm}. \tag{47}$$

The photon lifetime can be calculated as

$$\tau_p = \left(\frac{1}{\alpha_c}\right)\left(\frac{n_{\text{eff}}}{c}\right) = \left(\frac{1}{25.4}\right)\left(\frac{3.5}{3 \times 10^{10}}\right) = 4.6 \text{ ps}. \tag{48}$$

Assuming an operating wavelength of 850 nm, the Q-factor of the cavity becomes

$$Q = \omega\tau_p = 2\pi\left(\frac{c}{\lambda}\right)\tau_p = 10,200. \tag{49}$$

Comparison with Other Gain Media

One of the distinct advantages of semiconductor lasers is its ability to produce large gain values. Semiconductor gain values can easily exceed 500/cm. YAG crystals, for example, has a gain of around 2/cm, and HeNe amplifiers have a gain of around 0.005/cm. As we can verify from equation (30), a small length will require a high gain values. The high gain from semiconductors is what allows lasers to be made in microchip configuration, while other laser systems require much longer lengths. There is, however, a trade-off. The small volumes of semiconductor lasers result in very high optical and electrical power densities inside the chip. High optical density can lead to non-linearity and material damage. High electrical current densities can lead to high temperatures. As a result, semiconductor lasers often have lower output power levels than other types of lasers.

Semiconductor Chip Lasers & Heterostructures

A semiconductor laser is typically constructed as a slab PN junction diode with the current flowing vertically through the slabs as shown in Fig 7. The mirrors that create the optical cavity are normally the cleaved crystal facets of the semiconductor chip. The native reflection from a semiconductor/air interface is around 31%, as calculated in the example above, but this value can be modified by adding thin film coatings on the facets.

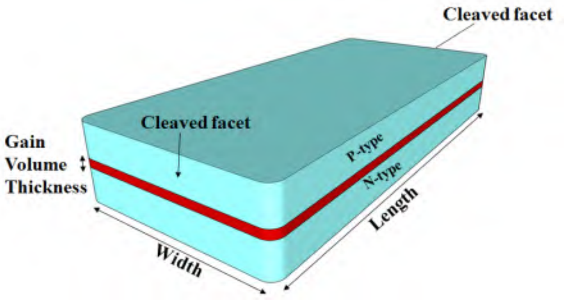

Figure 7: Typical slab configuration of semiconductor diode lasers.

The layer structure of the junction diode is rarely made from the same semiconductor material. Even though such junctions (homojunctions) are commonly used for electronic components, the first generation of semiconductor lasers made from homojunctions exhibited very poor efficiencies due to the large currents required to reach the gain threshold. As a result, subsequent laser structures used materials with dissimilar bandgaps (but with the same crystal structure) to confine the carriers in the junction and reduce the current required to reach threshold. These are known as hetero-structures, one of which is shown in Fig 8. The p-type and n-type regions are made from dissimilar semiconductors, such that there is a discontinuity in the conduction and valence bands. This discontinuity is designed to block (or reduce) the diffusion of electrons and holes, resulting in a large carrier concentration at the interface at very low forward currents.

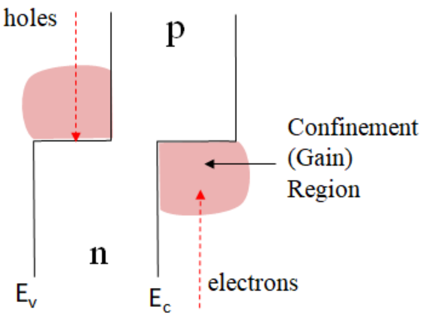

Figure 8: Single heterostructure used in early semiconductor diode lasers.

Nevertheless, the single heterostructure as shown in Fig 8, still had relatively poor carrier confinement. Due to diffusion, electrons and holes spread out over a wide distance, resulting in a relatively lower carrier concentration (which is the key factor that determines optical gain). The next major development in semiconductor lasers was the use of double hetero-structures (DH), as shown in Fig 9. This is a sandwich structure. The p-type and n-type regions are made from a wide bandgap material (such as Al$_x$Ga$_{1-x}$As) and the central region is made up of a thin layer of a narrower bandgap material (such as GaAs). This structure allowed the electrons and holes to be confined on both sides, which enabled very high carrier concentrations to be built-up at relatively low forward currents. These structures lead to some of the first room temperature semiconductor lasers. The photon energy was from the central low-bandgap layer. In the case of GaAs, whose bandgap is 1.42eV, the emission wavelength would be approximately $\frac{1.24}{E_g} = 0.87\mu$m.

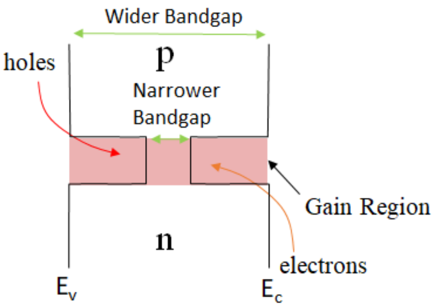

Figure 9: Double Heterostructures (DH) semiconductor laser

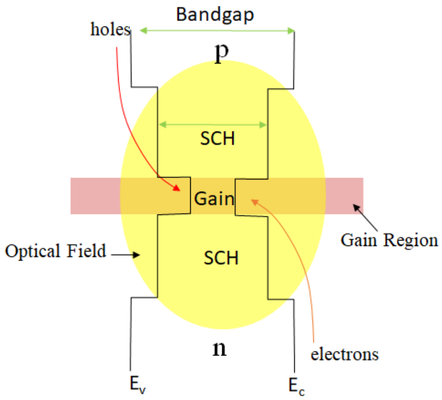

Figure 10: Separate confinement heterostructure (SCH) semiconductor laser

This DH semiconductor laser structure still had a weakness. Even though the electrons and holes were sufficiently confined, there was no confinement of photons. As a result, optical field was spread out over a large distance, and only a small portion of it overlapped with the gain region, resulting in a poor overall gain of the optical field. This was partially resolved by the introduction of the Separate Confinement Heterostructure (SCH), which is depicted in Fig 10. This actually contains two double heterostructures (four interfaces). The central double-heterostructure region is for confining electrons and holes as before. The outer double-heterostructure is for confining photons. The refractive index of semiconductors is inversely related to the bandgap. The SCH structure, therefore, has a larger refractive index in the central DH region. However, because the laser wavelength is typically much longer than the width of the central DH region, a separate wider DH is needed to act as an optical waveguide. This is the primary purpose of the outer DH. The optical field profile is illustrated in Fig 10 by the yellow region. Even in the SCH structure, a large portion of the optical field does not overlap the gain region (which is limited to central DH region). The fraction of photons in the optical field that experience gain is referred to as the confinement factor. The volumetric confinement factor, Γ_v, is defined as

$$\Gamma_v = \frac{\text{Gain Volume}}{\text{Photon Volume}} = \frac{V_\gamma}{V_p}, \tag{50}$$

where V_p is the volume of the optical mode and V_γ is the volume of the gain region. This can also be expressed as an overlap integral between the optical field distribution $\phi(r)$ and the gain distribution $G(r)$:

$$\Gamma_v = \frac{\int_0^\infty |\phi(r)|^2 \, dr}{\int_0^\infty |\phi(r)|^2 G(r) \, dr}. \tag{51}$$

We can also define a cross-sectional confinement factor

$$\Gamma_A = \frac{A_\gamma}{A_p}, \tag{52}$$

where A_γ is the cross sectional area the gain region that overlaps with the optical field, and A_p is the optical mode field area.

The confinement factor can be evaluated based on the geometry of the SCH structure (optical waveguide) and the gain region. Typical values are in the range of 0.01 to 0.1. The confinement factor has a profound impact on the threshold gain of the laser. Instead of the threshold gain G_{th} being equal to the total cavity loss α_c (equation (28)), it will now be higher by a factor of $1/\Gamma_A$,

$$\begin{aligned} A_\gamma G_{th} &= A_p \alpha_c && (53)\\ G_{th} &= \frac{A_p}{A_\gamma}\alpha_c && (54)\\ &= \frac{1}{\Gamma_A}\alpha_c. && (55) \end{aligned}$$

As a result, the threshold gain will typically be 10 to 100 times larger than the cavity loss α_c.

Example (Cont'd)

In the previous example, we calculated the cavity loss (and hence the threshold gain) to be 25.4/cm. If the confinement factor of this structure was specified as 0.05, the actual threshold gain will become

$$G_{th} = \frac{\alpha_c}{0.05} = 508 \text{ /cm}. \tag{56}$$

How Optical Gain is Produced in a Semiconductor

Electrons in the valence band can absorb a photon and be elevated into the conduction band creating an electron and a hole. Similarly, electrons already in the conduction band can recombine with the holes in the valence band to produce photons, annihilating the electron and hole.

Absorption, by definition, is a stimulated process - it cannot occur without the presence of a photon. Emission, on the other hand, can take on two forms - stimulated or spontaneous. An electron and a hole can recombine at random to produce a photon. This is known as spontaneous emission. On the other hand, an existing photon can stimulate an electron to recombine with a hole to produce a new photon, which is exactly the reverse of the absorption process. This is known as stimulated emission.

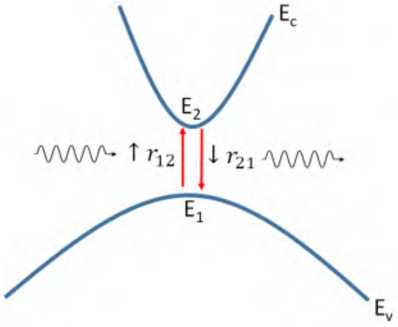

Figure 11: Upward and downward transitions absorbing/emitting photons

The stimulated absorption process can be written as

$$\uparrow r_{12}\left(E\right) = \left[\rho_c^{-1}\left(E_2\right) + \rho_v^{-1}\left(E_1\right)\right]^{-1} \; f\left(E_1\right) \;\; \left[1 - f\left(E_2\right)\right] \; \Phi\left(E\right) \;\; B_{12}, \tag{57}$$

where

- $\uparrow r_{12}\left(E\right)$ is the number of photons absorbed per unit time per unit volume per unit energy.
- $\left[\rho_c^{-1}\left(E_2\right) + \rho_v^{-1}\left(E_1\right)\right]^{-1}$ is the joint density of states, often written as $\rho_J\left(E\right)$, in the units of number of states per unit volume per unit energy.
- $f\left(E_1\right)$ is the Fermi distribution in the valence band (i.e., the probability of a valence state at energy E_1 being occupied).
- $\left[1 - f\left(E_2\right)\right]$ is the probability of a conduction band state with energy E_2 being vacant.

- $E_2 - E_1$ is the photon energy.
- $\Phi(E)$ is the photon density in the units of number of photons per unit volume per unit energy.
- B_{12} is the transition rate constant.

The emission process can be written as

$$\begin{aligned} \downarrow r_{21}(E) &= \downarrow r_{st}(E) + \downarrow r_{sp}(E) \qquad (58) \\ &= \underbrace{\rho_J(E)\ f(E_2)\ \left[1 - f(E_1)\right]\ \Phi(E)\ B_{21}}_{\text{Stimulated Emission}} + \\ &\quad \underbrace{\rho_J(E)\ f(E_2)\left[1 - f(E_1)\right]\ A_{21}}_{\text{Spontaneous Emission}}. \qquad (59) \end{aligned}$$

The joint density of states $\rho_J(E)$ is the combined density states between the conduction band ($\rho_c(E_2)$) and the valence band ($\rho_v(E_1)$) that interacts with a photon energy of $E = E_2 - E_1$. Densities are never added together to find the net density. Instead, they are added in inverse to find the net density. This is not specific only to electronic bands; even the densities we encounter in everyday life should be added in inverse.

Absorption and stimulated emission can be thought of as opposite effects. They both require the presence of a photon. Spontaneous emission, on the other hand, occurs by itself without any photon. However, there is no spontaneous absorption process - obviously one can't absorb a photon that does not exist.

At thermodynamic equilibrium, the upward and downward transitions will be balanced:

$$\uparrow r_{12}(E) = \downarrow r_{21}(E), \qquad (60)$$

with the photon density determined by Planck's law of black body radiation

$$\Phi(E) = \frac{24\pi E^2}{h^3 c^3} \frac{1}{e^{E/kT} - 1}. \qquad (61)$$

Notice that we have used the units of density per unit energy for $\Phi(E)$, which leads to a slightly nonstandard expression compared to the blackbody radiation spectrum that we discussed earlier. This results in the following relationship between the transition rate coefficients:

$$\begin{aligned} B_{12} &= B_{21} = B \qquad (62) \\ A_{21} &= B_{21} \frac{24\pi E^2}{h^3 c^3} = A. \qquad (63) \end{aligned}$$

The spontaneous emission coefficient A has the units of inverse time. In terms of A, the B coefficient becomes:

$$B = \frac{h^3 c^3 A}{24\pi E^2}. \qquad (64)$$

The difference between the emission and absorption is the net increase in photon density per unit time. This can be written as:

$$\frac{d\Phi(E)}{dt} = \downarrow r_{21}(E) - \uparrow r_{12}(E). \qquad (65)$$

Substituting for $\uparrow r_{12}(E)$ and $\downarrow r_{21}(E)$ from equations (57) and (59), and neglecting the spontaneous emission component results in:

$$\frac{d\Phi(E)}{dt} = B\rho_J(E)\left[f(E_2) - f(E_1)\right]\Phi(E) \tag{66}$$

$$= \mathcal{G}(E)\,\Phi(E)\,, \tag{67}$$

where $\mathcal{G}(E)$ is the photon gain per unit time per unit energy:

$$\mathcal{G}(E) = B\rho_J(E)\left[f(E_2) - f(E_1)\right]. \tag{68}$$

Spontaneous emission has a very small photon density compared to stimulated emission, hence we have neglected it in the above expression.

Temporal gain and spatial gain are related as

$$G = \mathcal{G}\left(\frac{n_{\text{eff}}}{c}\right) \tag{69}$$

where c is the speed of light and n_{eff} is the effective index of the medium.

The most important point to note in the gain expression is that the only factor that determines if gain is positive or negative is $f(E_2) - f(E_1)$. Under thermal equilibrium, the Fermi function $f(E_2)$ will be smaller than $f(E_1)$. Therefore, the gain will be negative. In other words, the semiconductor will exhibit net absorption under thermal equilibrium conditions. If we can get $f(E_2) > f(E_1)$, then it should be possible to get a positive optical gain. For this to happen, the probability of an electron occupying a higher energy state E_2 has to be *larger* than the probability of an electron occupying a lower energy state E_1. This is clearly an unnatural state, which is why optical gain and lasing does not occur naturally. This is known as population inversion. However, this condition can be induced inside a semiconductor, particularly in a PN junction, by current injection.

Under thermal equilibrium, the electrons in the conduction band and the holes in the valence band are characterized by a single Fermi level. However, under forward bias, we saw that the Fermi level splits into E_{FN} and E_{FP} near the junction, producing an unbalanced electron-hole distribution. The electron concentration is determined by E_{FN} and the hole concentration is determined by E_{FP}. These are known as quasi-Fermi levels. The Fermi function corresponding to these quasi-Fermi levels were

$$f_c(E_2) = \frac{1}{e^{(E_2 - E_{FN})/kT} + 1} \tag{70}$$

$$f_v(E_1) = \frac{1}{e^{(E_1 - E_{FP})/kT} + 1}. \tag{71}$$

Fig 12 illustrates three different Fermi functions. $f(E)$ is the Fermi function at thermal equilibrium. At the junction, this will be at the intrinsic level E_i. $f_c(E)$ and $f_v(E)$ are the Fermi functions that corresponds to the quasi-Fermi levels under injection. $f_c(E)$ is the Fermi function that applies for electrons in the conduction band, and $f_v(E)$ is the Fermi function that corresponds to the holes in the valence band. Both of these have different quasi-Fermi levels E_{FN} and E_{FP}, respectively. The density of states functions $\rho_c(E)$ and $\rho_v(E)$, along with the conduction and valence band edges E_c and E_v are also shown. We should be able to verify from this figure that the number of conduction band electrons has risen compared to the thermal equilibrium state, and the number of holes in the valence band has also risen compared to the thermal equilibrium state. However, population inversion has not been reached. This is because at the lowest level of the conduction band is higher than E_{FN}, which makes the probability of occupation smaller than 0.5. The highest level of the valence band is lower than E_{FP}, which makes the probability of occupation higher than 0.5. This means that the upper state (E_2) clearly has a lower occupation probability than the lower state (E_1). This is known as low-level injection.

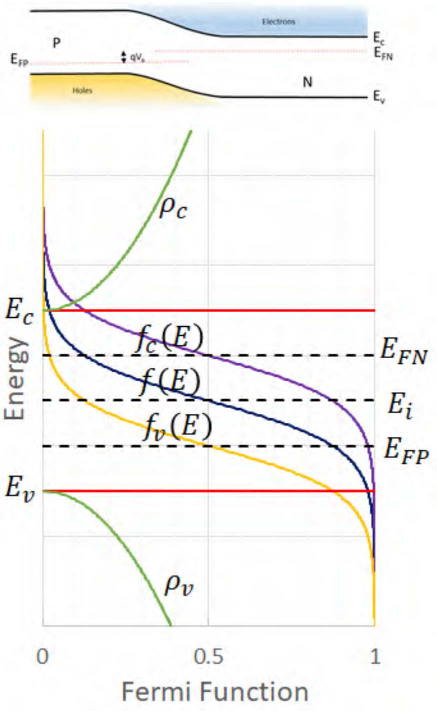

Figure 12: Depiction of Fermi functions under low level injection

Next, consider a higher level of injection. In this case, the quasi-Fermi level E_{FN} has crossed into and above the lowest level in the conduction band, and E_{FP} has crossed into and below the highest level in the valence band. This can only be achieved with very heavily doped junctions such that the Fermi levels fall into the bands. States in the region between E_c and E_{FN} will therefore have a higher probability of occupation than the states in the region between E_v and E_{FP}. These regions are shown shaded in Fig 13. As a result, $f(E_2) - f(E_1)$ will be positive for transition between these regions, resulting in optical gain for those photons..

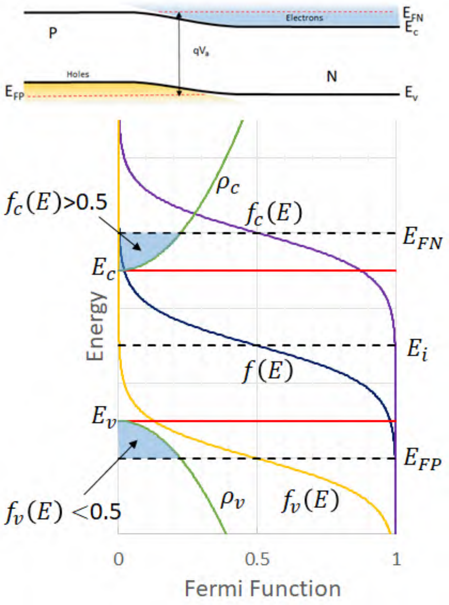

Figure 13: Depiction of Fermi functions under high level injection

We can summarize by saying that optical gain will be produced if the injection is high enough to move the quasi-Fermi levels past the band edges. All photons with energy greater than the bandgap energy E_g, and smaller than $E_{FN} - E_{FP}$ will experience positive gain.

Since $E_{FN} - E_{FP} = qV_a$, where V_a is the applied voltage, we need a voltage that exceeds the bandgap voltage. This will result in an extremely large forward current, most of which are collected by the terminals. Those that participate in optical emission will be a tiny fraction of the carriers. As a result, early semiconductor diode lasers had very poor efficiency, and had to be cooled aggressively to mitigate the heating due to the large current.

Achieving Population Inversion in Diode Lasers

The maximum injection of both types of carriers occurs inside the space charge layer. However, the maximum carrier concentration that can be reached due to injection cannot exceed the majority carrier concentration on the injecting side. This occurs at $V_a = V_{bi}$, also known as the flat band condition. Therefore, in order to achieve population inversion, the n- and the p-type regions have to be doped such that their Fermi levels fall inside the conduction and valence bands, respectively. This usually requires very large values of doping. For instance, the carrier concentration at the metallurgical junction can be written as

$$n_j = N_D e^{-(V_{bi}-V_a)/(2V_t)} \tag{72}$$

$$p_j = N_A e^{-(V_{bi}-V_a)/(2V_t)}. \tag{73}$$

As $V_a \rightarrow V_{bi}$ the carrier concentration at the junctions will approach N_D and N_A. If N_D and N_A correspond to E_{FN} and E_{FP} inside the conduction and valence bands, respectively, then we can achieve population inversion at the metallurgical junction. This, unfortunately, is not practical because as V_a approaches V_{bi}, not only does the diode current flow equation becomes less and less valid, the current flow also increases exponentially to impractical values. As a result, the diode will be destroyed before any population inversion is achieved.

This is where heterostructure diodes became important. Heterostructures make it easier to achieve population inversion at a smaller voltage and a smaller forward current. This concept was introduced in Figs 8, 9 and 10.

Fig 14 shows the energy band diagram of a double heterostructure diode. This is the same structure discussed previously in Fig 9. Also, as discussed before, the carrier concentration at the metallurginal junction will be at its intrinsic value, and the Fermi level will be at the center of the bandgap. The outer regions (larger bandgap regions) are moderately doped, and the narrow bandgap region is typically left undoped.

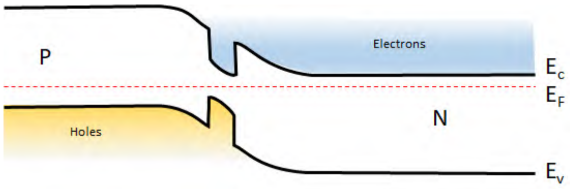

Figure 14: Approximate energy band diagram of a double heterostructure under zero bias

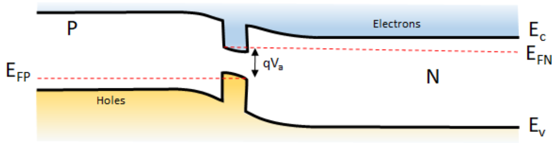

Figure 15: Approximate energy band diagram of a double heterostructure under forward bias

Under forward bias (low level injection), the band structure will be as shown in Fig 15. The important point worth noting is that the narrow bandgap at the center makes it easier to achieve population inversion without having to dope the p and p regions excessively, or having to apply very large voltages. Due to the smaller bandgap, the quasi-Fermi levels are able to penetrate above the band edges of that material even at relatively low forward bias voltages. There are additional benefits to this structure as well. The smaller bandgap region acts like a carrier trap that significantly reduces the diffusion current. Nearly all of the forward current arises from recombination in the smaller bandgap region. Therefore, we can approximate the forward current density as

$$J = \frac{n_i t \left(e^{V_a/V_t} - 1\right)}{\tau} \tag{74}$$

where n_i is the intrinsic carrier concentration in the smaller bandgap material, t is the thickness of that material, and τ is the recombination life time in that material. This equation assumes that the carrier profiles inside the smaller bandgap region is flat, or alternatively that the diffusion length of carriers is much larger than t. Since almost all of the carrier recombination and photon emission takes place in this region, it is also referred to as the *active region* of the laser.

When stimulated emission is present, its effect will be to create an additional current component. This will make the total current density equal to

$$J = \frac{n_i t \left(e^{V_a/V_t} - 1\right)}{\tau} + t \int_0^\infty \mathcal{G}\left(E\right) \Phi\left(E\right) dE \tag{75}$$

where $\mathcal{G}$ is the temporal gain, Φ is the photon density. This assumes that all of the stimulated emission takes place in the active region.

Calculation of Gain in Semiconductors

Optical gain can be calculated from equation (68), or by using equation (69) it can be convered to a gain per unit length coefficient. As a result,

$$G = \left(\frac{n_{\text{eff}}}{c}\right) \frac{h^3 c^3 A}{24\pi E^2} \rho_J\left(E\right) \left[f_c\left(E_2\right) - f_v\left(E_1\right)\right]. \tag{76}$$

Since gain should be expressed as a function of photon energy (or wavelength), we need to convert E_1 and E_2 in terms of the photon energy E. We did this earlier under LEDs. The result was:

$$E = E_g + \frac{\hbar^2 k^2}{2m_r}, \tag{77}$$

where E_g is the material bandgap and m_r is a reduced effective mass defined as

$$\frac{1}{m_r} = \frac{1}{m_c} + \frac{1}{m_v}. \tag{78}$$

This allowed us to get an expression for E_1 and E_2 in terms of the photon energy E.

$$E_2 = E_c + \left(E - E_g\right) \frac{m_r}{m_c} \tag{79}$$

$$E_1 = E_v - \left(E - E_g\right) \frac{m_r}{m_v}. \tag{80}$$

Furthermore, we were able to express the quasi-Fermi levels as

$$E_{FN} = E_i + \Delta E_{FN} \tag{81}$$

$$E_{FP} = E_i - \Delta E_{FP}, \tag{82}$$

where E_i is the intrinsic level, and ΔE_{FN} and ΔE_{FP} are the shifts in the Fermi level due to the applied bias. As a result, we can write the expressions for the Fermi distributions in terms of the photon energy E:

$$f_c\left(E_2\right) = \frac{1}{e^{\left(E_c - E_i - \Delta E_{FN} + (E - E_g)\frac{m_r}{m_c}\right)/kT} + 1} \tag{83}$$

$$f_v\left(E_1\right) = \frac{1}{e^{\left(E_v - E_i + \Delta E_{FP} - (E - E_g)\frac{m_r}{m_v}\right)/kT} + 1}, \tag{84}$$

and derive the complete expression for the gain coefficient:

$$\begin{aligned} G \quad = \quad & \left(\frac{n_{\text{eff}}}{c}\right)\frac{Ah^3c^3}{24\pi E^2}\left[\underbrace{\frac{\pi^2\hbar^3}{\sqrt{2}m_c^{3/2}\left(E-E_c\right)^{1/2}}}_{\rho_c^{-1}(E)}+\underbrace{\frac{\pi^2\hbar^3}{\sqrt{2}m_v^{3/2}\left(E_v-E\right)^{1/2}}}_{\rho_v^{-1}(E)}\right]^{-1} \\ & \left[\underbrace{\frac{1}{e^{\left(E_c-E_i-\Delta E_{FN}+(E-E_g)\frac{m_r}{m_c}\right)/kT}+1}}_{f_c(E_2)}-\underbrace{\frac{1}{e^{\left(E_v-E_i+\Delta E_{FP}-(E-E_g)\frac{m_r}{m_v}\right)/kT}+1}}_{f_v(E_1)}\right]. \end{aligned} \tag{85}$$

The joint density of states can be simplified in terms of the reduced effective mass m_r. In fact, it has the exact form as the conduction or valence band expression:

$$\rho_J\left(E\right)=\frac{\sqrt{2}m_r^{3/2}}{\pi^2\hbar^3}\left(E-E_g\right)^{1/2}. \tag{86}$$

As a result, the gain expression can be further simplified to:

$$\begin{aligned} G \quad = \quad & \left[\frac{\sqrt{2}}{3}\right]\left[\frac{An_{\text{eff}}c^2\left(m_r\right)^{3/2}\left(E-E_g\right)^{1/2}}{E^2}\right] \\ & \left[\underbrace{\frac{1}{e^{\left(E_c-E_i-\Delta E_{FN}+(E-E_g)\frac{m_r}{m_c}\right)/kT}+1}}_{f_c(E_2)}-\underbrace{\frac{1}{e^{\left(E_v-E_i+\Delta E_{FP}-(E-E_g)\frac{m_r}{m_v}\right)/kT}+1}}_{f_v(E_1)}\right]. \end{aligned} \tag{87}$$

Fig 16 shows the density of states in the conduction and valence bands as well as the Fermi functions $f_c(E)$ and $f_v(E)$, using GaAs as the example. The energy level has been arbitrarily set to zero at the top of the valence band, such that the bottom of the conduction band is at 1.42eV (which is the bandgap energy). In this case $E_{fc} - E_c$ and $E_v - E_{fv}$ have been chosen to be 0.1eV, such that the Fermi energy falls 0.1eV above the conduction band minimum and 0.1eV below the valence band maximum to ensure population inversion.

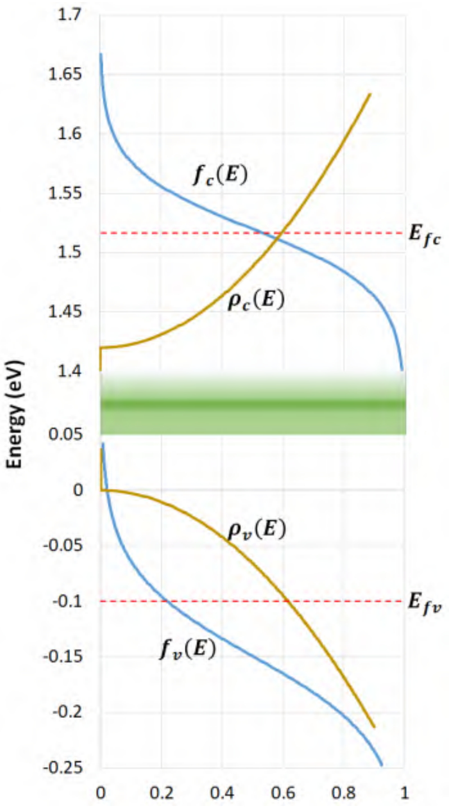

Figure 16: Density of states and the Fermi functions in the conduction and valence bands.

The lower part of the figure shows the valence band density of states $\rho_v(E)$, and the corresponding Fermi function $f_v(E)$. We can verify that population inversion will exist for transitions from all states within 0.1eV above the conduction band minimum to all states 0.1eV below the valence band maximum.

Fig 17 shows the calculated gain from equation (87), using $(E_{FN} - E_c) =$ 5meV and $(E_v - E_{FP}) = 5$ meV, with $1/A = 0.5$ ns. The second curve shows the gain when $(E_{FN} - E_c) = (E_v - E_{FP}) = 10$ meV. We can see that this results in a positive gain between the bandgap energy of 1.42 eV and $(E_{FN} - E_{FP})$ (which is 1.43 eV and 1.44 eV for the two curves). The bandwidth of the gain is exactly equal to the quasi-Fermi level offsets (10 meV and 20 meV). The negative value of gain means that the material exhibits a net absorption, which occurs for all energies higher than the quasi Fermi level offsets. We can also see that the gain spectrum is similar to the one that we earlier assumed in Fig. 4 (except in that figure we sketched gain as a function of wavelength).

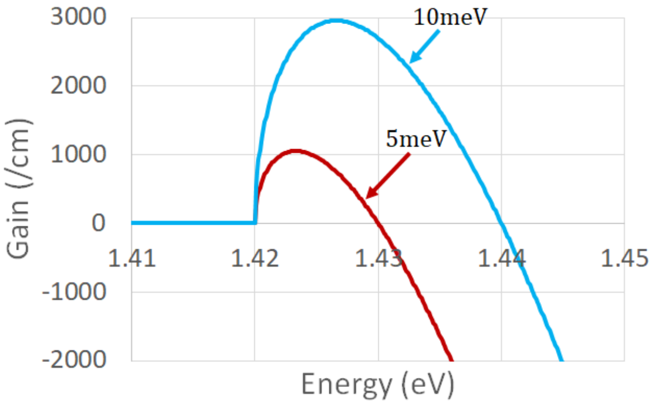

Figure 17: Calculated gain as a function of photon energy for $(E_{fc} - E_c) = (E_v - E_{fv}) = 1$ meV and 10 meV, and $A^{-1} = 0.5$ns.

The electron and hole concentrations that correspond to $E_{FN}-E_c=5$ meV and $E_v-E_{FP}=5$ meV can be found by integrating the density of states function with their respective Fermi functions:

$$\begin{aligned} n &= \int_{E_c}^{\infty}\rho_c\left(E\right)f\left(E\right)dE && (88) \\ p &= \int_{0}^{E_v}\rho_v\left(E\right)\left(1-f\left(E\right)\right)dE. && (89) \end{aligned}$$

For a bulk semiconductor, we had the density of states

$$\begin{aligned} \rho_c\left(E\right) &= \frac{\sqrt{2}m_c^{3/2}}{\pi^2\hbar^3}\left(E-E_c\right)^{1/2} && (90) \\ \rho_v\left(E\right) &= \frac{\sqrt{2}m_v^{3/2}}{\pi^2\hbar^3}\left(E_v-E\right)^{1/2}. && (91) \end{aligned}$$

Although it is possible to make some approximations to express the integrals in equations (88) and (89) in closed form, it is not difficult to carry out this integral numerically. Using a numerical integration, we can calculate the electron and hole concentrations that correspond to the bias condition of $E_{FN}-E_c=5$ meV and $E_v-E_{FP}=5$ meV. This results in

$$\begin{aligned} n &= 4.4\times10^{17}\text{ cm}^{-3} && (92) \\ p &= 7.04\times10^{19}\text{ cm}^{-3}. && (93) \end{aligned}$$

In this case the n and p concentrations are different. This is due to the difference in the density of states between the conduction and valence bands. As a result, even though we had assumed identical Fermi level offsets $E_{FN}-E_c$ and E_v-E_{FP}, the carrier concentrations from the integrals in equations (88) and (89) resulted in different values. Conversely, we could have started by assuming that $n=p$, and worked out the values for $E_{FN}-E_c$ and E_v-E_{FP} (which would be different). This would have resulted in a slightly different gain profile as compared to Fig 17.

Additionally, since lasing action occurs only near the peak value of the gain, we can plot this peak value from Fig 17 as a function of n (assuming p is equal to n). The result is Fig 18. At low carrier concentrations this can be modeled as a linear function,

$$G_p\left(n\right)=G_p'\left(\frac{n}{n_T}-1\right) \qquad (94)$$

where n_T is known as the transparent carrier density, and G_p' is the slope of the gain vs n curve, also known as the differential gain. At high concentrations, $G_p\left(n\right)$ becomes nonlinear with n. In quantum-wells and quantum-dots, which are commonly used as the active medium in semiconductor lasers, this function has different characteristics, primarily because their density of states functions have different forms. Therefore, in order to keep our description general, we will leave the peak gain vs carrier concentration function as $G_p\left(n\right)$.

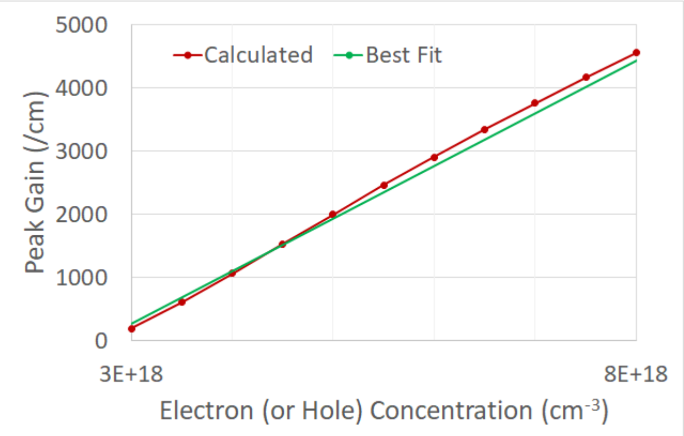

Figure 18: Peak gain vs electron (or hole) carrier concentration in GaAs using $A^{-1}=0.5$ns. Best fit values are: $G_p'=2217$/cm, $n_T=2.4\times10^{18}$/cm^3.

Carrier Rate Equation

Now we can generalize the carrier and photon interactions via the rate equation model. These are simply charge conservation and power conservation equations.

In the laser diode structure, the injected carriers are confined to the active region (smaller bandgap region) whose volume is V_γ, and the photons are confined to a larger heterostructure region (discussed earlier in Fig 10 under Separate Confinement Heterostructure) whose volume is V_p.

From current continuity, we can write:

$$V_\gamma \frac{dn}{dt} = \frac{I}{q} - V_\gamma \int_0^\infty \downarrow r_{21} dE - V_\gamma \int_0^\infty \downarrow r_{nr} dE. \tag{95}$$

This equation states that the difference between the rate of of electrons being injected into the conduction band, and the rate of electrons being removed from the conduction band results in a net increase in electrons.

Furthermore, the downward radiative transitions $\downarrow r_{21}$ consists of stimulated and spontaneous emission, as

$$\downarrow r_{21} = \downarrow r_{st} + \downarrow r_{sp}. \tag{96}$$

We can represent $\downarrow r_{sp}$ and $\downarrow r_{nr}$ as

$$\downarrow r_{sp} = \frac{n - n_i}{\tau_r} \tag{97}$$

$$\downarrow r_{nr} = \frac{n - n_i}{\tau_{nr}}. \tag{98}$$

From equation (67), total stimulated emission rate r_{st} can be written as

$$\int_0^\infty \downarrow r_{st} dE = \int_0^\infty \mathcal{G}(E)\, \Phi(E)\, dE. \tag{99}$$

Since lasing takes place only near the gain peak and not everywhere there is gain, we can represent this integral by a small band of energy

$$\int_0^\infty \mathcal{G}(E)\, \Phi(E)\, dE = \mathcal{G}(E)\ \bar{\Phi}(E)\, \Delta E \tag{100}$$

$$= G(E) \left(\frac{c}{n}\right) \bar{\Phi}(E)\, \Delta E \tag{101}$$

As a result, the carrier rate equation (95) becomes

$$\frac{dn}{dt} = \frac{I}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right) G(E)\, \Phi(E)\, \Delta E - \frac{n - n_i}{\tau_r} - \frac{n - n_i}{\tau_{nr}}. \tag{102}$$

For simplicity we can represent $\Phi(E)\, \Delta E = \phi$, and $G(E)\, \Phi(E)\, \Delta E = G_p \phi$ where G_p is the peak gain. We can also combine the radiative and nonradiative lifetimes, and write the final carrier rate equation as

$$\boxed{\frac{dn}{dt} = \frac{I}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right) G_p(n)\, \phi - \frac{n - n_i}{\tau}.} \tag{103}$$

Photon Rate Equation

The conservation of power can be written as

$$V_p \frac{d}{dt}\left(\Phi\left(E\right)\Delta E\right) = V_\gamma \int_0^\infty \downarrow r_{st} dE + V_\gamma \int_0^\infty \downarrow r_{sp} dE - V_p \int_0^\infty \frac{\Phi\left(E\right)}{\tau_p} dE \tag{104}$$

which basically states that the difference between the photon amplification and photon loss is the rate of increase in photon density. In this expression, τ_p is the photon lifetime. It is related to the total cavity loss as defined by equation (39).

Additionally, using the previous derivations and approximations, we can rewrite this equation as

$$V_p \frac{d\Phi\left(E\right)\Delta E}{dt} = V_\gamma \left(\frac{c}{n_{\text{eff}}}\right) G\left(E\right)\Phi\left(E\right)\Delta E + V_\gamma \frac{n}{\tau_r} - V_p \frac{\Phi\left(E\right)\Delta E}{\tau_p}. \tag{105}$$

The term $\frac{n}{\tau_r}$ is the spontaneous emission rate. This is a random emission of photons in all directions and polarizations. As we saw in the study of light emitting diodes, only a very small fraction of these photons will exit the cavity through the output mirror. We will represent this fraction (extraction efficiency of spontaneous emission, or spontaneous emission coupling factor) by the symbol Θ. Therefore, equation (105) becomes modified to

$$\frac{d\Phi\left(E\right)\Delta E}{dt} = \Gamma \left(\frac{c}{n_{\text{eff}}}\right) G\left(E\right)\Phi\left(E\right)\Delta E + \Gamma_v \Theta \frac{n-n_i}{\tau_r} - \frac{\Phi\left(E\right)\Delta E}{\tau_p}, \tag{106}$$

where Γ_v is the volumetric photon confinement factor defined earlier in equation (50). Again, we can set $\Phi\left(E\right)\Delta E = \phi$, and $G\left(E\right)\Phi\left(E\right)\Delta E = G_p \phi$. This allows equation (106) to be written as

$$\boxed{\frac{d\phi}{dt} = \Gamma_v \left(\frac{c}{n_{\text{eff}}}\right) G_p\left(n\right)\phi + \Gamma_v \Theta \frac{n-n_i}{\tau_r} - \frac{\phi}{\tau_p}.} \tag{107}$$

This is the photon rate equation. Note that we have included τ in the carrier rate equation (103), but only τ_r is included in the photon rate equation (106). This is because all recombination processes contribute to the decay in the carrier density, but only the radiative decay processes contribute to an increase in photons in the laser cavity.

The photon life time τ_p includes all loss mechanisms in the cavity. As defined in equations (39) and (40), the ”useful” loss is due to τ_{pm1}. Everything else contributes to an unrecoverable photon loss. Therefore, we can write the output power as

$$P_o = h\nu \frac{\phi V_p}{\tau_{pm1}}. \tag{108}$$

Steady State Solution of the Rate Equations

The carrier rate equation (103), and the photon rate equation (107) are coupled. Solving these two equations is not trivial, especially if time-varying currents are involved. However, we can solve for the steady-state condition if we set the derivatives $\frac{dn}{dt}$ and $\frac{d\phi}{dt}$ to zero. This results in

$$\frac{I}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right) G_p\left(n\right)\phi - \frac{n-n_i}{\tau} = 0 \tag{109}$$

$$\Gamma_v \left(\frac{c}{n_{\text{eff}}}\right) G_p\left(n\right)\phi + \Gamma_v \Theta \frac{n-n_i}{\tau_r} - \frac{\phi}{\tau_p} = 0. \tag{110}$$

The solution we are seeking is for the photon density ϕ as a function of the input current I. Even this is non-trivial unless we make a few additional assumptions. We can identify two cases:

- **Below lasing threshold:** is when the current is low and the gain is below the threshold to starting the lasing action. As a result, the current due to recombination will dominate the overall current, resulting in
$$G_p(n)\,\phi \ll \frac{n-n_i}{\tau}. \tag{111}$$
- **Above lasing threshold** is when the current due to stimulated emission far exceeds the recombination current. This can be expressed as:
$$G_p(n)\,\phi \gg \frac{n-n_i}{\tau}. \tag{112}$$

For below threshold, we can rewrite equation (109) as

$$n-n_i \approx \tau\frac{I}{V_\gamma q}. \tag{113}$$

We can substitute this into equation (110) to produce

$$\phi \quad = \quad \Gamma_v\Theta\frac{\tau\tau_p}{V_\gamma\tau_r}\frac{I}{q}. \tag{114}$$

Combining this with equation (108), we can get an expression for the output power

$$P_o = \Theta\left(\frac{h\nu}{q}\right)\left(\frac{\tau}{\tau_r}\right)\left(\frac{\tau_p}{\tau_{pm1}}\right)I. \tag{115}$$

Additionally, we have $\frac{h\nu}{q}=V_g$ (equivalent voltage of the bandgap), and $\frac{\tau}{\tau_r}=\eta_i$ (internal quantum efficiency). We also defined the extraction efficiency η_{ext} in equation (29), in terms of photon lifetime as well as in terms of loss coefficients. Therefore, the responsivity (optical power out/electrical current in) becomes

$$\mathcal{R} = \Theta\;V_g\;\eta_{ext}\;\eta_i. \tag{116}$$

Since the spontaneous emission coupling factor Θ is a very small number, this will result in a very small responsivity value. This is the expected behavior for lasers below their lasing threshold.

Lasing threshold will be reached when the gain is equal to the total cavity losses.

$$G_p(n_{th}) = \frac{\alpha_c}{\Gamma_v}, \tag{117}$$

where n_{th} is the carrier density at the threshold gain. From this, we can get the threshold current. Using equation (109) and using the below-threshold approximation, we can get

$$I_{th} = q\frac{V_\gamma n_{th}}{\tau}. \tag{118}$$

For above threshold operation, we can write equation (109) as

$$\left(\frac{c}{n_{\text{eff}}}\right)G_p(n)\,\phi \approx \frac{I}{V_\gamma q}. \tag{119}$$

Substituting this into equation (110), we can get

$$\Gamma_v \frac{I}{V_\gamma q} = \frac{\phi}{\tau_p}. \tag{120}$$

Combining this with equation (108) results in

$$P_o = \left(\frac{h\nu}{q}\right)\left(\frac{\tau_p}{\tau_{pm1}}\right) I. \tag{121}$$

Since this is for current values above the threshold current, we can offset the injection current by I_{th} and write it as

$$\begin{aligned} P_o &= \left(\frac{h\nu}{q}\right)\left(\frac{\tau_p}{\tau_{pm1}}\right)(I - I_{th}) && (122) \\ &= V_g\, \eta_{ext}\, \left(I - I_{th}\right). && (123) \end{aligned}$$

and the responsivity becomes:

$$\mathcal{R} = V_g\, \eta_{ext}. \tag{124}$$

This responsivity is significantly larger than equation (116), because it does not contain the spontaneous emission coupling factor Θ, or the internal quantum efficiency η_i.

External Quantum Efficiency

In addition to responsivity, we can also define an external quantum efficiency, η_e. This is the number of photons produced for each injected electron. From the above description, we can get the following results:

$$\begin{aligned} \eta_e &= \Theta\, \eta_{ext}\, \eta_i && \text{(below threshold)} && (125) \\ \eta_e &= \eta_{ext} && \text{(above threshold)}. && (126) \end{aligned}$$

Example (Cont'd)

Let's reconsider the example from before, which was a semiconductor laser with $L = 300\mu$m, $\lambda = 0.85\mu$m, $n_{\text{eff}} = 3.5$, $\alpha_s + \alpha_a = 5$/cm. We calculated $\alpha_c = 25.4$/cm, and $\alpha_{m1} = 19.6$/cm. This would result in an extraction efficiency

$$\eta_{ext} = \frac{\tau_p}{\tau_{pm1}} = \frac{\alpha_{m1}}{\alpha_c} = 0.77. \tag{127}$$

Furthermore, lets assume that the the internal quantum efficiency is $\eta_i = 0.5$, and the total recombination lifetime is $\tau = 2$ns. Also assume that the spontaneous emission coupling factor is $\Theta = 0.01$. We can now calculate the responsivity below threshold, using equation (116),

$$\mathcal{R} = \Theta\, V_g\, \eta_{ext}\, \eta_i = 0.01 \times 1.42 \times 0.77 \times 0.5 = 5.46 \text{ mW/A}. \tag{128}$$

Above threshold, the responsivity becomes

$$\mathcal{R} = V_g\, \eta_{ext} = 1.42 \times 0.77 = 1.09 \text{ W/A}, \tag{129}$$

which as expected, is significantly higher than the below-threshold value.

The threshold current can be calculated if we know a few additional parameters. The confinement factor was given as $\Gamma_v = 0.05$, and the gain volume (width × height × length of the gain region) is $V_\gamma = 5\mu$m $\times\ 0.1\mu$m $\times\ 300\mu$m $= 150\mu$m^3. Following the calculated results presented in Fig 18, we can look up the carrier density corresponding to the required gain value of $G_{th} = \frac{\alpha_c}{\Gamma} = 508$/cm. This is 3.3×10^{18} cm^{-3}. Then we can use equation (118) to calculate the threshold current:

$$I_{th} = q\frac{V_\gamma n_{th}}{\tau} = 39 \text{ mA}. \quad (130)$$

The light-current (L-I) curve of this laser diode is shown in Fig 19, which shows two distinct regions for below-threshold, and above-threshold operation. In practice, the transition from below-threshold to above-threshold will not be abrupt, nor would the above-threshold responsivity remain constant. At high currents, the responsivity will decline due to internal heating. The effects of increasing temperature on gain can be verified from equation (87). Additionally, the spontaneous emission coefficient, A, will also decline with increasing temperature. This has the effect of reducing the internal quantum efficiency. The results of these effects are shown by the dashed red line in Fig 19.

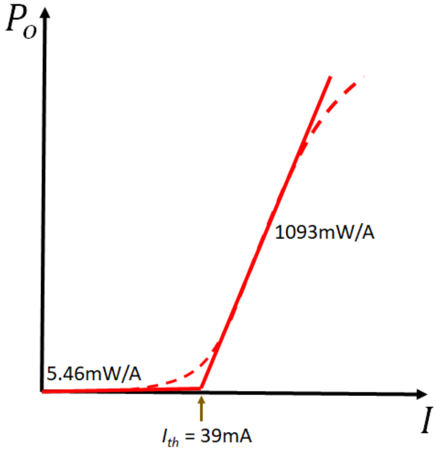

Figure 19: Light-current (L-I) curve of the laser diode example.

We can also note that the carrier concentration will increase linearly with current, as determined by equation (113). When the laser is above threshold, the carrier concentration will remain pinned at the threshold value as per equation (117). All of the injected carrier above that level will go to support stimulated emission. Using the example parameters from above, we can calculate these carrier density values. For below-threshold operation:

$$n = \tau\frac{I}{V_\gamma q}.$$

The carrier density will saturate at the threshold value of 3.3×10^{18} cm^{-3}. The behavior of carrier concentration vs current is shown in Fig 20. Similar to the L-I curve, we can identify two distinct regions of operation - a linearly rising carrier density below-threshold, and a constant value above-threshold. In practice, however, the carrier density will not remain perfectly flat above threshold. As discussed previously, the peak gain value will decline due to internal heating. This will produce a slight increase in the carrier density with injected current above-threshold.

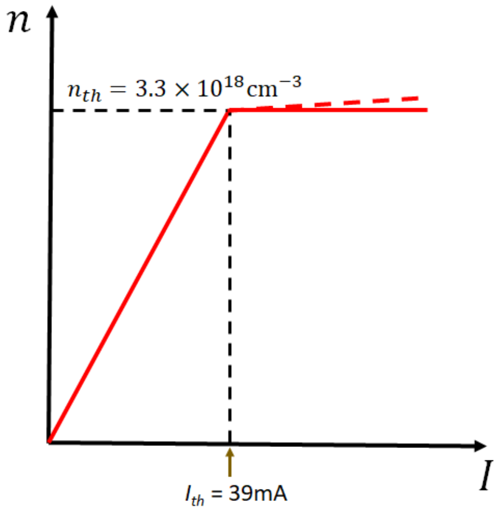

Figure 20: Carrier concentration vs injected current for the laser diode example.

Figure 21 shows the gain versus current curve for the same example. Because of the linear relationship between carrier density and peak gain (which was discussed in Figure 18) it follows the same behavior as Figure 20. Initially, the gain value increases as the current increases. Once threshold is reached, it remains constant at the threshold gain value. The additional current beyond threshold goes towards increasing optical output power.

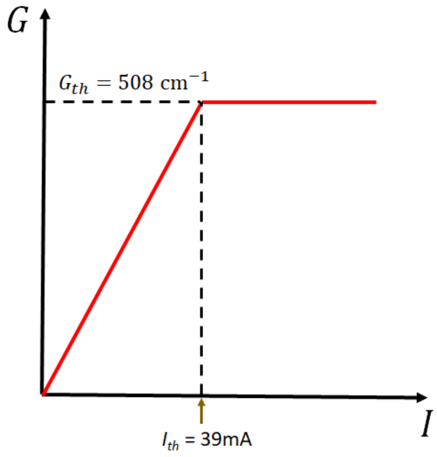

Figure 21: Gain vs injected current for the laser diode example.

Small-Signal Current Modulation of Laser Diodes

One of the distinct advantages of diode lasers as compared to other lasers (such as gas lasers, fiber lasers and other optically pumped solid state lasers) is its ability to be directly modulated by the input current. This allows analog or digital information to be encoded into the light output, allowing it to be utilized as an optical transmitter in communication systems. In the case of LEDs, the modulation bandwidth was determined primarily by the total recombination lifetime τ. With laser diodes, carrier lifetime does not play a role in the maximum modulation bandwidth. Instead, photon lifetime, differential gain and photon density are the primary factors that determine its modulation bandwidth. With proper design, it is possible to get modulation bandwidths well into the GHz range. We can study this effect by solving the two rate equations that we derived earlier. We will also assume that the laser is biased well above threshold such that the effects of sponatenous emission can be ignored. The two rate equations then become:

$$\frac{dn}{dt} = \frac{I}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right) G_p(n)\,\phi \tag{132}$$

$$\frac{d\phi}{dt} = \Gamma\left(\frac{c}{n_{\text{eff}}}\right) G_p(n)\,\phi - \frac{\phi}{\tau_p}. \tag{133}$$

The solution we seek is to obtain an expression for ϕ due to a time varying current I. Unfortunately, the term $G_p(n)\,\phi$ contains a product of n and ϕ which make these equations not only coupled, but also nonlinear. However, we can find a solution if we make certain approximations to linearize these equations. We will assume a small signal imposed on the dc operating condition, such that

$$I = I_o + \delta I e^{j\omega t} \tag{134}$$

$$n = n_o + \delta n e^{j\omega t} \tag{135}$$

$$G_p = G_{po} + \delta G_p e^{j\omega t} \tag{136}$$

$$\phi = \phi_o + \delta\phi e^{j\omega t}, \tag{137}$$

where I_o, n_o and ϕ_o satisfy the steady state condition discussed in the previous section. Even though the carrier density was shown to level off at n_{th} and the gain at G_{th}, under modulation, they experience transient excursions above these values. Therefore, the values for δn and δG_p in equations (135) and (136) will be nonzero.

The DC operating point satisfies the following conditions:

$$\begin{aligned} \frac{I_o}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right) G_{po}\phi_o &= 0 && (138)\\ \Gamma\left(\frac{c}{n_{\text{eff}}}\right) G_{po}\phi_o - \frac{\phi_o}{\tau_p} &= 0, && (139)\\ \Rightarrow \Gamma\left(\frac{c}{n_{\text{eff}}}\right) G_{po} &= \frac{1}{\tau_p}. && (140) \end{aligned}$$

Substituting the small signal modulations (134) - (137) into the rate equations (132) and (133) results in:

$$\begin{aligned} \frac{dn}{dt} &= \frac{I_o + \delta I e^{j\omega t}}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po} + \delta G_p e^{j\omega t}\right)\left(\phi_o + \delta\phi e^{j\omega t}\right) && (141)\\ \frac{d\phi}{dt} &= \Gamma\left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po} + \delta G_p e^{j\omega t}\right)\left(\phi_o + \delta\phi e^{j\omega t}\right) - \frac{\left(\phi_o + \delta\phi e^{j\omega t}\right)}{\tau_p}. && (142) \end{aligned}$$

Next, we will remove all terms involving the product of two small signal terms, such as $\delta G_p \delta\phi$. This also removes the nonlinear terms containing twice the frequency components:

$$\begin{aligned} \frac{dn}{dt} &= \frac{I_o + \delta I e^{j\omega t}}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po}\phi_o + G_{po}\delta\phi e^{j\omega t} + \phi_o \delta G_p e^{j\omega t}\right) && (143)\\ \frac{d\phi}{dt} &= \Gamma\left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po}\phi_o + G_{po}\delta\phi e^{j\omega t} + \phi_o \delta G_p e^{j\omega t}\right) - \frac{\left(\phi_o + \delta\phi e^{j\omega t}\right)}{\tau_p}. && (144) \end{aligned}$$

We can also group the terms that contain the dc solutions, resulting in:

$$\begin{aligned} \frac{dn}{dt} &= \underbrace{\frac{I_o}{V_\gamma q} + \left(\frac{c}{n_{\text{eff}}}\right) G_{po}\phi_o}_{\text{Eqn (138)}} + \frac{\delta I e^{j\omega t}}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po}\delta\phi e^{j\omega t} + \phi_o \delta G_p e^{j\omega t}\right) && (145)\\ \frac{d\phi}{dt} &= \underbrace{\Gamma\left(\frac{c}{n_{\text{eff}}}\right) G_{po}\phi_o - \frac{\phi_o}{\tau_p}}_{\text{Eqn (140)}} + \Gamma\left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po}\delta\phi e^{j\omega t} + \phi_o \delta G_p e^{j\omega t}\right) - \frac{\delta\phi e^{j\omega t}}{\tau_p}. && (146) \end{aligned}$$

Since equations (138) and (140) are equal to zero, this results in further simplifications of the above two equations:

$$\begin{aligned} \frac{dn}{dt} &= \left[\frac{\delta I}{V_\gamma q} - \left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po}\delta\phi + \phi_o \delta G_p\right)\right] e^{j\omega t} && (147)\\ \frac{d\phi}{dt} &= \left[\Gamma\left(\frac{c}{n_{\text{eff}}}\right)\left(G_{po}\delta\phi + \phi_o \delta G_p\right) - \frac{\delta\phi}{\tau_p}\right] e^{j\omega t}. && (148) \end{aligned}$$

Clearly, these two equations are coupled. However, it is possible to unlink them by eliminating n and expressing the one equation in terms of ϕ only. This can be done by taking the second derivative of the photon rate equation and substituting the carrier rate equation to eliminate all terms involving n. Taking the derivative of equation (148), and substituting (140) we can get:

$$\begin{aligned}
\frac{d^2\phi}{dt^2} &= j\omega\left[\Gamma\left(\frac{c}{n_{\textsf{eff}}}\right)\left(G_{po}\delta\phi+\phi_o\delta G_p\right)-\frac{\delta\phi}{\tau_p}\right]e^{j\omega t} &(149)\\
-\omega^2\delta\phi e^{j\omega t} &= j\omega\left[\Gamma\left(\frac{c}{n_{\textsf{eff}}}\right)G_{po}\delta\phi+\Gamma\left(\frac{c}{n_{\textsf{eff}}}\right)\phi_o\delta G_p-\frac{\delta\phi}{\tau_p}\right]e^{j\omega t} &(150)\\
j\omega\delta\phi &= \left[\cancel{\frac{\delta\phi}{\tau_p}}+\frac{\phi_o\delta G_p}{G_{po}\tau_p}-\cancel{\frac{\delta\phi}{\tau_p}}\right]. &(151)
\end{aligned}$$

Next, since the peak gain $G_p\left(n\right)$ is a function of n, we can express the small signal gain amplitude δG_p in terms of δn:

$$\delta G_p = G_p'\delta n, \tag{152}$$

where the prime represents

$$G_p' = \frac{dG_p}{dn}. \tag{153}$$

G_p' is also known as the differential gain because represents the derivative of gain with respect to carrier concentration. Substituting this into equation (151),

$$j\omega\delta\phi = \left(\frac{G_p'}{G_{po}}\right)\left(\frac{\phi_o}{\tau_p}\right)\delta n. \tag{154}$$

Next, we can substitute the time dependence of $\frac{dn}{dt}$ into equation (147), along with (140) to get:

$$\begin{aligned}
j\omega\delta n &= \frac{\delta I}{V_\gamma q}-\left(\frac{c}{n_{\textsf{eff}}}\right)\left(G_{po}\delta\phi+\phi_o\delta G_p\right) &(155)\\
&= \frac{\delta I}{V_\gamma q}-\frac{\delta\phi}{\Gamma\tau_p}-\left(\frac{\phi_o}{\Gamma\tau_p}\right)\left(\frac{G_p'}{G_{po}}\right)\delta n &(156)\\
\delta n\left[j\omega+\left(\frac{\phi_o}{\Gamma\tau_p}\right)\left(\frac{G_p'}{G_{po}}\right)\right] &= \frac{\delta I}{V_\gamma q}-\frac{\delta\phi}{\Gamma\tau_p} &(157)\\
\delta n &= \frac{\frac{\delta I}{V_\gamma q}-\frac{\delta\phi}{\Gamma\tau_p}}{j\omega+\left(\frac{\phi_o}{\Gamma\tau_p}\right)\left(\frac{G_p'}{G_{po}}\right)}. &(158)
\end{aligned}$$

Substituting (158) into (154) results in:

$$j\omega\delta\phi = \left(\frac{G_p'}{G_{po}}\right)\left(\frac{\phi_o}{\tau_p}\right)\left[\frac{\frac{\delta I}{V_\gamma q}-\frac{\delta\phi}{\Gamma\tau_p}}{j\omega+\left(\frac{\phi_o}{\Gamma\tau_p}\right)\left(\frac{G_p'}{G_{po}}\right)}\right]. \tag{159}$$

Now the entire equation is in terms of ϕ only. We can divide by $\delta\phi$ to get:

$$j\omega = \left(\frac{G_p'}{G_{po}}\right)\left(\frac{\phi_o}{\tau_p}\right)\left[\frac{\frac{\delta I/q}{\delta\phi V_\gamma}-\frac{1}{\Gamma\tau_p}}{j\omega+\left(\frac{\phi_o}{\Gamma\tau_p}\right)\left(\frac{G_p'}{G_{po}}\right)}\right]. \tag{160}$$

Then we can re-arrange the terms:

$$\begin{aligned}
-\omega^2+j\omega\left(\frac{\phi_o}{\Gamma\tau_p}\right)\left(\frac{G_p'}{G_{po}}\right)+\left(\frac{G_p'}{G_{po}}\right)\left(\frac{\phi_o}{\tau_p}\right)\left(\frac{1}{\Gamma\tau_p}\right) &= \left(\frac{G_p'}{G_{po}}\right)\left(\frac{\phi_o}{\tau_p}\right)\left(\frac{\delta I/q}{\delta\phi V_\gamma}\right) &(161)\\
-\omega^2\left(\frac{\tau_p}{\phi_o}\right)\left(\frac{G_{po}}{G_p'}\right)+j\omega\left(\frac{1}{\Gamma}\right)+\left(\frac{1}{\Gamma\tau_p}\right) &= \left(\frac{\delta I/q}{\delta\phi V_\gamma}\right) &(162)
\end{aligned}$$

The ratio of small signal photons to small signal carrier injection rate can therefore be expressed as:

$$\frac{q\delta\phi}{\delta I}=\frac{1}{\left(\frac{V_\gamma}{\Gamma\tau_p}\right)-\omega^2V_\gamma\left(\frac{\tau_p}{\phi_o}\right)\left(\frac{G_{po}}{G'_p}\right)+j\omega\left(\frac{V_\gamma}{\Gamma}\right)}. \tag{163}$$

Since the confinement factor is $\Gamma=\frac{V_\gamma}{V_p}$

$$\frac{q\delta\phi}{\delta I} = \frac{1}{\left[\left(\frac{V_p}{\tau_p}\right)-\omega^2V_\gamma\left(\frac{G_{po}}{G'_p}\right)\left(\frac{\tau_p}{\phi_o}\right)\right]+j\omega V_p} \tag{164}$$

$$\frac{\delta\phi}{\delta I}\left(\frac{qV_p}{\tau_p}\right) = \frac{1}{1-\left(\frac{\omega}{\omega_r}\right)^2+j\omega\tau_p} \tag{165}$$

where we have defined

$$\omega_r^2=\left(\frac{1}{\Gamma}\right)\left(\frac{\phi_o}{\tau_p^2}\right)\left(\frac{G'_p}{G_{po}}\right). \tag{166}$$

ω_r is known as the relaxation frequency of the laser cavity. It is an important paramater because it is entirely controlled by the laser cavity design. It is related to the photon density at the DC bias (ϕ_o), the differential gain (G'_p) as well as other design parameters such as total cavity loss (τ_p) and confinement factor (Γ) and the DC gain (G_{po}).

In equation (166), the photon density ϕ_o is the DC bias value. Hence, it can be more conveniently expressed in terms of the DC bias current above threshold $I-I_{th}$. Using the dc rate equations (138) and (140):

$$\phi_o=\frac{(I_o-I_{th})}{qV_p}\tau_p. \tag{167}$$

Therefore, the relaxation oscillation frequency can also be described in terms of the dc bias condition:

$$\omega_r^2=\left(\frac{G'_p}{G_{po}}\right)\left(\frac{I_o-I_{th}}{qV_\gamma\tau_p}\right). \tag{168}$$

Furthermore, the output optical power is related to the photon density, which we defined earlier via equation (108). Using this, we can get

$$\frac{\delta P}{\delta I}=\frac{\left(\frac{h\nu}{q}\right)\left(\frac{\tau_p}{\tau_{pm1}}\right)}{1-\left(\frac{\omega}{\omega_r}\right)^2+j\omega\tau_p} \tag{169}$$

The term in the numerator is the DC responsivity, $\mathcal{R}$ that we defined in equation (124). Therefore, we can define the AC responsivity r as:

$$r=\frac{\delta P}{\delta I}=\frac{\mathcal{R}}{1-\left(\frac{\omega}{\omega_r}\right)^2+j\omega\tau_p}. \tag{170}$$

Finally, the magnitude of this response can be written as

$$\boxed{|r|=\frac{\mathcal{R}}{\sqrt{\left(1-\frac{\omega^2}{\omega_r^2}\right)^2+\omega^2\tau_p^2}}}. \tag{171}$$

We can see from this expression that the responsivity when $\omega=0$ is exactly equal to the DC responsivity $\mathcal{R}$. At the resonance frequency of ω_r, the responsivity reaches a peak value of $\mathcal{R}/\left(\omega_r\tau_p\right)$. At frequencies higher than ω_r, the responsivity will decline similar to a LED.

Example (Cont'd)

Continuing the example from the previous sections, we can now calculate the modulation frequency response of the laser. We had $\alpha_c = 25.4$/cm, and $n_{\text{eff}} = 3.5$. From this, we can get

$$\tau_p = \frac{n_{\text{eff}}}{c\alpha_c} \tag{172}$$

$$= 4.58 \text{ ps}. \tag{173}$$

We will assume a DC bias current of 250 mA. The calculated threshold current was 39 mA, $V_\gamma = 150\mu\text{m}^3$, and $\Gamma = 0.05$. Next we need to evaluate the ratio between the gain at the DC bias point, G_{po} and the differential gain G'_p. For this, we refer to the calculate gain vs carrier density plot in Fig 18. At 250 mA, we will be well above threshold. The threshold carrier density was previously calculated as 3.3×10^{18} cm^{-3}. The gain value corresponding to this carrier density is 508/cm. The slope at this bias point can also be calculated from the gain data, which is 1.7×10^{-18}cm^3.

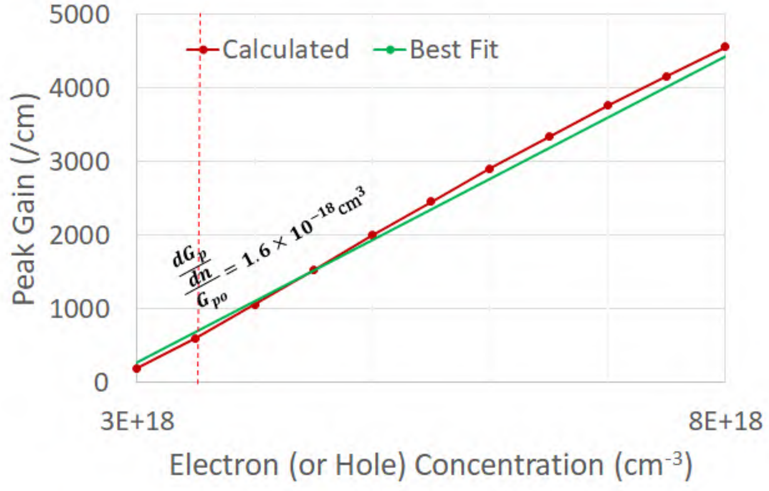

Figure 22: Peak gain vs carrier concentration and the differential gain at the bias point.

Using these values, we can calculate the relaxation oscillation angular frequency by utilizing equation (168):

$$\omega_r = \sqrt{\left(\frac{G'_p}{G_{po}}\right)\left(\frac{I_o - I_{th}}{qV_\gamma \tau_p}\right)} \tag{174}$$

$$= 56.9 \times 10^9 \text{ rad/s}, \tag{175}$$

which corresponds to a frequency of

$$f_r = 8.8 \text{ GHz}. \tag{176}$$

This can often be considered the useful bandwidth of the laser. However, that does not prevent the use of this laser at higher modulation frequencies.

The relaxation oscillation frequency is a resonance effect between the carrier density and the photon density, coupled through the gain coefficient. At this resonance frequency, we can see an increase in responsivity. Beyond this frequency, the responsivity will decline rapidly, primarily due to the photon lifetime τ_p. In cases where $1/\tau_p$ is significantly smaller than the ω_r,

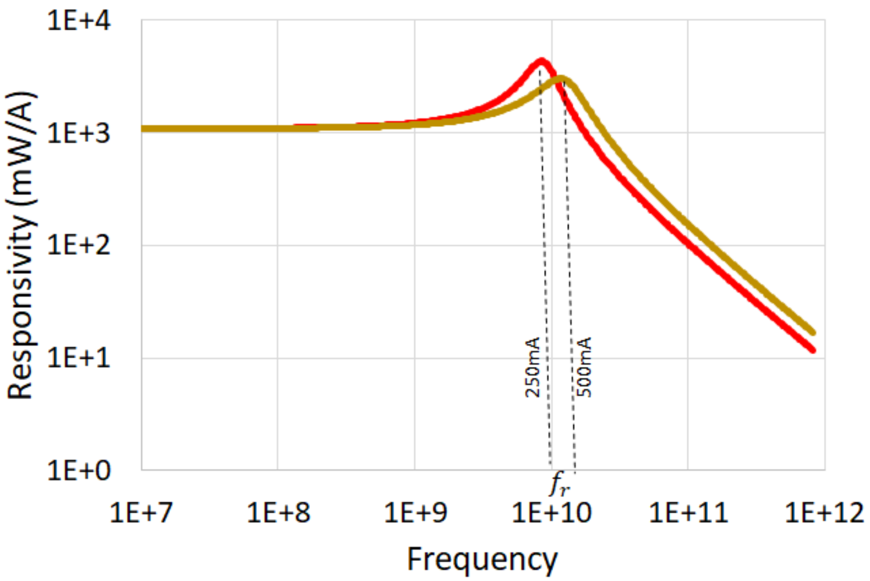

Figure 23: Calculated dynamic responsivity vs modulation frequency for 250mA and 500mA bias.

the resonance effect may not be seen at all. In that sense, τ_p can be seen a contributing to a damping effect to a resonance frequency of ω_r.

Factors Affecting the Modulation Frequency of Lasers

In LED's we saw that the only parameter that determined the maximum modulation frequency was the carrier lifetime τ. This is a basic material property. Therefore, it is not possible to increase this frequency by design methods. In laser diodes, the maximum modulation frequency is determined by a number of design factors that can be engineered.

The photon lifetime is the primary factor that affects the modulation frequency. This is determined by α_c, which is a combination of intrinsic losses, mirror reflectivities and the cavity length. A higher value for α_c (lossier cavity) would result in a higher modulation bandwidth. However, this does not mean that a lossier cavity is necessarily better. The trade-off between intrinsic losses and output mirror loss has to be managed such that a low photon lifetime τ_p is achieved without compromising the external quantum efficiency or responsivity. For example, simply increasing the absorption and scattering losses $\alpha_a + \alpha_s$ will reduce the external quantum efficiency. However, reducing the cavity length and reducing output mirror reflectivity will result in a smaller τ_p without degrading the quantum efficiency. However, this would also increase the threshold current, so obviously this process requires a number of design considerations that must be weighed carefully.

The relaxation oscillation frequency is also an important factor that contributes to the overall modulation bandwidth. This can be increased, for example, by operating the laser at a higher bias current it is possible to increase the frequency response. Its effect is illustrated in Fig 23, where we have shown two curves corresponding to a bias current of 250 mA and 500 mA (threshold current is 41 mA).

The differential gain factor $\frac{G_p'}{G_{po}}$ also plays an important role. Although this is primarily a material effect, it too can be modified by engineering the bandstructure of the gain material. Quantum well lasers, for example, exhibit a step-like density of states function instead of the parabolic function described in Fig 13. This results in a larger change in Fermi level for a given change in carrier density, resulting in a higher differential gain factor. Combining all of these effects, it is possible to design diode laser structures that exhibit modulation frequencies well into the GHz range with low threshold currents. These are commonly used in fiber optic communication systems.

Diode Laser Configurations

Diode lasers are fabricated by growing epitaxial thin films on a lattice-matched substrate. Therefore, all the layers that make up the device will be planar in nature. However, the resonator can be configured to exhibit two distinctly different geometries edge-emitting configuration, or vertical-emitting configuration. Within each configuration, there are a number of variations. Some of these are discussed below.

Broad-Area (BA) Lasers

This is the simplest and most used semiconductor laser configuration for general applications. As discussed in Fig 7, this is a Fabry-Perot resonator consisting of two cleaved facets that form the mirrors. Also, as discussed in Fig 10, the optical field will be confined by the separate confinement heterostructure (SCH) while the electrons will be confined within the gain region (active region). The SCH is generally designed to support a single optical mode in order to increase the optical confinement factor Γ. However, in the lateral direction there is no optical confinement (Fig 24). The width can range from a few tens of microns to hundreds of microns. The gain volume V_γ and the optical volume V_p will scale with the width, while the confinement factor Γ will remain nearly the same. The large volume allows high optical powers to be obtained from this laser.

One of the main drawbacks of this structure is the lack of lateral beam control. The lateral direction can be considered as a multi-mode waveguide with a very large number of optical modes. Each mode will have a different field distribution. The output beam will therefore consist of many different optical modes, each of them incoherent with one another. As a result, the beam will have poor spatial coherence. When focused in an optical system, each mode will come to a focus at different points in space, making it difficult to achieve diffraction-limited performance. Furthermore, due to the lack of a dielectric waveguide structure in the lateral direction, the mode distribution will be highly influenced by the temperature profile and carrier distribution, both of which affect the refractive index. This can result in an unstable beam that can shift and change shape depending on the operating conditions. The light-current curve will exhibit discontinuities, and effects such as self-focusing and hot spots (also known as filamentation) can create instabilities.

In the spectral domain, the output will contain a large number of Fabry-Perot modes, similar to the one shown in Fig 6. However, since each lateral mode will have its own effective index n_{eff} and group index n_g, this will result in a less defined distribution of spectral emission lines. Nevertheless, the location of these peaks will be largely determined by the location of the gain

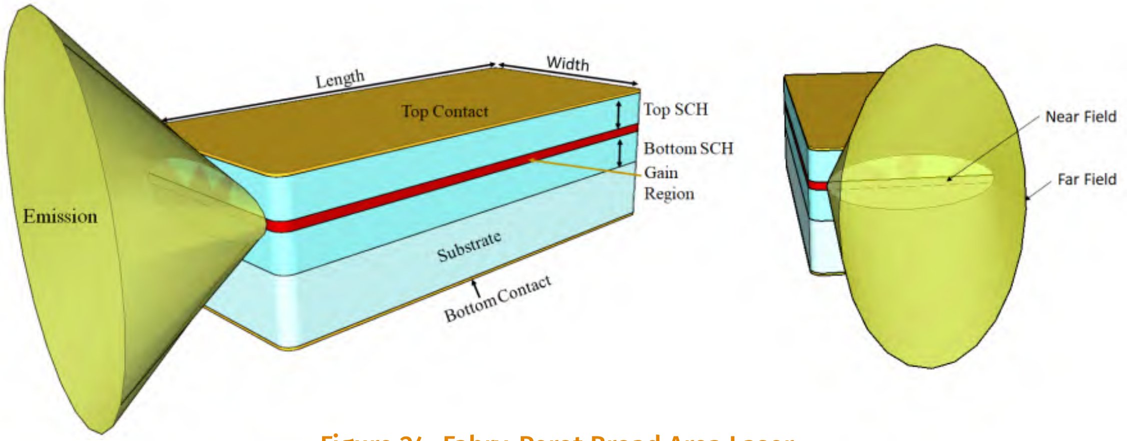

Figure 24: Fabry-Perot Broad Area Laser

peak, which is primarily determined by the material bandgap and population inversion.

The output beam will generally be thin (coincident with the single-mode SCH layer), and wide (due to the large width). This results in the bean being significantly elliptical in the near-field (thin and wide). In the far field, this translates to a larger divergence in the vertical direction and a smaller divergence in the horizontal direction, as illustrated in Fig 24.

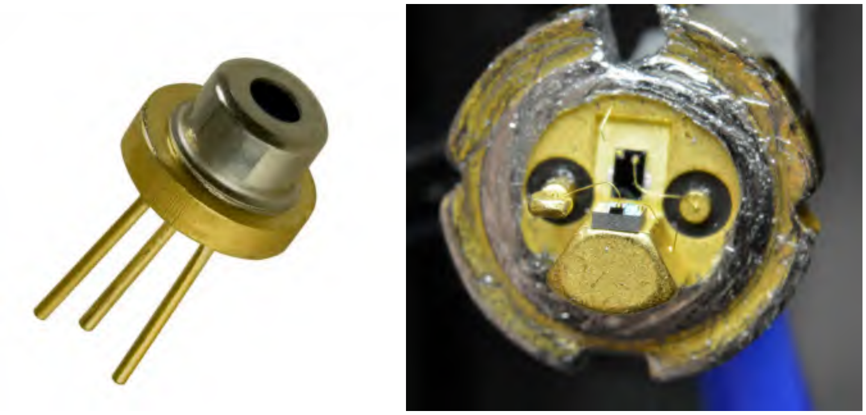

Figure 25: TO package laser with the cover cut away. Source: wikiwand.com

Broad-area lasers are inexpensive to manufacture (compared to the other types of lasers discussed next). They can be thought of as an LED with much higher extraction efficiency, and a much higher brightness (smaller beam divergence). They are used where high optical powers are required with less stringent requirements on the spatial coherence of the beam. Example applications include pumping other lasers, such as YAG and fiber lasers. Laser pointers, flash ladar as well as numerous other illumination applications also utilize broad-area lasers.

Waveguide Lasers

Lateral waveguide mode control can be introduced into a laser chip by a number of different techniques. One technique is to randomize the crystal structure in all areas except the waveguide region. This is done by ion implantation. Ions are driven into the semiconductor laser

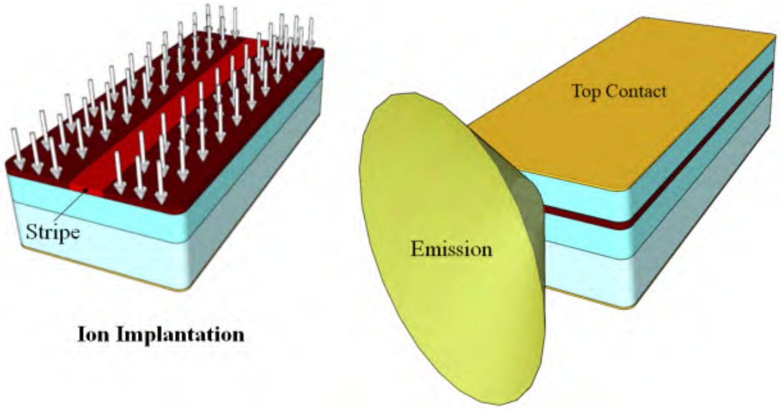

Figure 26: Ion Implanted Gain Guided Waveguide Laser

which cause impact damage, as shown in Fig 26. Therefore, these areas will not be able to produce optical gain. Only the central striped region will be able to support lasing. This is referred to as a gain-guided laser structure, because technically the optical mode is not defined by a dielectric discontinuity, but instead by a gain discontinuity. Nevertheless this results in a reasonably well defined lateral mode control. However, since the outside regions will now have a higher scattering and absorption loss, the overall cavity loss α_c will become higher, leading to a lower external quantum efficiency.

Another type of gain guidance can be achieved by patterning the top electrical contact as a narrow stripe. This will produce current injection along a narrow stripe, and hence population inversion will be achieved only within this stripe. This leads to a similar effect to that of the ion implanted structure. However, the carriers can diffuse laterally, which will broaden the stripe width which makes it difficult to precisely control the lateral mode. Carrier diffusion can also vary with injection conditions, which will make the beam profile dependent on the current level. Nevertheless, since there is less crystal damage in this structure compared to ion implantation, so it can lead to an improved external quantum efficiency.

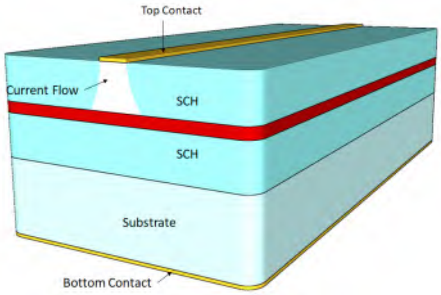

Figure 27: Striped Contact Gain Guided Laser

A better method of controlling the lateral mode is by producing a strong dielectric contrast. This can be achieved with a ridge waveguide configuration. The top SCH region can be etched during fabrication to provide a relatively strong mode control in the lateral direction. Because the primary mode control is by the refractive index geometry, it is relatively immune from instabilities due to varying currents and temperatures. Nevertheless, it still has some aspects of gain guidance because the bottom SCH is unpatterned. Hence,

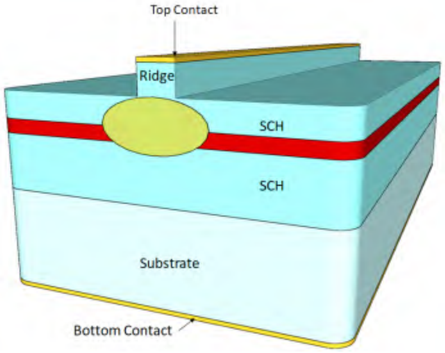

Figure 28: Ridge Waveguide Laser

the spatial profile of the carrier injection does have some role in the optical mode.

Far Field Profile of Edge Emitting Lasers

All semiconductor lasers are fabricated as epitaxial thin film structures. During growth, it is possible to precisely control the film thicknesses down to a few nanometers. As a result, the SCH layer structure can be easily made to support a single optical mode. The lateral geometry, on the other hand, is achieved by lithographic patterning, which has a much worse resolution, on the order of a micron. As a result the waveguides are nearly always wider than their height. It is common for a waveguide structure to have 0.5μm thickness and 10μm width. This results in a near-field beam profile that is highly elliptical, as illustrated in Fig 28. This is a general characteristic that is observed in nearly all edge-emitting lasers except vertically emitting lasers (such as VCSELs). This characteristic also leads to an elliptical divergence pattern that is fast in the vertical axis and slow in the horizontal axis. These are noted as $\theta_{||}$ and $\theta_{\perp}$. Typical numbers are $\theta_{||} = 9°$ and $\theta_{\perp} = 17°$ (from Thorlabs L840P200 840nm Laser Diode). Working with such beams require astigmatic lenses, which are expensive and difficult to align.

A circular beam shape will be far more attractive, but it is difficult to achieve using an edge-emitting configuration. However, in surface-emitting configurations where the beam output is normal to the surface of the semiconductor, it is possible to get a nearly circular beam. The Vertical Cavity Surface Emitting Laser (VCSEL) is an example of this.

Distributed Bragg Reflector (DBR) Lasers

As discussed previously, the SCH waveguide structure is typically designed to be single mode. Even if the lateral structure is designed to be single mode (as in the ridge waveguide laser), in the longitudinal direction there will be a large number of modes (thousands) spaced by $\Delta\lambda = \frac{\lambda^2}{2n_g L}$. Which of these modes will actually lase will be determined by the gain spectrum of the semiconductor. Typically a few dozen modes close to the gain peak will be able to reach their threshold gain. In some special circumstances a single longitudinal mode may lase, but it not be dynamically stable. It will quickly switch to another mode as conditions change, such as current or temperature. Hence these are referred to as multi-mode lasers.

To achieve a a spatially and spectrally coherent beam, one needs to achieve not only a single lateral mode, but also single longitudinal mode. This is necessary in some communication applications, especially those that employ coherent detection techniques. It is obviously not possible to achieve this with a Fabry-Perot laser structure. Reducing the cavity length to a sub-wavelength regime is clearly not possible because it will raise the threshold gain to unrealistically large values (see equation (30)). However, if we can replace the mirror with a frequency-selective mirror, then we can at least reduce the number of longitudinal modes. This is the main idea behind DBR lasers.

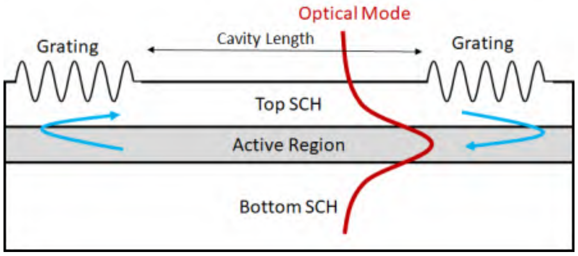

Figure 29: First Order Distributed Bragg Reflection (DBR) Laser

Distributed Bragg Reflectors (DBR) use a periodic structure (grating) to generate reflection. This is a fundamentally different way of producing reflection compared to a discrete mirror. It is essentially a diffraction grating, but it is designed such that the first order diffraction propagates backwards to act like a reflected beam. Two gratings are etched into the SCH regions of the waveguide on either end of the cavity, as shown in Fig 29.

Part of the waveguide mode will sample the grating and be reflected back into the cavity similar to a discrete mirror. This effect can be conveniently depicted in a scattering diagram as shown in Fig 30. All of the possible scattered modes are represented by a circle whose radius is the propagation constant β of the waveguide mode. The forward traveling waveguide mode is on the positive axis with a value of $+\beta$, and the backward traveling waveguide mode is on the negative axis with a value of $-\beta$. The grating etched into the waveguide couples the two modes together. The coupling vector K is the distance in k-space between $+\beta$ and $-\beta$. Therefore, the grating vector has to be

$$K = 2\beta. \tag{177}$$

Since $K = \frac{2\pi}{\Lambda}$ where Λ is the grating pitch, and $\beta = \frac{2\pi}{\lambda} n_{\text{eff}}$ where λ is the wavelength and n_{eff} is the effective index of the waveguide mode, we can get

$$\Lambda = \frac{\lambda}{2n_{\text{eff}}}. \tag{178}$$

For example, consider a GaAs waveguide with $\lambda = 850$nm, $n_{\text{eff}} = 3.4$. The required grating pitch that will produce a reflection in the waveguide will be 123.5nm.

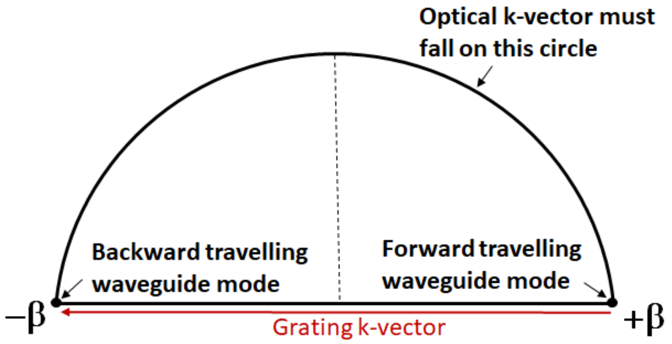
Optical k-vector must fall on this circle
Backward travelling waveguide mode
Forward travelling waveguide mode
−β
+β
Grating k-vector

Figure 30: Scattering diagram of a first order diffraction from a grating producing reflection

Alternatively, we can also create a long pitch grating (smaller K) such as $K = \beta$. This will produce a guided mode that falls exactly at the center of the circle. However, since the wave vector of the optical wave has to be preserved, the diffracted field that corresponds to this mode will appear at the top of the circle, as depicted in Fig 31. This is a radiation mode that will be directed upwards from the waveguide.

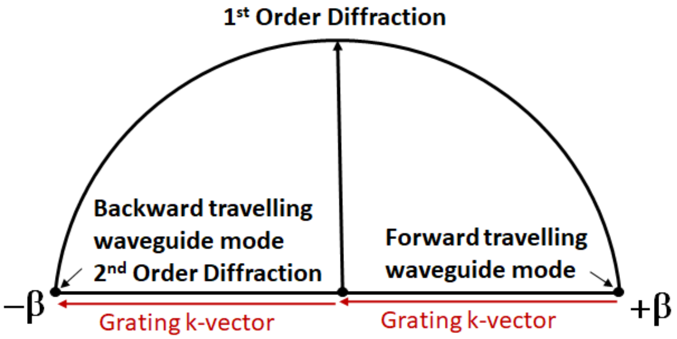

Figure 31: Scattering diagram of a second order grating producing the reflection feedback

While the first order is directed vertically away from the grating, the second order diffraction will lead to the backward propagating mode with $-\beta$. Therefore, in this case, there will be two diffraction orders. The first diffracted order will couple light away from the grating normal to the surface. The second order will couple with the reflected mode. Therefore, this grating can be considered as a combined reflector/out-coupler, as depicted in Fig 32.

The strength of reflection from this grating is related to the refractive index contrast between the high and low regions of the grating, Δn_{eff}, and the length of the grating. The principle involves the application of coupled-mode theory, which we will not pursue here, but the reflection can be approximately represented as

$$R = \tanh^2(\kappa L)\,, \tag{179}$$

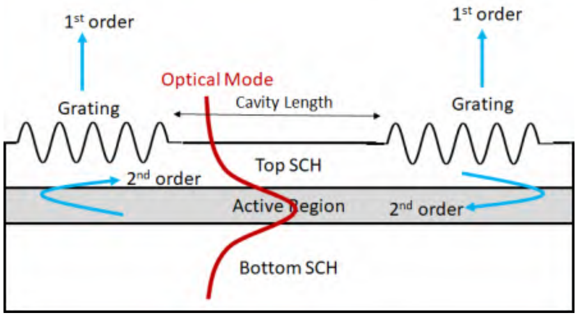

Figure 32: Second Order Distributed Bragg Reflection (DBR) Laser

where κ is the coupling coefficient which is related to the effective index contrast as

$$\kappa = \frac{\pi \Delta n_{\text{eff}}}{\lambda}, \tag{180}$$

and L is the length of the grating.

For example, if small grooves are etched into the top cladding (SCH) region such that $\Delta n_{\text{eff}} = 0.01$, and the length of the grating is 200μm, we can get a maximum reflection of 99.6% at the Bragg wavelength (the Bragg wavelength is defined by the condition that satisfies equation (178)). More importantly, this reflection is a strong function of wavelength. The reflection value of $\tanh^2 \kappa L$ occurs at the Bragg wavelength, but it quickly declines at other wavelengths. The bandwidth of reflection depends on the coupling coefficient κ. For example, the reflection bandwidth of a 200μm long grating with $\Delta n_{\text{eff}} = 0.01$ designed for a Bragg wavelength of 850nm is shown in Fig 33 (which was derived from coupled-mode theory). We can see that the peak reflection of 99% exists only within a narrow spectral bandwidth of around 2nm, and declines rapidly on both sides. The phase of the reflection is also shown on the right hand axis of Fig 33. Within the reflection band, the phase gradually changes from 0 to $\pm\frac{\pi}{2}$ from one edge of the reflection band to the center.

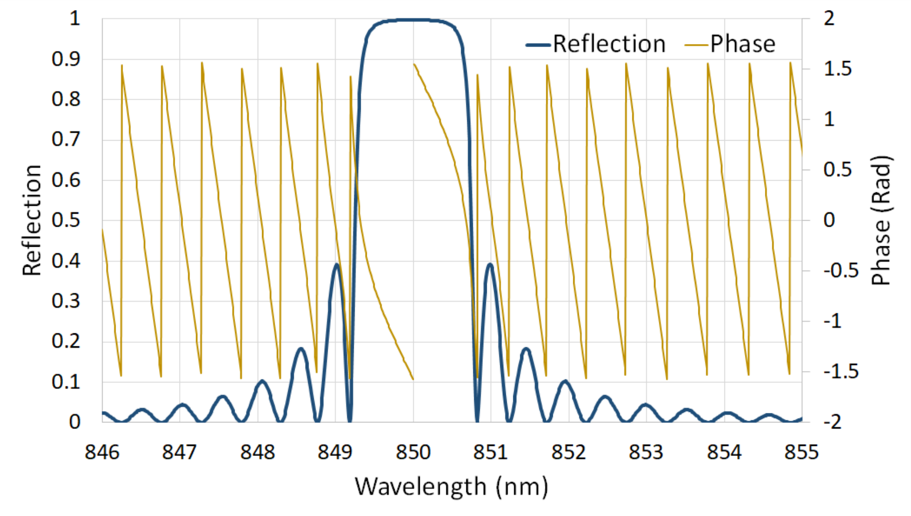

Figure 33: DBR Reflection Spectrum

DBR lasers can be analyzed by using the Fabry-Perot theory developed earlier. The threshold gain can be calculated using the same equation (55):

$$G_{th} = \frac{1}{\Gamma}\alpha_c \quad (181)$$

$$= \frac{1}{\Gamma}\left(\frac{1}{2L}\ln\left(\frac{1}{R_1R_2}\right)+\alpha_s+\alpha_a\right). \quad (182)$$

The only difference is, R_1 and R_2 are now due to the DBR mirrors, but otherwise the approach is identical to that of the Fabry-Perot cavity. An important difference is, however, only a very small number of longitudinal modes will satisfy the round trip phase condition due to the narrow reflection band. The round trip phase must also include the reflection phase from the DBR, which is not constant with wavelength as depicted in Fig 33.

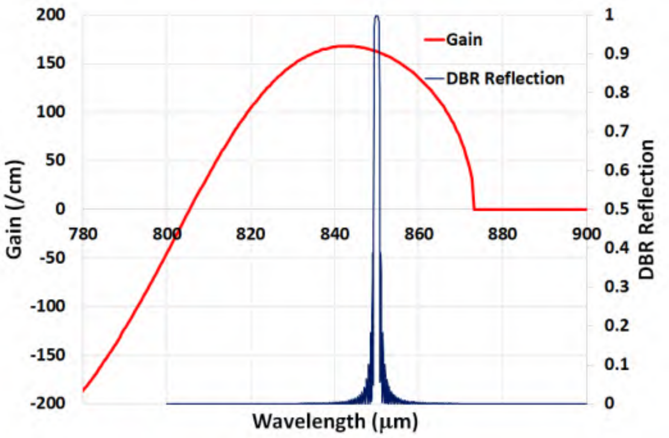

Figure 34: Semiconductor Gain Spectrum for GaAs Combined with the DBR Reflection Spectrum

For example, assuming a cavity length of 300μm, $n_g = 4.5$ gives a $\Delta\lambda = 0.27$ nm, so only about four or five longitudinal modes will fit within the reflection band. To illustrate this, we can overlay the DBR reflection spectrum from Fig 33 with the previously calculated gain spectrum (Fig 17). This is shown in Fig 34. We can clearly see that the reflection exists only over a very small portion of the gain spectrum. In addition to reducing the number of longitudinal modes, this has additional benefits as well. In the case of Fabry-Perot lasers, since the mirror reflectivity is flat across all wavelengths, the the lasing modes will always be close to the gain peak. However, small changes in current or temperature can cause the gain spectrum to fluctuate. As a result, the number of lasing modes and their wavelengths will also fluctuate with these changes. In the case of DBR lasers, the lasing modes will remain fixed by the DBR reflection regardless of these fluctuations. The reflection spectrum of the DBR is determined only by the grating pitch and the effective index of the waveguide. Although these parameters can also experience some changes due to temperature or current, they are far more stable than the gain spectrum.

Distributed Feedback (DFB) Lasers

Distributed Feedback (DFB) lasers can be considered as a specific case of DBR lasers with the cavity length (in Figs 29 and Fig 32) is reduced to zero. In other words, the entire laser cavity consists of just a grating structure. Although a zero cavity length may appear to be unusual, in terms of phase, a cavity whose

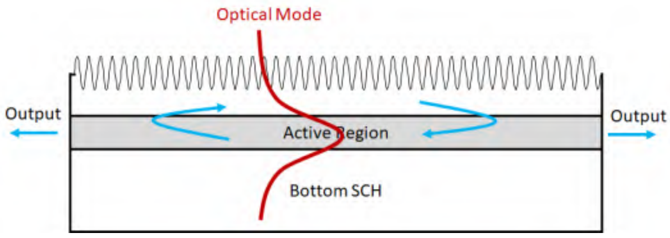

Figure 35: Distributed Feedback (DFB) Laser Cavity

length is zero is not really much different than a cavity whose phase length is a multiple of 2π. With a zero cavity length, the round trip phase condition will be determined entirely by the reflection phase of the DBR reflectors. From Fig 33, we can see that the reflection phase at the Bragg wavelength (850nm in this case) is $\pm\pi/2$ (either sign has the same effect on the field). After two reflections, this will become a phase of $\pm\pi$. Therefore, the round-trip phase will not be a multiple of 2π at the Bragg wavelength. As a result, unlike the DBR laser, no lasing will be observed at the Bragg wavelength. If we move away from the Bragg wavelength, either to longer, or shorter wavelengths, we can find a point where the reflection phase is equal to zero. In Fig 33, these points coincide with the reflection reaching zero. These are the resonance points of the DFB (single grating) cavity.

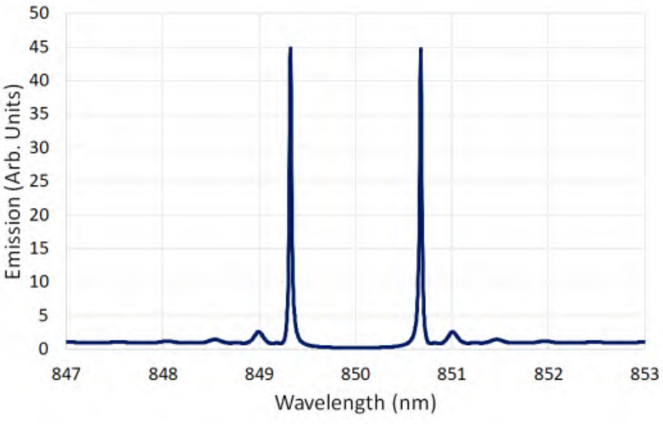

Figure 36: Typical lasing spectrum from a DFB laser

It may also appear odd that lasing can occur where the reflection from the DBR is zero. This is opposite of the DBR lasing condition illustrated in Fig 34. However, the reflection we are referring to in this case is the net reflection from the whole cavity, not from one mirror alone. If we refer to the cavity resonance conditions we examined earlier (equations (**??**) - (**??**)), we noted that the reflection from the overall system dropped to a minimum at the resonance wavelengths, and to exactly zero in the case of a symmetric cavity ($R_1 = R_2$). We can consider the DFB cavity as a symmetric system if we split the grating in half and consider the cavity as being in the center. This is the reason why the zero reflection points from the whole grating structure also correspond to their resonances. It is also possible to calculate the Q-factor of the cavity and the corresponding photon lifetime, but that analysis is beyond the scope of this discussion.

Fig 36 shows the typical emission spectrum from a DFB laser. We can identify two lasing peaks corresponding to the two zero reflection points of the grating. The distance between the two lasing peaks is equal to the reflection bandwidth of the grating, which is typically larger than the Fabry-Perot line spacing (equation 18). The symmetry in the emission spectrum arises from the grating having an exactly symmetric configuration. However, in practice, the structure is never perfectly symmetric. The shape of the gain spectrum will generally not be symmetric about the Bragg wavelength. Additionally, the termination of the grating on one end will not be identical to the other end. The crystal facets will also contribute to the cavity resonance, and they are unlikely to be identical. As a result, only one emission line will be observed in practice, although it is difficult to predict which one. This may not be acceptable in some applications that require a very strict wavelength control, such as in spectroscopy or wavelength division multiplexed (WDM) communication systems. Hence some modifications can be done to improve this design and make it a truly single-frequency laser.

Quarter-Wave Shifted Distributed Feedback (QW-DFB) Lasers

This is a modified version of the standard DFB laser where a tiny cavity is introduced at the center of the grating to create a phase shift of $\frac{\pi}{2}$ (quarter-wave) at the Bragg wavelength. This is shown in Fig 37. The round-trip phase from this cavity will be π. The effect of this phase shift will be to exactly offset the reflection phase of π discussed in Fig 36. As a result, the resonance moves from the two points on either side of the reflection band to the central Bragg wavelength.

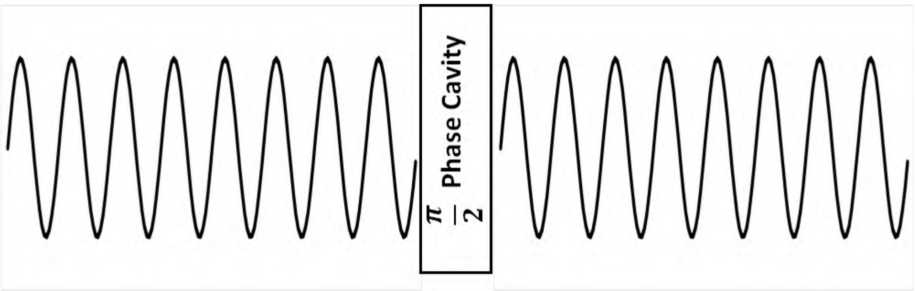

Figure 37: Quarter-wave shifted DFB grating structure

The typical emission spectrum from a QW-DFB is shown in Fig 38. We can see that the emission line has now moved to exact center of the reflection band, at 850 nm, making it a truly single-frequency emission. These lasers are ideal for applications requiring precise and narrow emission wavelengths. They are also known as dynamically single-frequency lasers. This is because any drift in the gain spectrum due to temperature shifts will not change the emission wavelength. The emission wavelength is determined only by the grating period and the refractive index. Nevertheless, changes in temperature can and do affect the refractive index and the grating period (due to thermal expansion), but these are relatively minor effects compared to the behavior of the gain function.

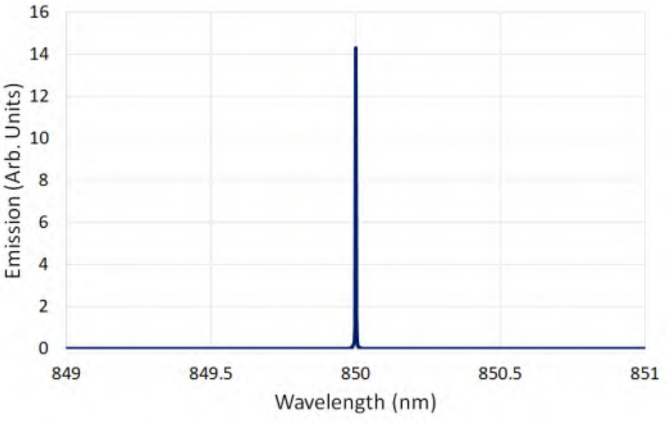

Figure 38: Typical lasing spectrum from a QW-DFB laser

Vertical Cavity Surface Emitting Laser (VCSEL)

The usual configuration of edge emitting lasers is such that all the feedback reflections occur parallel to the substrate, making it a horizontal cavity. The vertical cavity surface emitting laser (VCSEL) is essentially the same as the quarter-wave shifted DFB laser laying on its side. The feedback occurs normal to the substrate instead of horizontally, and the emission is also normal to the substrate. However, surface-emission is not the only unique characteristic in this laser cavity. While the grating in a DFB laser is produced by etching a sinusoidal feature into the films that make up the waveguide, in a VCSEL, the grating is produced by alternating layers of thin films. Hence the grating in a VCSEL is produced during crystal growth rather than lithographic patterning. This makes the VCSEL tremendously more economical and flexible comapred to edge emitting lasers.

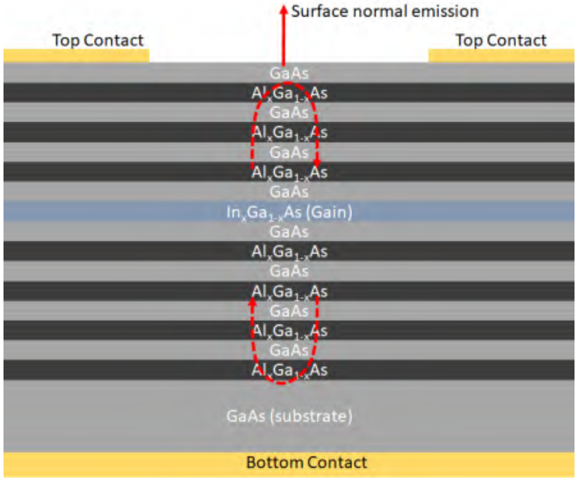
Surface normal emission
Top Contact
Top Contact
GaAs
Al_xGa_{1-x}As
GaAs
Al_xGa_{1-x}As
GaAs
Al_xGa_{1-x}As
GaAs
In_xGa_{1-x}As (Gain)
GaAs
Al_xGa_{1-x}As
GaAs
Al_xGa_{1-x}As
GaAs
Al_xGa_{1-x}As
GaAs
Al_xGa_{1-x}As
GaAs (substrate)
Bottom Contact

Figure 39: A generic VCSEL structure on a GaAs substrate using GaAs/Al$_x$Ga$_{1-x}$As layers and a quarter-wave In$_x$Ga$_{1-x}$As gain layer.

A generic VCSEL structure is shown in Fig 39. In this example, the substrate is GaAs. A number of GaAs/Al$_x$Ga$_{1-x}$As layers are grown on the substrate to produce the grating structure similar to a DFB. The fact that all compositions of Al$_x$Ga$_{1-x}$As are lattice matched to GaAs is what makes this possible. The grating structure is interrupted near the center of the stack with a quarter-wave thick In$_x$Ga$_{1-x}$As layer followed by more layers of GaAs/Al$_x$Ga$_{1-x}$As on top. The In$_x$Ga$_{1-x}$As layer provides the quarter-wave shift to the DFB as well as the optical gain. Since the bandgap of In$_x$Ga$_{1-x}$As is smaller than either GaAs or Al$_x$Ga$_{1-x}$As, most of the carriers will accumulate in this layer. Though it is not perfectly lattice-matched to GaAs, it is still possible to grow this layer as long as it is relatively thin. Furthermore, the top Bragg reflector is generally doped p-type and the bottom stack will be doped n-type, creating a vertical PN junction. This allows the structure to behave as a double heterostructure. The separate confinement heterostructure (SCH) is not really necessary here because the optical field in the vertical direction is largely confined by the grating. The current is injected into the structure with a ring contact at the top, and a uniform contact at the bottom. The ring contact is necessary to allow the emission to exit the structure without being obstructed by the metal.

The peak reflection of the DBR mirrors due to the alternating layers can be shown to be

$$R = \left| \tanh\left[m \ln\left(\frac{n_1}{n_2} \right) \right] \right| \tag{183}$$

where m is the number of layer pairs, and n_1 and n_2 are their refractive indices. This equation is similar to equation (179), except κL has been replaced with $m\left(\frac{n_1}{n_2}\right)$, which is more accurate for gratings with high index contrasts. Typically, the cavity structure is designed with more layers in the bottom stack than the top stack to allow most of the emission to come out of the top side. For example, assuming 30 pairs of layers (60 layers) for the bottom stack, assuming the refractive index of Al$_x$Ga$_{1-x}$As is 2.95 and the refractive index of GaAs is 3.45, we can calculate the reflection to be 0.9998. With 20 layer pairs (40 layers) for the top stack, the reflection will be 0.9962. It is necessary for these reflection values to be very high because the optical confinement factor Γ_v is very small in these structures. This was defined earlier in equation (50) as the ratio between the gain volume and the optical field volume. In this case, gain is provided only by the central In$_x$Ga$_{1-x}$As layer (in fact, it is even smaller than the full layer thickness because quantum wells embedded inside the quarter-wave shift layer are often used for producing the gain). As a result, the optical field will be significantly larger than the gain volume, resulting in a very small value for Γ_v. In order to keep the threshold gain reasonably low, the mirror losses, which are the dominant component of loss in VCSELs, have to be kept small by increasing their reflectivity. Assuming the gain region is 20 nm thick, the mirror loss becomes

$$\alpha_m = \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2} \right) = 1000/\text{cm}. \tag{184}$$

This gain is relatively high. Hence, alternative methods to increase the optical gain is necessary. This is accomplished, at least partially, through the use of quantum wells (QW) as the active region. The density of states of quantum wells is different than a bulk material, and this makes it possible to achieve a significantly higher gain value for the same carrier density as compared to a heterostructure. The trade-off, however, is the small thickness of quantum wells (20 nm in the above example). In the case of VCSELs, L will be small, leading to a high threshold gain (1000/cm in the above example). In the case of edge-emitters, the thin quantum wells will produce a very small cross-sectional confinement factor Γ_A, also leading to a high threshold gain. This can be offset, to some extent, by utilizing multiple quantum wells (MQW). By repeating the quantum well structure many times, it is possible to make it appear to be thicker. This has become the standard configuration in many diode laser structures, and is often abbreviated as MQW lasers.

Just like with edge-emitting lasers, the generic VCSEL structure shown in Fig 39 has no transverse mode confinement. As a result, it will behave like a broad-area edge-emitting laser with numerous transverse modes. This, of course, is undesirable, so a number of techniques are utilized to define a transverse structure in VCSELs.

A common technique is to use ion implantation to create a cylindrical gain-guided structure, as shown in Fig 40. This is similar to the edge-emitting structure in Fig 26. A portion of the upper Bragg reflector is implanted with high energy ions such that the conductivity in those areas is lowered due to crystal disruption. This acts as a funnel to redirect the current through the central portions. As a result, gain will be localized to the central portions only. Therefore, the transverse mode will be defined entirely by the gain profile. However, the crystal damage from ion implantation will increase the nonradiative recombination rates in the implanted regions. This will result in an increase in the threshold current and decrease in the overall quantum efficiency. Nevertheless, this is a commonly used technique that is relatively inexpensive to implement.

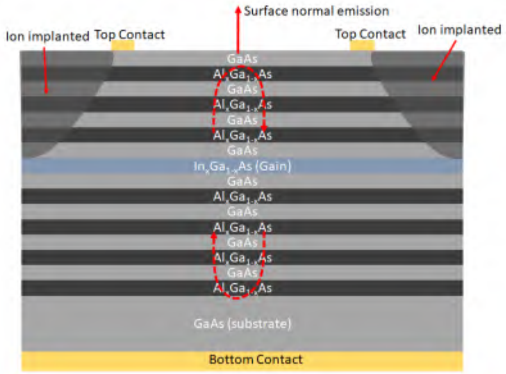

Figure 40: Implant-disrupted VCSEL structure

An altertive technique is to etch a portion of the upper Bragg reflector to define a cylindrical waveguide, as shown in Fig 41. This overcomes several problems in the ion implanted structures, but still only provides a partial solution because the cylindrical waveguide only exists in the upper reflector. The gain layer and the bottom reflector are unrestricted. Therefore, there will be diffraction losses due to the optical mode mismatch between these regions. Additionally, current can flow very close to the etched side walls, producing surface recombination and leakage.

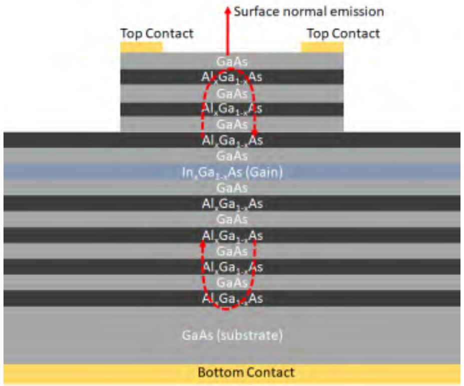

Figure 41: Etched-post VCSEL structure

By far the most commonly used technique in VCSELs is known as oxide-aperturing. This is done by designing one or a few layers of the upper Bragg reflector with high aluminum content Al$_x$Ga$_{1-x}$As layers (which will have an indirect bandgap when $x > 0.45$). After etching the cylinder, the exposed Al$_x$Ga$_{1-x}$As layers are partially oxidized in an oxygen atmosphere at elevated temperatures. This oxidation turns the Al$_x$Ga$_{1-x}$As layer into a dielectric, greatly reducing its refractive index and turning it into an insulator. The timing of the oxidation step is carefully controlled such that only the outer portions of the film is oxidized. The resulting refractive index profile allows better control of the transverse optical mode. The oxide aperture also keeps the current within the central region. Nevertheless, the bottom Bragg reflector is largely unconfined. Although in principle it is possible to etch through the entire structure, in practise, etching through the active region often results in large nonradiative recombination effects.

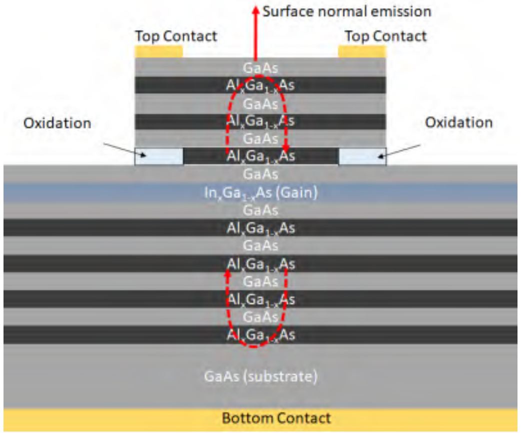

Figure 42: Etched-post VCSEL structure

VCSELs vs Edge-Emitting Lasers

There are many trade-offs between edge-emitting lasers and vertical cavity lasers. Some of these are discussed below:

- Even though it is possible to make single-mode waveguides in edge-emitters, the output beam shape will generally be elliptical. It is impossible to create a circularly symmetric waveguide mode from a stratified thin film structure. In VCSELs, it is possible to produce a circularly symmetric structure by patterning the planar layers from the top. However, achieving a single waveguide mode is more difficult in VCSELs than in edge-emitters. The etched cylinder has to be very small to become single mode. This is due to the high refractive index of the materials, and also due to the high refractive index contrast between the waveguide core (etched cylinder) and the clad (air).

- A cleaved facet is not required in VCSELs because the mirrors are entirely formed by DBR mirrors composed of thin film stacks, whereas high quality facets are required in edge-emitting Fabry-Perot lasers. Even in DFB and DBR lasers, some type of high-quality termination is required to allow the output beam to pass through. As a result, edge-emitters cannot be diced (with a saw) into chips. They have to be cleaved, or otherwise separated much more carefully. VCSELs, on the other hand, are immune from the quality of terminations along the sides. This makes VCSELs a lot more manufacturable in high volumes, significantly reducing their cost.

- Modifying the mirror reflectivities of edge-emitters requires the lasers to be separated or cleaved first followed coating their edges. Edge-coating is a lot more difficult due to the unique mounting and holding configurations than face-coating. VCSELs, on the other hand, can be coated in one step at the wafer level without separating them into individual chips.

- VCSELs can be tested for their full operation before they are diced into chips. Edge-emitters must be separated into discrete devices before they can be evaluated. Known as on-wafer testing, this allows each laser to be marked and evaluated for performance before determining which ones to dice and sort. This significantly increases the manufacturability of VCSELs.

- VCSELs can be fabricated in a two-dimensional array format quite easily, whereas edge emitters are limited to a one-dimensional array only.

It may appear, therefore, that VCSELs are far superior to edge-emitting lasers. However, that is not true. A significant weakness of VCSELs is their small cavity volume (V_p) and the gain volume (V_γ), especially if a cylinder is etched for single-mode performance. This limits the maximum power that can be obtained from a VCSEL. For the same power level, the photon density and current density in a VCSEL will be far greater than in an edge-emitter. This leads to higher internal temperatures and nonlinear optical effects. Most commercially available VCSELs are limited to a few mW of optical output, whereas edge-emitters can easily produce several hundred mW. Therefore, high power applications generally use edge-emitter, while VCSELs are ideal for low-cost low-power applications.

Applications of Diode Lasers

While there are many laser systems, ranging from gas plasma lasers (HeNe, HeCd, Argon ion, CO_2, Excimer), solid state lasers (neodymium-doped yttrium aluminum garnet Nd:YAG, ruby, titanium-doped sapphire, erbium-doped fiber lasers), diode lasers have the largest market penetration in consumer applications. This is primarily due to their low manufacturing cost, high gain values, miniature size and mechanical robustness. They can be manufactured using methods similar to electronic diodes, which is a well established process. Additionally, they are powered directly by electrical current, unlike many other lasers which have to be power by another light source, or even by another laser. Electrical powering makes integration with electronics simpler. The high gain values means the lasing threshold can be achieved at a relatively low injection current. As a result diode lasers can be run using a battery, making it possible to use in portable devices. These are some of the key factors for its wide use in consumer electronics. Although the output beam quality of diode lasers is generally inferior compared to other lasers, it is adequate for a large number of consumer applications. These range from low-end applications such as CD/DVD players, barcode readers, laser printers, laser pointers to high-end applications such as fiber optic and free-space communication systems, laser radar and ranging applications. Some larger lasers, such as Nd-YAG or fiber lasers use laser diodes to pump their amplifiers to induce population inversion. These are generally referred to as ”diode-pumped” lasers.

Coupling Laser Diodes to Optical Fibers

Earlier we examined the coupling efficiency from an LED to a multi-mode optical fiber. We can extend those principles to laser diodes as well. The principles are essentially the same, except for the asymmetry of the beam (especially with edge-emitting lasers). The different divergence angles $\theta_\perp$ and $\theta_\parallel$ will make the integral somewhat more complicated. One way of defining the beam divergence is to use the $1/e^2 = 13.5$% as the extent of the beam size. Representing the intensity function along the $\perp$ direction as $\cos^{n_\perp}\theta$, and along the $\parallel$ direction as $\cos^{n_\parallel}\theta$, we can write

$$
\begin{aligned}
0.135 &= \cos^{n_\perp}\theta_\perp &(185)\\
0.135 &= \cos^{n_\parallel}\theta_\parallel . &(186)
\end{aligned}
$$

This results in

$$
\begin{aligned}
n_\perp &= \frac{\ln(0.135)}{\ln(\cos\theta_\perp)} = \frac{-2}{\ln(\cos\theta_\perp)} &(187)\\
n_\parallel &= \frac{\ln(0.135)}{\ln(\cos\theta_\parallel)} = \frac{-2}{\ln(\cos\theta_\parallel)} . &(188)
\end{aligned}
$$

Then we can empirically represent the intensity function as

$$I(\theta,\phi) = I_o \cos^m\theta, \tag{189}$$

where

$$m = \left(n_\parallel - n_\perp\right)\sin^2\phi + n_\perp . \tag{190}$$

We can verify that $m = n_\perp$ when the azimuthal angle $\phi = 0$, and $m = n_\parallel$ when $\phi = \pi/2$. Coupling efficiency, therefore, can be calculated by performing the integration:

$$
\begin{aligned}
\eta_c &= \frac{\int_0^{\theta_a}\int_0^{\pi/2}(r\sin\theta d\phi)(rd\theta)(I_o\cos^m\theta)}{\int_0^{\pi/2}\int_0^{\pi/2}(r\sin\theta d\phi)(rd\theta)(I_o\cos^m\theta)} &(191)\\
&= \frac{\int_0^{\theta_a}\int_0^{\pi/2}\cos^m\theta\sin\theta\, d\phi\, d\theta}{\int_0^{\pi/2}\int_0^{\pi/2}\cos^m\theta\sin\theta d\phi d\theta} . &(192)
\end{aligned}
$$

Unfortunately, this cannot be evaluated analytically, but numerical integration is relatively straight forward.

Example

Given an edge emitting laser with $\theta_\perp = 28°$ and $\theta_\parallel = 8°$, we can get

$$
\begin{aligned}
n_\perp &= \frac{\ln(0.135)}{\ln(\cos 28°)} = 16 &(193)\\
n_\parallel &= \frac{\ln(0.135)}{\ln(\cos 8°)} = 204. &(194)
\end{aligned}
$$

Comparing this with LEDs, we can see that the beam is significantly narrower. The values of n for LEDs were in the range of 1-5, whereas for lasers it is in the range of 100. For the above example we can get:

$$m = 188\sin^2\phi + 16. \tag{195}$$

By numerically integrating the equation (192), we can calculate the coupling coefficient of 55%. Therefore, even without any optics, the coupling efficiency from a laser diode to a fiber can be significantly higher than from an LED. Unlike an LED, there is more latitude to use a lens to increase coupling, even when the laser diode emission diameter is larger than the fiber core diameter. Due to the very small divergence angle of the leaser beam, we can afford to de-magnify it despite the resulting enlargement in divergence angles. However, some aspects of this are complicated by the fact that the beam is asymmetric. This generally requires the use of multiple cylindrical lenses.

Back Facet Power Monitoring

So far we have implicitly assumed that the optical output emerges from mirror #1 (front mirror) and the mirror #2 (back mirror) to be highly reflective. In edge-emitting lasers, it is possible to use the power that comes out of the backside mirror to assess the power out of the front side mirror. Because the reflectivity of both mirrors are controlled and precisely known, we can derive a relationship between the two output powers. Referring back to equation (108), we can write the following:

$$P_{o1} = h\nu \frac{\phi V_p}{\tau_{pm1}} \tag{196}$$

$$P_{o2} = h\nu \frac{\phi V_p}{\tau_{pm2}}. \tag{197}$$

Therefore,

$$\frac{P_{o1}}{P_{o2}} = \frac{\tau_{pm2}}{\tau_{pm1}} = \frac{\ln R_2}{\ln R_1} \tag{198}$$

This relationship allows us to determine the output power by measuring the backside power. A photodetector can be attached to backside facet, known as the back facet monitor, to measure the power output from the laser in real time without interrupting the output beam. This is very convenient because all other methods of measurement requires splitting the output beam and redirecting one portion to a detector.

Biasing and Modulating Laser Diodes

Laser diodes can be driven just like LEDs or any other diodes, with a limiting resistor and a voltage source, or a constant-current source. However, due to the large responsivity of the laser diode above threshold compared to LEDs, laser diodes require a very stable power supply. Voltage fluctuations can lead to very large fluctuations in photon density, and if the photon density increases too much, even momentarily, it can lead to catastrophic damage to the laser facets. As a result, highly stable power supplies are necessary, especially with laser diodes that have high responsivity values. In many cases, laser diode drivers also have integrated power monitoring capability (from the backside facet) and temperature controllers.

A typical circuit for biasing a laser diode is shown in Fig 43. For example, if the laser diode has a forward voltage of $V_f = 1.6$V and the desired operating current is $I = 100$mA, and the supply voltage source is $V1 = 5$V, the required series resistance can be calculated as

$$R1 = \frac{V1 - Vf}{I} = 34\Omega. \tag{199}$$

$V2$ is the small signal modulation voltage that is superimposed on the DC bias. As with LEDs, to reduce nonlinearity, the amplitude of the modulation voltage has to be kept fairly small.

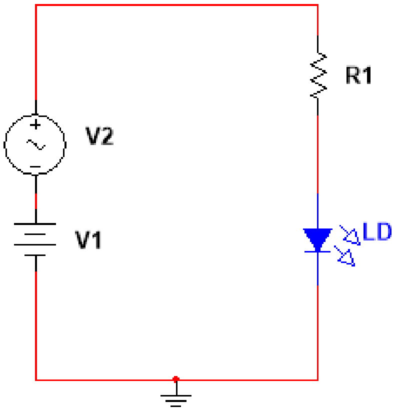

Figure 43: A laser diode biasing circuit with a current-limiting resistor

Homework 6

1. Consider an $In_{0.53}Ga_{0.47}As$/InP Fabry-Perot laser diode with a cavity length of $L = 250\mu$m. The active region ($In_{0.53}Ga_{0.47}As$) has a bandgap $E_g = 0.76$eV, an effective refractive index of 3.5 and a group refractive index of 4.0. The optical confinement factor of the waveguide is 10%.

   - The laser facets are coated such that $R_1 = 0.4$ and $R_2 = 0.95$. Assuming the internal loss, $\alpha_a + \alpha_s$, is 10/cm, calculate the threshold gain value.
     Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val L = 250.0
    val R1 = 0.4
    val R2 = 0.95
    val alpha = 10.0
    val Gamma = 0.1
    val Gth = (1.0/(2.0*L) * ln(1.0/(R1*R2))*1.0E4 + alpha)/Gamma
    println(”Gth = ${”%.2f”.format(Gth)}/cm”)
}

>>Gth = 293.52/cm
```

   - Calculate the total photon lifetime in the cavity.
     Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
  val L = 250.0
  val R1 = 0.4
  val R2 = 0.95
  val alpha = 10.0
  val neff = 3.5
  val c = 3.0e10
  val alphaC = 1.0/(2.0*L) * ln(1.0/(R1*R2))*1.0E4 + alpha
  val tauP = 1.0/alphaC*neff/c
  println(”Photon Lifetime = ${”%.2f”.format(tauP*1.0e12)} ps”)
}

>>Photon Lifetime = 3.97 ps
```

   - Calculate the extraction efficiency from this laser cavity (above threshold).
     Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val L = 250.0
    val R1 = 0.4
    val R2 = 0.95
    val alpha = 10.0
    val alphaM1 = 1.0/(2.0*L)*ln(1.0/R1)*1.0E4
    val alphaC = 1.0/(2.0*L) * ln(1.0/(R1*R2))*1.0E4 + alpha
    val etaExt = alphaM1/alphaC
    println(”Extraction Eff = ${”%.2f”.format(etaExt)}”)
}

>>Extraction Eff = 0.62
```

- This diode laser has a back facet power monitor. If the power measured on the back facet power monitor is 500μW, what would the output power through the front facet be?
  Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val R1 = 0.4
    val R2 = 0.95
    val backfacetP = 0.5
    val PowerRatio = ln(R1)/ln(R2)
    println(”Output Power = ${”%.2f”.format(backfacetP*PowerRatio)} mW”)
}

>>Output Power = 8.93 mW
```

- Current is injected into this diode such that the difference between the quasi-Fermi levels is $E_{fc} - E_{fv} = 0.80$eV. Ignoring all internal losses, determine the maximum number of longitudinal modes that will experience positive gain.
  Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val Eg = 0.76
    val Efc_Efv = 0.8
    val lambdaG = 1.24/Eg
    val ng = 4.0
    val L = 250.0
    val DeltaLambda = lambdaG.pow(2)/(2.0*ng*L)
    val modes = ((1.24/Eg-1.24/Efc_Efv)/DeltaLambda).toInt()
    println(”DeltaLambda = ${”%.2f”.format(DeltaLambda*1000.0)} nm”)
    println(”Number of modes = $modes”)
}

>>DeltaLambda = 1.33 nm
>>Number of modes = 61
```

2. The above In$_{0.53}$Ga$_{0.47}$As/InP structure is being used as semiconductor optical amplifier. The minimum desired single-pass amplification is 100. Assuming both facets have equal reflectivity, with $\alpha_a + \alpha_s = 10$/cm, calculate the maximum permissible facet reflection to prevent oscillations (assuming both facets to be identical).

   If A is the amplification,
   $A = e^{\Gamma G L} \rightarrow \Gamma G = \frac{\ln(A)}{L}$.
   To prevent oscillations, we need
   $\Gamma G < \alpha_m + \alpha_a + \alpha_s$
   $\alpha_m > \Gamma G - (\alpha_a + \alpha_s)$
   $\frac{1}{2L}\ln\left(\frac{1}{R^2}\right) > \Gamma G - (\alpha_a + \alpha_s) = \frac{\ln(A)}{L} - (\alpha_a + \alpha_s)$
   $\ln\left(\frac{1}{R}\right) > \ln(A) - (\alpha_a + \alpha_s)L$
   $R < e^{-\ln(A) + (\alpha_a + \alpha_s)L}$

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val L = 250.0
    val alpha = 10.0
    val A = 100.0
    val R = exp(-ln(A)+alpha*L*1.0e-4)
    println(”R = ${”%.2e”.format(R)}”)
}

>>R = 1.28e-02
```

3.
   - The extraction efficiency and responsivity of a laser diode is much larger than that of LEDs. Explain the main reasons for this.

     > Laser resonance takes place between the mirror facets. Hence the beam is parallel to the mirrors. As a result, more of the photons fall within the escape cone in a laser. LED has a broad angular spread (Lambertian), and a large fraction of photons fall outside the escape cone.

   - The modulation response of a laser diode is typically much greater than LEDs. Explain the main reasons for this.

     > The coupling between photons and carriers make the response time in lasers related to the photon lifetime. This value is typically in picoseconds. In LEDs (or any diodes), the response time is related to the carrier recombination life time, which is in the nanoseconds to microseconds range.

   - Explain the differences between homostructures, single heterostructures, double heterostructures, and separate confinement heterostructures in the context of semiconductor diode lasers.

4. Consider a GaAs/Al$_x$Ga$_{1-x}$As VCSEL with a In$_{0.1}$Ga$_{0.9}$As active layer. The top and bottom reflectors have a reflection coefficient of 0.990 and 0.999, respectively. The thickness of the gain region is 25nm, and the confinement factor can be assumed to be 1.0 within the active layer.

   - Ignoring other material losses, calculate the threshold gain.
     Run this code

     ```
     import kotlin.math.*
     //Andrew Sarangan

     fun main() {
         val R1 = 0.99
         val R2 = 0.999
         val L = 25.0 //nm
         val alphaC = 1.0/(2.0*L) * ln(1.0/(R1*R2))*1.0E7
         val Gth = alphaC //since confinement = 1
         println(”Gth = ${”%.2f”.format(Gth)} /cm”)
     }

     >>Gth = 2210.17 /cm
     ```

   - Qualitatively explain the reason why VCSELs have a much narrower emission spectrum than Fabry-Perot lasers.

The emission wavelength in VCSELs is controlled by the grating mirror consisting of multiple layer of thin films. With a quarter-wave shifted grating, it is possible to create a single resonance wavelength. In a Fabry-Perot laser, the mirrors reflect a very broad range of wavelengths. Hence the emission spectrum is controlled primarily by the gain spectrum (which is relatively broad).

5. A diode laser has a beam divergence of $\theta_{\|} = 10^{\circ}$ and $\theta_{\perp} = 20^{\circ}$. Assuming a fiber with a numerical aperture of 0.15, and the near-field beam of the laser is an ellipse 10μm x 2μm, and that the fiber core has a diameter of 25μm, estimate the coupling efficiency (without a lens). Discuss how this coupling can be improved by using lenses.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun Double.toRad() = this*PI/180.0

fun main() {
    fun fn(theta:Double, phi:Double, m:Double) = cos(theta).pow(m) * sin(theta)
    val nperp = ln(1.0/exp(2.0))/ln(cos(20.0.toRad()))
    val nparl = ln(1.0/exp(2.0))/ln(cos(10.0.toRad()))
    val dTheta = 1.0e-3
    val dPhi = 1.0e-3
    val theta = DoubleArray((PI/2.0/dTheta).toInt()){it*dTheta}
    val phi = DoubleArray((PI/2.0/dPhi).toInt()){it*dPhi}
    val NA = 0.15
    val thetaA = asin(NA)

    var total = 0.0
    var coupled = 0.0
    phi.forEach{ phi ->
        val m = (nparl-nperp)*sin(phi).pow(2) + nperp
        theta.forEach{ theta ->
            total += fn(theta,phi,m)*dTheta*dPhi
            if (theta < thetaA){
                coupled += fn(theta,phi,m)*dTheta*dPhi
            }
        }
    }
    println(”${”%.2f”.format(coupled/total)}”)
}

>>0.50
```

Photodetectors

Photodetectors can be categorized into photon detectors and thermal detectors. Photon detectors rely on the creation of an electron-hole pair due to an absorbed photon. Hence, these devices require a semiconductor with an appropriate bandgap. However, unlike emitters like LEDs and laser diodes, the bandgap does not have to be direct. Photon absorption can take place even in indirect bandgap semiconductors, such as silicon and germanium.

There are two broad categories within photon detectors - photoconductors and photodiodes. The ones that produce a change in conductance (or resistance) due to an incident light are known as photoconductors (or as photocells, or photoresistors). Strictly speaking, these are variable resistors whose resistance is a function of incident radiation. As resistors, they convert electrical energy into heat. They do not produce electrical energy from electromagnetic energy. Photodiodes, on the other hand, can convert electromagnetic energy into electrical energy. But that is a feature, not a requirement. The majority of photodiodes are used simply as a sensor, not as an energy harvester. Photovoltaic devices (solar cells) are another class of photodiodes designed for the explicit purpose of converting light energy into electrical energy.

Thermal detectors are based on the heating effect due to light absorption. Bolometers operate by absorbing an incident radiation and converting it to heat, and then measuring the rise in temperature by using a temperature-sensitive resistor. A large temperature coefficient of resistance (TCR) is desired for this application. The rise in temperature can also be measured using crystals that have a pyroelectric effect, which produce a voltage difference due to a temperature difference.

The table below lists the common semiconductors used in photodetectors, their bandgaps and the corresponding cutoff wavelengths at room temperature (300K). Bandgaps generally decrease with increasing temperature, so those materials that are used at cryogenic temperatures (such as InSb and InAs) will exhibit larger bandgaps than listed here.

Material	Bandgap (eV)	Wavelength (um)	Type	Application
CdS	2.42	0.51	Direct	Photocells
CdSe	1.74	0.71	Direct	Photocells
CdTe	1.5	0.83	Direct	Solar cells
GaAs	1.42	0.87	Direct	
InP	1.27	0.97	Direct	
Si	1.1	1.12	Indirect	Visible and NIR detectors and cameras
$In_{0.53}Ga_{0.47}As$	0.74	1.67	Direct	SWIR detectors and cameras
Ge	0.66	1.87	Indirect	
PbS	0.37	3.35	Direct	MWIR detectors
InAs	0.36	3.44	Direct	MWIR detectors
PbSe	0.27	4.6	Direct	MWIR detectors
InSb	0.23 (at 77K)	5.4	Direct	MWIR detectors and cameras
$Hg_{1-x}Cd_xTe$	-0.3 to 1.6 (at 77K)		Direct	MWIR & LWIR detectors and cameras

Photodiodes

PN junction diodes can operate as photon detectors. In fact this is the most widely used type of photodetector. When illuminated to light, the electron-hole pairs generated by the photons can significantly modify the I-V characteristics of the diode. The number of incident photons can then be determined by measuring the change in voltage (at a fixed current), or a change in current (at a fixed voltage).

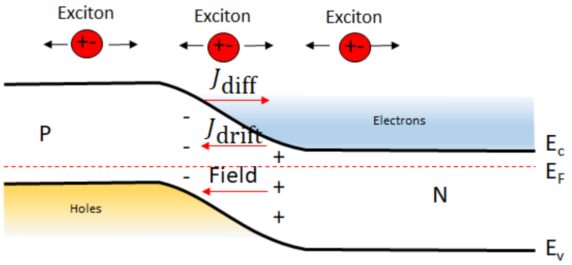

Figure 1: Potential band diagram of a PN junction under zero bias.

Consider, for example, the potential band diagram of a PN junction under zero bias as shown in Fig 1. An incident photon will create an electron-hole pair by elevating an electron from the valence band to the conduction band. The electron and hole will exist in the same space as a single neutral particle known as an exciton. Excitons are loosely bound electron-hole pairs, and are easily ionized by even a small electric field. When an exciton is exposed to an electric field, it will get ionized and become separated into a free electron and a free hole. The electron and hole will move in opposite directions, but it will constitute a single current flowing from the N-side to the P-side. This is in the same direction as the reverse bias current, and will present itself as an offset current to the normal bias current. Therefore, the I-V relationship becomes:

$$I = I_s\left(e^{V_a/V_t} - 1\right) - I_{ph}, \tag{1}$$

where I_{ph} is the photocurrent. I_{ph} constitutes only those excitons that are intercepted by the space charge region. All the other excitons that never intercept the space charge region will eventually recombine and disappear.

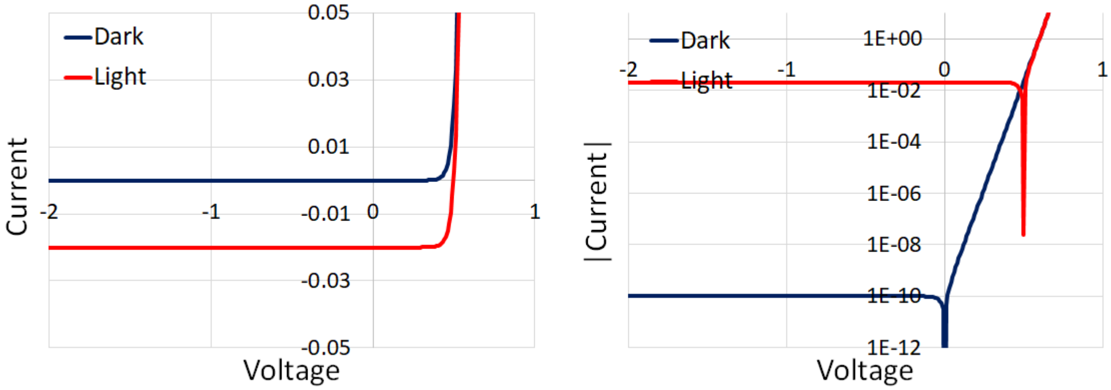

Figure 2: Terminal characteristics of a photodiode plotted as I-V (left) and as log(|I|)-V (right).

The I-V curves of a photodiode with and without illumination are shown in Fig 2. With illumination, the current will experience an offset that is equal to the photocurrent. Due to the large difference in magnitudes, especially in the reverse direction, this is best plotted on a log scale after taking the magnitude of the current. Given the small values of the reverse saturation current, the photocurrent offset produces the largest change in the reverse bias region. Therefore, photodetectors are typically operated in the reverse base region.

As stated before, excitons have to be exposed to an electric field in order to be separated into free carriers. We also know that nearly all of the electric field is contained in the space charge region. Therefore it may appear that only a small fraction of excitons that are generated inside the space charge region will become electrons and hole. But this is not actually the case. In practice, nearly all excitons will be ionized. This is due to the non-zero electric field that exists outside the space charge region. But only some of them will survive long enough to cross the space charge region and become majority carriers. To understand this, we have to revisit the description of current flow in diodes. We saw that minority carriers carry the current only to within a few diffusion lengths outside the space charge region. The current through the remainder of the structure is carried by drift current of the majority carrier. This requires a small electric field, and a small potential drop. This field is sufficient to ionize the excitons and create free electrons and holes. However, whether or not those minority carriers will reach the space charge layer (from which they will emerge as majority carriers) will depend on their carrier lifetime and the transit time.

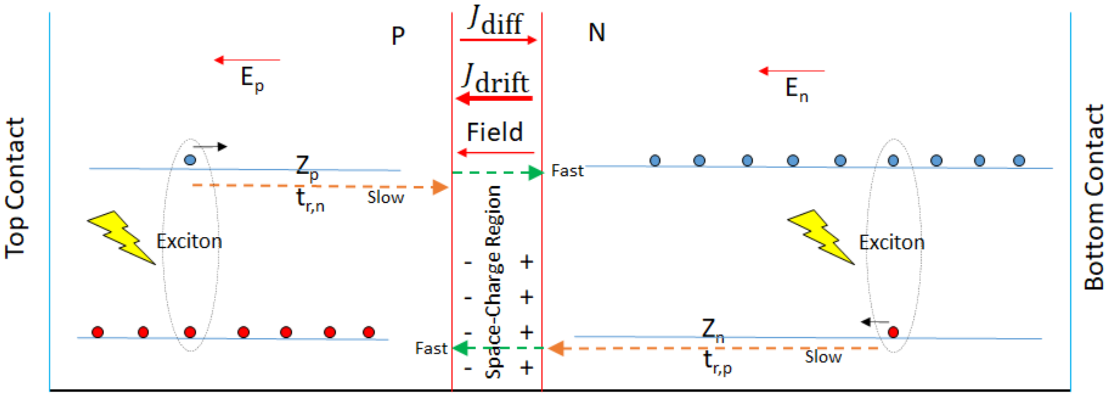

Figure 3: Transit time model of a photodiode

Referring to the illustration in Fig 3, an exciton generated in the p-side will produce a minority carrier electron and a majority carrier hole. The minority electron needs to survive the trip to the space charge layer. The transit time will be

$$t_{r,n} = \frac{Z_p}{v_n} \tag{2}$$

where Z_p is the distance (on the p-side) from that exciton to the space charge layer, and v_n is the drift velocity of the electron on the p-side. Since the field is very small, the drift velocity will be in the linear regime, allowing it to be written as

$$v_n = \mu_n E_p \tag{3}$$

where E_p is the field on the p-side. This electric field can be deduced from the current flowing through the diode as

$$I = qA\mu_p pE_p = qA\mu_p N_A E_p \tag{4}$$

where A is the cross-sectional area of the diode, and N_A is the acceptor doping. Therefore,

$$E_p = \frac{I}{qA\mu_p N_A} \tag{5}$$

from which we can get

$$
\begin{aligned}
t_{r,n} &= \frac{Z_p}{v_n} && (6)\\
&= \frac{Z_p}{\mu_n E_p} && (7)\\
&= \frac{Z_p}{\mu_n}\frac{qA\mu_p N_A}{I}. && (8)
\end{aligned}
$$

In these expressions, we are making the assumption that the drift current is uniform across the entire diode. Clearly, this is not exactly the case. Referring to the chapter on diodes, we can surmise that the drift current will grow as one moves away from the space charge region. This means that the field will be higher near the space charge layer. However, as a first order approximation, this is a reasonable assumption to make.

Similarly, for the excitons created on the n-side, we can get

$$
t_{r,p} = \frac{Z_n}{\mu_p}\frac{qA\mu_n N_D}{I}. \tag{9}
$$

In this model we have ignored the transit time through the space charge layer. This is a reasonable assumption because its value typically very small. Because of the high field in this region, the carriers will travel at high velocities. At even higher fields, they can reach velocity saturation. In silicon, this saturation velocity is on order of 10^7cm/s. Since space charge layer widths are usually on the order of $1-5\mu$m, the transit time through the space charge layer works out to be on the order of 10ps.

If the total transit time is longer than τ, then those carriers will recombine and disappear before reaching the space charge layer. Therefore, the maximum distance an exciton can travel on the p- and n-side can be written as

$$
\begin{aligned}
Z_{p,\text{max}} &= \frac{I\tau_n}{qAN_A}\frac{\mu_n}{\mu_p}, && (10)\\
Z_{n,\text{max}} &= \frac{I\tau_p}{qAN_D}\frac{\mu_p}{\mu_n}. && (11)
\end{aligned}
$$

Any excitons created outside of these boundaries will not contribute to any photocurrent. Exctions that are too close to the incident surface, or too far into the substrate will be lost before they can reach the space charge layer. $Z_{n,\text{max}}+x_n+x_p+Z_{p,\text{max}}$ is known as the collection width of the photodiode, where x_n and x_p are the space charge widths on the n and p-sides, respectively.

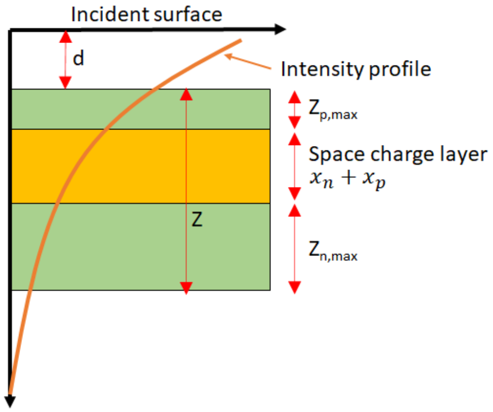

Figure 4: Attenuation of optical intensity and the collection volume.

It is worth pointing out that the drift fields E_p and E_n, depend on the photocurrent I. Since that is the quantity we are trying to calculate, this requires an iterative solution. We have to make an estimate of the photocurrent to get E_p and E_n, from which can calculate the responsivity and the photocurrent.

The incident light will be absorbed as it enters the semiconductor, and the intensity will decay as a function depth

$$I = I_i e^{-\alpha z} \tag{12}$$

where α is the attenuation coefficient. The attenuation can also be expressed in terms of the imaginary part of the refractive index

$$\alpha = 2k_0\kappa \tag{13}$$

where the refractive index is $n - j\kappa$. The intensity that is lost due to absorption can be written as a differential quantity dI. The collection efficiency is, therefore,

$$\eta_c = \frac{\int_d^{d+Z} \frac{dI}{dz} dz}{\int_0^{\infty} \frac{dI}{dz} dz} \tag{14}$$

where d is the distance below the incident surface where the collection region begins, and

$$Z = Z_{n,\text{max}} + Z_{p,\text{max}} + x_n + x_p \tag{15}$$

where $x_n + x_p$ is the space charge width. From this, we can get

$$\eta_c = e^{-\alpha d}\left(1 - e^{-\alpha Z}\right). \tag{16}$$

Quantum Efficiency of Photodiodes

The internal quantum efficiency (IQE), η_i is the number of electron-hole pairs produced per incident photon. With proper anti-reflection coating, it is possible to achieve a very high internal quantum efficiency in most semiconductor materials.

The fraction of excitons that fall between the bounds of Z_n and Z_p is the collection efficiency η_c. This was derived in equation (16). The external quantum efficiency (EQE) is, therefore

$$\eta_e = \eta_i \eta_c. \tag{17}$$

While the value for IQE can be very high, the collection efficiency varies greatly with geometry and applied bias. We can see from equations (10) and (11) that bias current (and hence bias voltage) has a linear effect on the collection width.

The responsivity can be expressed in terms of the external quantum efficiency as

$$\begin{aligned} q\frac{P_i}{h\nu}\eta_e &= I_{ph} && (18) \\ \mathcal{R} &= \eta_e \frac{q}{h\nu}. && (19) \end{aligned}$$

Dark Current

The dark current of a photodiode is the current that would flow with no illumination. Earlier, we had the photodiode I-V characteristic as:

$$I = I_s\left(e^{V_a/V_t} - 1\right) - I_{ph}. \tag{20}$$

Under reverse bias with a sufficiently large and negative V_a, and under dark conditions (zero I_{ph}), the diode current would become

$$I \approx -I_s. \tag{21}$$

The value of I_s (reverse saturation current) depends on semiconductor parameters such as intrinsic concentration, doping concentrations, diffusion coefficients and diffusion lengths, which we examined in an earlier chapter on basic diode properties. Using this model, typical values for a 1 cm x 1 cm diode will be smaller than a pico amp. But in practice, the measured saturation currents are significant higher, on the order of nano amps. This arises due to additional current components not accounted for in the expression for I_s, some of which include the following:

- In deriving the diode I-V characteristics, we had assumed that the recombination lifetime was a constant value τ. In reality, this lifetime will be a function of carrier concentration, leading to a nonlinear recombination rate with carrier density. Different mechanisms contribute to these recombination events. The three dominant mechanisms are (1) radiative (band-to-band) recombinations, (2) trap-assisted recombinations (also known as Shockly-Read-Hall recombinations) and (3) Auger recombinations (which dominates at high carrier densities). Under steady-state, the generation rate will be equal to the recombination rate. Under forward bias, the recombination rate will exceed the generation rate. Under reverse bias, generation rate exceeds recombination rate. In the I-V diode equation (20), the lifetime is contained in the saturation current I_s. The nonlinear nature of the lifetime will become most apparent in the reverse bias region, resulting in an increase in current with bias voltage.
- Defects in the semiconductor crystal will lead to an additional current component that *leaks* through the PN junction. Some of this could also arise due to non-uniform doping density, especially in thermally diffused doping profiles. In such structures, doping levels will be high near the surface, resulting in very narrow space charge widths. This can lead to a tunneling current through the PN junction.
- Even though the surface area A of the photodiode is typically taken as the top surface area, in practice, the junction is a three-dimensional structure. It will contain sidewall areas in addition the horizontal area. This will lead a larger effective surface area, which will produce a higher I_s than calculated.

Due to these effects, the total dark current can be significantly larger than what we calculated assuming a constant value for τ. A simple model to account for this is by modifying the diode equation (1) with a parallel current path for I_s, using an equivalent shunt resistor:

$$I = \left(I_s + \frac{|V_a|}{R_{\text{shunt}}}\right)\left(e^{V_a/V_t} - 1\right) - I_{ph}. \tag{22}$$

This is a highly simplified model, but it does provide a means to account for the excess currents without complicated equations. Additionally, R_{shunt} should not be taken as a constant, because it will be bias-dependent. Typical values for R_{shunt} will be on the order of 100 MΩ. At 5V reverse bias, this will lead to a current of about 50 nano amps instead of pico amps.

Example

Consider a silicon PN junction photodiode with an internal quantum efficiency of 0.9, and with a surface area of 100μm x 100μm. It is irradiated with an incident intensity of 10mW/cm^2 at a wavelength of 600nm. We will assume that the surface is AR coated such that all of the light enters the semiconductor. Furthermore, assume that the junction is at a depth of 2μm below the surface. The top side is doped p-type with $N_A = 10^{17}$/cm^3. The substrate is doped n-type with $N_D = 10^{15}$/cm^3. The diode is reverse biased, with a dark current of 10nA at a bias voltage of -5V.

The lifetimes of the minority carriers are a strong function of doping density. These values can be looked up from repositories. We can get the following lifetime values for $N_A = 10^{17}$/cm^3 and $N_D = 10^{15}$/cm^3:

$$\tau_n = 20\ \mu\text{s} \tag{23}$$

$$\tau_p = 200\ \mu\text{s}. \tag{24}$$

Additional values can be looked up as well: $\mu_n = 1450$cm^2/V.s and $\mu_p = 200$cm^2/V.s, $D_n = 37$ cm^2/s and $D_p = 5$ cm^2/s. Assuming a long diode approximation for the n-side and a short-diode approximation for the p-side, we can calculate the value for I_s,

$$I_s = \underbrace{qA\left(\frac{D_n}{W_n}\frac{n_i^2}{N_A} + \frac{D_p}{L_p}\frac{n_i^2}{N_D}\right)}_{6.78\ fA} + \frac{|V_a|}{R_{\text{shunt}}} = 10\ \text{nA}, \tag{25}$$

where we have used W_n as the junction depth on the p-side. Obviously the linear saturation current is much smaller than the actual dark current, so we can ignore it and can get the shunt resistance of the diode as

$$R_{\text{shunt}} = \frac{5}{10\times 10^{-9}} = 500\ \text{M}\Omega. \tag{26}$$

The built-in voltage for this diode can be calculated as:

$$V_{bi} = V_t \ln\left(\frac{N_D N_A}{n_i^2}\right) = 0.026 \ln\left(\frac{10^{15}\times 10^{17}}{(1.5\times 10^{10})^2}\right) = 0.697\ \text{V}. \tag{27}$$

Next, we can calculate the total space charge width at the given bias voltage of -5V. Silicon has a dielectric constant value of $11.68\epsilon_o$. Therefore,

$$x_p = \sqrt{\frac{2\epsilon_s}{q}\frac{N_D}{N_A}\frac{1}{N_A+N_D}(V_{bi}-V_a)} = 27\ \text{nm} \tag{28}$$

$$x_n = \sqrt{\frac{2\epsilon_s}{q}\frac{N_A}{N_D}\frac{1}{N_A+N_D}(V_{bi}-V_a)} = 2.7\ \mu\text{m}. \tag{29}$$

As a result, the total space charge width is 2.73μm, extending primarily into the substrate.

In order to proceed further, we need to assume a diode current (sum of dark current and photocurrent). Since that is the end-result we are trying to calculate, this has to be done recursively. We will start by assuming $I = 200$nA. Using this, we can get the collection widths on either side of the junction:

$$Z_{p,\text{max}} = \frac{I\tau_n}{qAN_A}\frac{\mu_n}{\mu_p} = 90\ \text{nm} \tag{30}$$

$$Z_{n,\text{max}} = \frac{I\tau_p}{qAN_D}\frac{\mu_p}{\mu_n} = 6.9\ \mu\text{m}. \tag{31}$$

The total collection width becomes:

$$Z = Z_{n,\text{max}} + Z_{p,\text{max}} + x_n + x_p = 9.7\ \mu\text{m}. \tag{32}$$

We can note that the collection width is significantly larger than the space charge width. This is a typical with most photodiodes. One way to increase the collection width in photodiodes is by using a substrate with a longer recombination lifetime τ_n or τ_p (which is typically done by selecting a high-quality low-defect substrate with low doping values, or ideally undoped).

We can also calculate the electric fields in the collection regions:

$$E_p = \frac{I}{qA\mu_p N_A} = 3.1 \times 10^{-4} \text{ V/cm} \tag{33}$$

$$E_n = \frac{I}{qA\mu_n N_D} = 8.6 \times 10^{-3} \text{ V/cm}. \tag{34}$$

The value of the electric field inside the space charge region is much greater. Although this field is not constant, the average field can be estimated to be

$$\bar{E}_{\text{sc}} = \frac{V_{bi} - V_a}{x_n + x_p} = 2.1 \times 10^4 \text{ V/cm}, \tag{35}$$

which validates are earlier assumption that the field outside the space charge region is much smaller.

Next, we have to look up the refractive index of silicon at $\lambda = 600$nm. This value is $n - j\kappa = 3.93 - j0.018521$. Therefore, the attenuation coefficient works out to $\alpha = 3879$/cm.

Finally, the collection efficiency can be calculated:

$$\eta_c = e^{-\alpha d}\left(1 - e^{-\alpha Z}\right) = 0.45, \tag{36}$$

from which we can get the external quantum efficiency (EQE) of

$$\eta_e = \eta_i \eta_c = 0.40. \tag{37}$$

Finally, the responsivity becomes

$$\mathcal{R} = \eta_e \frac{q}{h\nu} = 0.40 \frac{0.6}{1.24} = 0.195 \text{ A/W}. \tag{38}$$

Since the incident intensity is 10mW/cm^2, the incident power becomes

$$P_i = I_i A = 10 \times \left(100 \times 10^{-4}\right)^2 = 1\mu\text{W}. \tag{39}$$

Therefore, the detected photocurrent will be

$$I = \mathcal{R} P_i = 195 \text{ nA}, \tag{40}$$

which is very close to the 200nA we had assumed at the start of this calculation. Therefore, it is not necessary to iterate the calculation.

Spectral Response of Photodiodes

Using the model developed in the previous section, we can predict the spectral behavior of photodiodes. This is going to be primarily dictated by the dispersion characteristics of the semiconductor material. In most semiconductors, α becomes smaller as the wavelength approaches the band gap wavelength, and falls to zero for wavelengths longer than the bandgap wavelength. Therefore, the collection efficiency (equation (16)) will decline as one approaches the band gap wavelength from shorter wavelengths. At the other end of the spectrum, at very short wavelengths, α will be extremely large. This also results in a reduced collection efficiency. This arises because most of the absorption will now take place in the region above the collection region.

The shape of the spectral response will be largely dictated by two factors. Most of the short-wavelength photons are absorbed close to the incident surface, and most of the longer-wavelength photons are absorbed deeper within the semiconductors. Therefore, the short-wavelength response will be a strong function of the depth of the junction below the surface (d in Fig 4). Long-wavelength response will be a strong function of the depth of the collection region. Silicon photodiodes optimized for blue wavelengths are fabricated with extremely shallow junctions (on the order of a few hundred nanometers), and those optimized for infrared are designed with a very thick collection region (such as PIN diodes, discussed a littler later in this section).

Example (Cont'd)

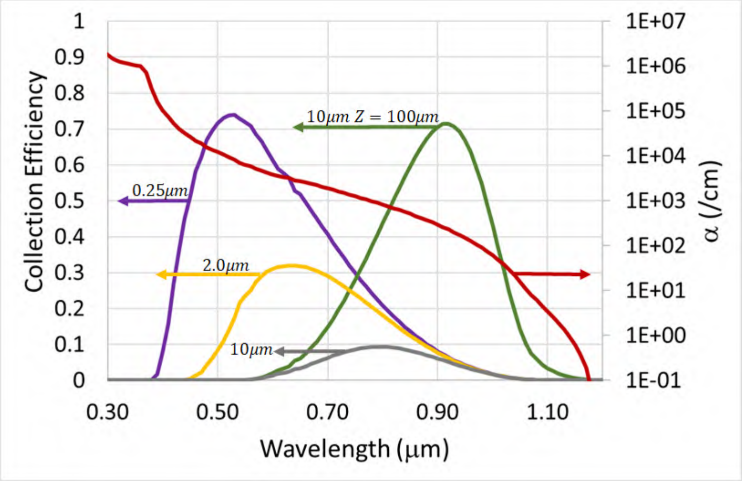

Figure 5: Plot of attenuation coefficient α vs wavelength and the corresponding value of collection efficiencies for junction depths of 0.25μm, 2μm and 10μm.

Continuing the example of the silicon photodiode from before, we can look up the dispersion characteristics of silicon, from which we can plot the behavior of α vs wavelength. This is shown on the right hand axis of Fig 5. Then, we can calculate the collection efficiency using equation (16), which is shown on the left hand axis Fig 5. Three different junction depths are used: 0.25μm, 2μm and 10μm. We can see that α is very large at wavelengths below 400nm, which leads to a nearly zero collection efficiency in all cases. As α declines with increasing wavelength, collection efficiency goes up, reaching a peak value, and then declines. This peak occurs at different wavelengths depending on the junction depth. For the shallow 0.25μm junction, the peak is 0.8 at a wavelength of 550nm. For the junction at 2μm below the surface, the peak value is 0.38 at a wavelength of 680nm, and for 10μm deep junction, the peak value is 0.12 at a wavelength of 800nm. Beyond 1μm, the value of α is very small, and the collection efficiency declines to extremely small values in all three cases. We can also notice that except for the shift towards longer wavelengths, the peak values of the deeper junctions are no higher than that of the shallow junction. This is because the top layer is simply acts like a long-pass cut-off filter. A thinner top layer allows more of the shorter wavelengths to pass through, but it does not necessarily improve the collection of long-wavelength photons. To improve the collection of long wavelength photons, the maximum depth of the collection area has to be increased. To illustrate this effect, Fig 5 also shows the case for a device with the junction 10μm below the surface, but extending deeper, with $Z = 100\mu$m. We can see the peak value for this case is significantly higher in the infrared, reaching a maximum of 0.71

at a wavelength of 930nm.

Fig 6 shows the responsivity plots for the same scenarios shown in Fig 5. All of the curves exhibit essentially the same spectral shape as the collection efficiency, except for a small shift due to the photon energy $h\nu$ in equation (19). The responsivity is nearly zero below 400nm for all cases. The peak responsivity values are progressively at longer wavelengths as the junction is placed deeper below the substrate surface. For the case with $Z = 100\mu$m, the peak responsivity occurs at 930nm.

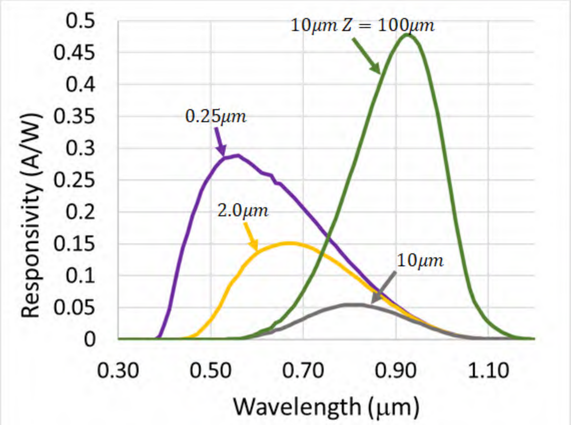

Figure 6: Plot of responsivity vs wavelength.

Silicon PN Photodiode Structures

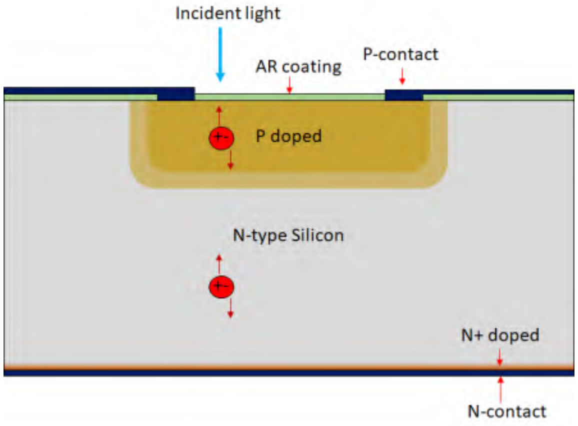

Figure 7: Generic silicon photodiode structure

The generic structure of a silicon photodiode is shown in Fig 7. In this configuration, the device is constructed on an n-type substrate by selectively doping an area p-type. This is known as compensation doping, where the opposite dopant type is used to compensate and reverse the polarity of the substrate. Obviously, this will require the p-type concentration to be larger than the n-type. Then electrical contacts are placed on the p-side. Since the metal contacts should allow the light to enter the semiconductor, the contact it has to be designed in a ring configuration. Anti-reflection coatings are also applied to the top surface of the device. The n-type contacts are made on the back side of the substrate. Since light is not entering from this side, it can be a uniform metal film. However, a metal on a lightly doped semiconductor often creates what is known as a Schottky barrier. This will distort the I-V characteristics of the PN junction diode. Therefore, a thin highly-doped n-type layer is created first, followed by the metal film. This allows the semiconductor/metal interface to behave as an ohmic contact. This type of geometry, where the light is incident from the junction-side of the device is known as front-side-illuminated photodiodes. In general, a lightly-doped substrate is will improve the collection efficiency because the minority carrier lifetime will then be longer on the substrate side. A shorter carrier lifetime on the front side is inevitable because it will usually be doped higher than the substrate.

PIN Photodiodes

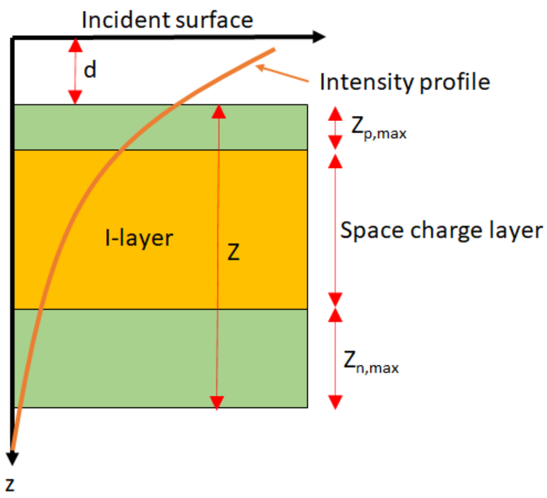

Figure 8: Optical absorption and collection in a P-I-N photodiode.

P-I-N photodiodes consist of an extra intrinsic layer (hence the ”I”) sandwiched between the p-side and the n-side of the diode. This effectively creates a double junction: one at the P-I interface and another at the I-N interface. In practice, the I-layer will never be intrinsic, but it will be lightly-doped either p-type or n-type.

In the previous example we saw that the space charge layer primarily exists in the low-doped side of the junction. The lower the doping, the larger the space charge width. In the case of P-I interface, the space charge layer will be large and all of it will exist in the I-layer. Similarly, all of the space charge of the I-N junction will also exist in the I-layer. In fact, the extent of the space charge layer in the I-layer will be several hundred microns. When the calculated space charge width exceeds the I-layer thickness, the entire I-layer will be depleted (it will not spill into the adjacent doped layers), merging both the P-I and I-N junctions together into a single junction.

There are several advantages to the PIN photodiode compared to a PN photodiode. The space charge width will be much larger than in a regular PN photodiode (tens of microns instead of a few microns). This will increase the collection efficiency of the photodiode. Additionally, the large space charge width also reduces the breakdown voltage of the diode under reverse bias. This enables us to bias the photodiode at a high reverse bias, which can produce greater linearity.

Example (Cont’d)

Consider a similar example as before, but with a silicon PIN photodiode configuration with a I-region thickness of 300μm. The start of the I-region is at 0.25μm distance from the surface of the semiconductor.

$Z_{n,\text{max}}$ and $Z_{p,\text{max}}$ will remain the same because the p and n-type doping concentrations are the same as before, at 90nm, and 6.9μm, respectively. The built-in voltage V_{bi} will also remain unchanged at 0.697V. At a reverse bias voltage of 5V, we can get the space charge widths for the P-I junction as:

$$x_p = \sqrt{\frac{2\epsilon_s}{q}\frac{n_i}{N_A}\frac{1}{N_A+n_i}(V_{bi}-V_a)} = 0.1 \text{ nm} \tag{41}$$

$$x_i = \sqrt{\frac{2\epsilon_s}{q}\frac{N_A}{n_i}\frac{1}{N_A+n_i}(V_{bi}-V_a)} = 700\ \mu\text{m}. \tag{42}$$

where we have used an intrinsic carrier density of $n_i = 1.5\times10^{10}\text{cm}^{-3}$. At the I-N junction, we can get

$$x_i = \sqrt{\frac{2\epsilon_s}{q}\frac{N_D}{n_i}\frac{1}{n_i + N_D}(V_{bi} - V_a)} = 700\ \mu\text{m} \tag{43}$$

$$x_n = \sqrt{\frac{2\epsilon_s}{q}\frac{n_i}{N_D}\frac{1}{n_i + N_D}(V_{bi} - V_a)} = 10\ \text{nm}. \tag{44}$$

From this, we can see that the sum of the calculated space charge widths in the I-layer is significantly larger than the I-layer thickness. As a result, the entire I-layer will be depleted. The space charge width on the p-side and n-sides are negligibly small.

Next, we can calculate the total collection width:

$$Z = Z_{n,\text{max}} + Z_{p,\text{max}} + x_n + x_p = 307\ \mu\text{m}. \tag{45}$$

Using these values, the collection efficiency can be calculated at a wavelength of 600nm :

$$\eta_c = e^{-\alpha d\left(1-e^{-\alpha Z}\right)} = 0.91, \tag{46}$$

from which we can get the external quantum efficiency of

$$\eta_e = \eta_i \eta_c = 0.82. \tag{47}$$

Finally, the responsivity becomes

$$\mathcal{R} = \eta_e \frac{q}{h\nu} = 0.40\ \text{A/W}, \tag{48}$$

which is larger than what was achieved with the PN junction photodiode.

Fig 9 shows the calculated responsivity of the PIN photodiode. For this example, we have assumed a shallow junction depth of 0.25μm and a I-layer thickness of 300μm. This will give us a very good short-wavelength performance as well as a good long-wavelength performance.

One of the tradeoffs with PIN photodiodes is the response speed. Even though the entire I-layer is depleted, the transit time through this layer will still not be insignificant. This can reduce the speed of the device.

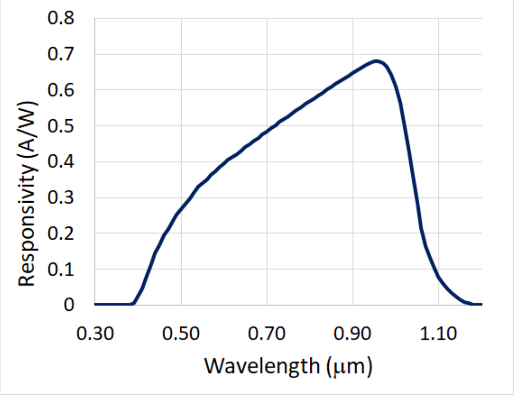

Figure 9: Calculated responsivity curve of a silicon P-I-N photodiode with a 0.25μm junction depth and a 300μm I-layer.

Silicon PIN Photodiode Structures

In silicon, PIN structures are fabricated by starting from an intrinsic (or low-doped) substrate. These are referred to as float-zone (FZ) silicon in reference to the manufacturing technique used to grow these ultra-low-doped silicon crystals. The p-type and n-type dopings are applied from opposite sides of the substrate to create the PIN geometry. Although many different configurations are used, one common configuration is shown in Fig 10. Since the doping depths are usually limited to a few microns, the thickness of the substrate essentially defines the thickness of the I-layer. Substrate thicknesses of $200 - 500\mu$m are fairly common in these devices.

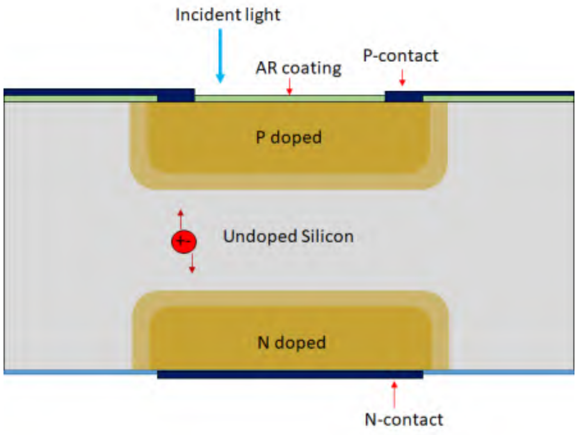

Figure 10: PIN photodiode structure

Fig 11 shows the published responsivity curve for a commercially available silicon photodiode. Comparing it with Fig 9, we can see that their are very similar.

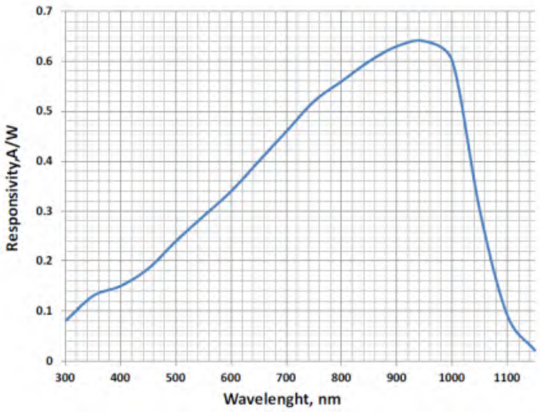

Figure 11: Spectral responsivity curve of a commercial photodiode. Source: Luna Optoelectronics.

Backside Illuminated Photodiodes

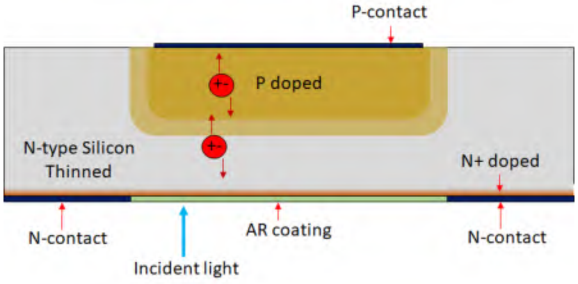

Figure 12: Backside illuminated silicon PN photodiode

An alternate configuration known as backside-illuminated photodiode is shown in Fig 12. This is almost identical to Fig 7, except the incident light enters the device from the substrate-side. Despite the similarity to front-side illuminated photodiodes, the construction of these devices is fundamentally different. Starting from the n-type substrate, the p-type regions are doped first. However, since substrates are generally several hundred microns thick, illuminating this photodiode through the substrate will result in an extremely poor collection efficiency. Most of the photons will be absorbed in the substrate too far from the space charge layer. This problem is avoided by thinning down the substrate to a sufficiently small value to allow the photogenerated carriers to reach the space charge region. Then the anti-reflection coating and contacts are applied to the backside. This is a complex task because the thinning process has to be accurately controlled to reach within a few microns of the space charge layer. Although this may seem like an unnecessarily complicated process, backside illumination is highly favored in imaging applications. When electronic components have to be integrated with photodetectors (which is always the case with cameras) all of the electronics including the doping areas required for photodetectors can be made on one side, allowing the backside free for optical coatings and optical interfacing. Despite the need to thin down the substrate, backside illuminated detectors actually have a larger external quantum efficiency. Many of the research-grade silicon image sensors in the market today use backside illuminated geometries.

Infrared Photodiodes

As evident from Fig 6, silicon responds wavelength between 400nm and 1μm. For detecting longer (or shorter) wavelengths, other semiconductor materials are required. A number of common materials are used for other wavelengths, which are listed below:

- In$_{0.53}$Ga$_{0.47}$As grown on a InP substrate gas a bandgap of 0.74eV. The corresponding wavelength is 1.67μm. Therefore, it is commonly used for detecting wavelengths up to 1.6μm., especially for fiber telecommunication applications. Because the In$_{0.53}$Ga$_{0.47}$As layer is grown epitaxially on the InP substrate, and has a smaller bandgap than InP (1.34eV), it will exhibit absorption only in the In$_{0.53}$Ga$_{0.47}$As layer and in the substrate. If the In$_{0.53}$Ga$_{0.47}$As layer is designed to be fully depleted, this can result in high speed response from the photodiode. Nearly all of the SWIR cameras are built using an array of In$_{0.53}$Ga$_{0.47}$As detectors.
- InSb is a semiconductor very similar to GaAs. Unlike In$_{0.53}$Ga$_{0.47}$As, it is manufactured as substrates. As a result, the device geometries are very similar to that of silicon. The bandgap of InSb is 0.17eV, which corresponds to a wavelength of 7.3μm. However, the intrinsic carrier concentration of InSb at room temperature is 2×10^{16}cm-3, which is 6 orders of magnitude larger than silicon. Therefore, it behaves more like a metal than a semiconductor at room temperature, resulting in poor diode performance, large dark

currents and noise. This is overcome by operating the device at lower temperatures, for example at liquid nitrogen temperature (77K). The intrinsic carrier concentration drops to about 5×10^9cm^{-3}, which is much closer to silicon. The bandgap also increases to 0.23eV, which corresponds to a wavelength of 5.4μm. As a result, cooled InSb photodiodes can respond to wavelength between 3μm and 5μm, which ideally matches one of the atmospheric transmission windows for infrared wavelengths, as well as most thermal imaging spectra.

- Hg$_{1-x}$Cd$_x$Te, or simply MCT, is a material that is synthesized by combining HgTe and CdTe. Similar to Al$_x$Ga$_{1-x}$As, the entire range of composition has nearly the same lattice constant. HgTe is actually a metal with a negative bandgap of -0.3eV, so it is unusable as a detector material. However, CdTe has a bandgap of 1.5eV, corresponding to a wavelength of 830nm. When alloyed together, Hg$_{1-x}$Cd$_x$Te can be composed to span a wide range of wavelengths from about 1.5μm to about 10μm. Therefore, it is a very versatile material. However, one of the challenges with MCT detectors is the lack of suitably lattice-matched and inexpensive substrates. It is typically grown on CdTe or CdZnTe substrates, which are difficult to process and expensive. Similar to InSb, MCT detectors have to be used in a cryo cooled environment to reduce the intrinsic carrier density. Nevertheless, MCT is one of the most widely used detector material for infrared detection.

- Germanium shares many similarities with silicon. It has a bandgap of 0.66eV, which corresponds to a wavelength of 1.87μm. Therefore, it competes in the same the application space as In$_{0.53}$Ga$_{0.47}$As. Ge can be grown as a single crystal substrate, whereas In$_{0.53}$Ga$_{0.47}$As has to be grown as a crystalline thin film on an InP substrate. Despite this advantage for Ge, In$_{0.53}$Ga$_{0.47}$As has become a more dominant material in this spectral range because it has better noise characteristics. Additionally, In$_{0.53}$Ga$_{0.47}$As absorption spectrum extends past 1600nm, whereas the absorption in Ge reaches a peak near 1550nm, and then drops sharply. As a result, nearly all telecommunication products operating at 1.55μm use In$_{0.53}$Ga$_{0.47}$As instead of Ge.

Modulation Response of PN and PIN Photodiodes

The response times of photodiodes are determined primarily by the carrier transit time. It is dominated by the slowest carrier in the system. It can be written as

$$|r| = \frac{\mathcal{R}}{\sqrt{1 + (\omega t_r)^2}}, \tag{49}$$

where t_r is the largest transit time in the system. We also argued that the largest transit time is equal to the carrier lifetime τ_n or τ_p (the collection distances Z_n and Z_p are determined by the minority carrier lifetimes). One way to increase the modulation response of photodiodes is by decreasing the minority carrier lifetime. In silicon, this is on the order of 10μs, depending largely on doping concentration and material quality. In In$_{0.53}$Ga$_{0.47}$As, these lifetimes are on the order of 10ns. Therefore, In$_{0.53}$Ga$_{0.47}$As photodiodes are intrinsically faster than silicon photodiodes. Using these numbers, we can calculate that the 3dB frequency in a silicon photodetector is 27.5MHz, and with In$_{0.53}$Ga$_{0.47}$As it is 27.5GHz. The recombination lifetimes in germanium is also on the order of 10μs, which is another reason why In$_{0.53}$Ga$_{0.47}$As is preferred over Ge.

From equation (49), the 3dB frequency of the photodetector can be expressed as

$$f_{3dB} = \frac{\sqrt{3}}{2\pi t_r}. \tag{50}$$

The modulation response is identical to the LED modulation response we examined earlier.

Besides choosing a material with a smaller recombination lifetime, we can also modify the absorbing region in a semiconductor to reduce the transit time. The recombination lifetime becomes the maximum transit time only when excitons are created at large distances from the space charge layer. If we can limit the generation of excitons closer to the space charge layer, it will become possible to reduce the transit time regardless of the recombination lifetime. This was the main idea behind heterostructure photodiodes.

Heterostructure Photodiodes

We examined heterostructures under LEDs and laser diodes. Besides their ability to confine carriers and reach population inversion at lower injection currents, they have advantages in photodetectors as well. Specifically, they can be used to limit absorption only to areas of high fields (space charge regions). This can significantly increase the collection efficiency as well as the transit time of the collected carriers.

For example, consider an In$_{0.53}$Ga$_{0.47}$As/InP photodiode structure as shown in Fig 13. The In$_{0.53}$Ga$_{0.47}$As layer is sandwiched between an n-type InP substrate and a top layer of p-type InP. Photons with energy smaller than 1.27eV, but larger than 0.74eV will travel through the top InP layer without any absorption. Therefore, the effect of the junction depth has been eliminated, Absorption begins when the photons enter the In$_{0.53}$Ga$_{0.47}$As I-layer. Once the photons exit the In$_{0.53}$Ga$_{0.47}$As layer, they will travel without any further absorption. Therefore, we can consider the InP layers as transparent windows surrounding the active material. Additionally, since the I-region will be fully depleted, transit time through this layer will be the only factor contributing to the total transit time. Since the electric field is large in the space charge layer, this transit time will be very short. Since no photons will be absorbed in the p-type and n-type InP layers, we will not have any contributions from the transit time through the outlying n- and p- regions with low electric fields. In other words, Z_n and Z_p will be zero.

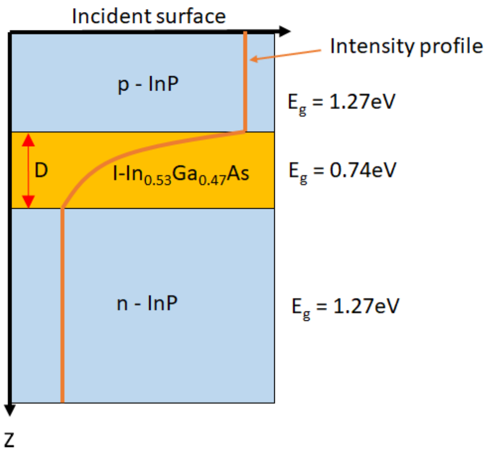

Figure 13: In$_{0.53}$Ga$_{0.47}$As/InP heterostructure PIN photodiode

The expression for collection efficiency will become

$$\eta_c = 1 - e^{-\alpha D}, \tag{51}$$

where D is the thickness of the In$_{0.53}$Ga$_{0.47}$As I-layer. Compared to equation (16), we can see that there is no contribution due to absorption in the region above the photodiode. The average electric field in the I-layer will be

$$\bar{E} = \frac{V_{bi} - V_a}{D}. \tag{52}$$

The drift velocity will be

$$v_p = \mu_p \bar{E}. \tag{53}$$

This equation is valied only at low field values. At high fields, the velocity will reach saturation. Nevertheless, once the drift velocity is known, the transit time can be written as

$$t_r = \frac{D}{v_p}. \tag{54}$$

Here we have assumed that holes travel slower than electrons, because the longest transit time will dominate the whole process.

One disadvantage of the collection widths Z_n and Z_p being zero is that it will also reduce collection efficiency. Any photons that are transmitted out the bottom InP will not be absorbed. Fortunately, compared to silicon, the absorption coefficient of In$_{0.53}$Ga$_{0.47}$As is very high, which partly compensates for this effect. Nevertheless, the transit time could be reduced even further by making the In$_{0.53}$Ga$_{0.47}$As I-layer thinner. This can result in very high speed photodetectors, at the expense of a lower collection efficiency. However, some of these photons can be recovered by utilizing a reflector on the substrate to recycle it through the In$_{0.53}$Ga$_{0.47}$As layer. This effectively creates a resonant cavity, and are known as resonant-cavity-enhanced (RCE) photodetetcors.

Example

For example, if the total voltage $V_{bi} - V_a$ is equal to 5V, and the In$_{0.53}$Ga$_{0.47}$As layer thickness is 5μm, the electric field in the space charge layer will be

$$E = \frac{V_{bi} - V_a}{D} = \frac{5}{5 \times 10^{-4}} = 1 \times 10^4 \text{ V/cm}. \tag{55}$$

The hole mobility in In$_{0.53}$Ga$_{0.47}$As is 250 cm^2/Vs. The drift velocity of holes in the space charge layer will be

$$v_p = \mu_p E = 2.5 \times 10^6 = \text{cm/s} \tag{56}$$

However, the saturation velocity of holes in In$_{0.53}$Ga$_{0.47}$As is 2×10^5cm/s. Clearly we cannot exceed the saturation velocity. Therefore, the hole velocity will be saturated, resulting in a transit time of

$$t_r = \frac{D}{v_p} = \frac{5 \times 10^{-4}}{2 \times 10^5} = 2.5 \text{ ns}. \tag{57}$$

The resulting 3dB modulation frequency is

$$f_{3dB} = \frac{\sqrt{3}}{2\pi\tau} = \frac{\sqrt{3}}{2\pi \times 2.5 \times 10^{-9}} = 110 \text{ MHz}. \tag{58}$$

Additionally, assuming a wavelength of 1.55μm, we can also calculate the collection efficiency. The attenuation coefficient in In$_{0.53}$Ga$_{0.47}$As at 1.55μm wavelength is about $\alpha = 4000$/cm. This results in a collection efficiency of

$$\eta_c = 1 - e^{-\alpha D} = 0.86. \tag{59}$$

Assuming an internal quantum efficiency of $\eta_i = 1.0$, the responsivity becomes:

$$\mathcal{R} = \eta_i \eta_c \frac{q}{h\nu} = 1.08 \text{ A/W}. \tag{60}$$

It is possible to reduce the I-layer width and increase the modulation bandwidth. For example, if we reduce the In$_{0.53}$Ga$_{0.47}$As layer thickness to 500nm, the results will be:

$$E = \frac{V_{bi} - V_a}{D} = \frac{5}{5 \times 10^{-5}} = 1 \times 10^5 \text{ V/cm}. \tag{61}$$

The hole velocity will still be saturated. Therefore, the transit time will be

$$t_r = \frac{D}{v_p} = \frac{5 \times 10^{-5}}{2 \times 10^5} = 0.25 \text{ ns}. \tag{62}$$

The resulting 3dB modulation and collection efficiency will be

$$f_{3dB} = \frac{\sqrt{3}}{2\pi\tau} = \frac{\sqrt{3}}{2\pi \times 0.25 \times 10^{-9}} = 1.1 \text{ GHz} \tag{63}$$

$$\eta_c = 1 - e^{-\alpha D} = 0.18. \tag{64}$$

As we can see, it is possible to achieve GHz response speed from heterostructure diodes by trading off the collection efficiency with speed.

Photodiode Packages

Photodiodes come in a large number of different packages. The TO package is a very common package. The symbol TO stands for Transistor Outline, but this terminology is primarily historic and has little to do with the component inside the package. A photodiode is clearly a two-terminal device, but some packages contain three pins. The third pin is the casing material (which is typically grounded). Some packages will also contain an integrated lens on the input window to focus the light onto the detector surface.

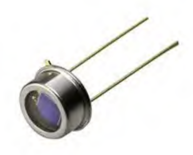

Figure 14: A TO-5 package photodiode. Source: Hamamatsu

It is also very common to find photodiodes in a plastic dome package similar to LEDs. The same features of the dome that improve the extraction efficiency of LEDs also improves the coupling efficiency of photodiodes. This configuration is particularly useful for detecting ambient light arriving from all angles. All angles within the frontal hemisphere will be converted a smaller cone of angles incident on the photodetector, which increases coupling compared to glacing angle incidence.

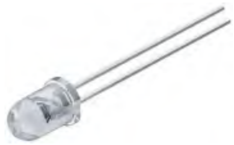

Figure 15: A plastic photodiode package similar to LEDs. Source: Osram

Homework 7

1. Consider a silicon PN junction photodiode with an internal quantum efficiency of 0.95, and with a surface area of 500μm x 500μm. The substrate is a low-doped float zone silicon that is n-type with $N_D = 10^{15}$/cm^3. The top side is doped p-type with $N_A = 10^{18}$/cm^3, forming a junction 1μm below the incident surface. The incident surface is AR coated for normal incidence such that all of the light is absorbed in the semiconductor. The photodiode is reverse biased at a voltage of -5V. Look up the relevant parameters for silicon from http://www.matprop.ru/.

   - Assuming standard material values for silicon, and a long-diode approximation for the n-side, an equivalent shunt resistance of 1 GΩ, calculate the dark current (reverse sauration current of this diode).
   - For an incident intensity of 100μW/cm^2, calculate the collection width Z, and the collection efficiency η_c at a wavelength of 632.8 nm.
   - Calculate the external quantum efficiency and the responsivity at a wavelength of 632.8 nm.
   - Estimate the 3dB modulation bandwidth of this photodiode.

     Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val etaI = 0.95
    val Area = 500e-4*500e-4
    val d = 1.0e-4
    val ND = 1.0e15
    val NA = 1.0e18
    val Va = -5.0
    val k = 8.6173303e-5
    val q = 1.602e-19
    val T = 300.0
    val Vt = k*T
    val muN = 1450.0      //at 1e15
    val muP = 300.0       //at 1e18
    val tauP = 200.0e-6     //at 1e15
    val tauN = 10.0e-6      //at 1e18
    val Dn = muN*Vt
    val Dp = muP*Vt
    val Lp = (Dp*tauP).pow(0.5)
    val ni = 1.0e10
    val Is = q*Area*(Dn/d*ni.pow(2)/NA + Dp/Lp*ni.pow(2)/ND)
    println(”Is = ${”%.2e”.format(Is)} A”)
    val Rshunt = 1.0e9
    val Idark = abs((Is + abs(Va)/Rshunt)*(exp(Va/Vt)-1.0))
    println(”Idark = ${”%.2e”.format(Idark)} A”)
    val Vbi = Vt*ln(NA*ND/ni.pow(2))
    println(”Vbi = ${”%.2f”.format(Vbi)} V”)
    val epsilon0 = 8.85e-14        //F/cm
    val epsilonr = 11.68
    val xp = (2.0*epsilonr*epsilon0/q*ND/NA*(Vbi-Va)/(NA+ND)).pow(0.5)
    println(”xp = ${”%.2f”.format(xp*1.0e7)} nm”)
    val xn = (2.0*epsilonr*epsilon0/q*NA/ND*(Vbi-Va)/(NA+ND)).pow(0.5)
    println(”xn = ${”%.2f”.format(xn*1.0e7)} nm”)
    val wavelength = 0.632         //microns
    val kappa = 0.015953
    val alpha = 4.0*PI/(wavelength*1.0e-4)*kappa
    println(”alpha = ${”%.2f”.format(alpha)} /cm”)
    val Intensity = 100.0e−6       //uW/cm2
    val Power = Intensity ∗ Area
    println(”Power = ${”.2f”.format(Power)} W”)
    var Responsivity = 0.25   //Initial assumption (A/W)
    println(”Responsivity = ${”%.2f”.format(Responsivity)} A/W”)
    var I = 1.0
    var i = 1
    while (abs((Responsivity ∗ Power − I)/I) > 1.0e−3){
        I = Responsivity ∗ Power
        println(”————————————————————————————————————————————”)
        println(”Iteration = $i”)
        println(”I = ${”%.2e”.format(I)} A”)
        val Znmax = I∗tauN/(q∗Area∗NA)∗muN/muP
        println(”Znmax = ${”%.2f”.format(Znmax∗1.0e7)} nm”)
        val Zpmax = I∗tauP/(q∗Area∗ND)∗muP/muN
        println(”Zpmax = ${”%.2f”.format(Zpmax∗1.0e7)} nm”)
        val Z = Znmax+Zpmax+xn+xp
        println(”Z = ${”%.2f”.format(Z∗1.0e4)} um”)
        val Ep = I/(q∗Area∗muP∗NA)
        println(”Ep = ${”%.2e”.format(Ep)} V/cm”)
        val En = I/(q∗Area∗muN∗ND)
        println(”En = ${”%.2e”.format(En)} V/cm”)
        val Eav = (Vbi−Va)/(xn+xp)
        println(”Eav = ${”%.2e”.format(Eav)} V/cm”)
        val etaC = exp(−alpha∗d)∗(1.0 − exp(−alpha∗Z))
        println(”eta_c = ${”%.2f”.format(etaC)} /cm”)
        val etaE = etaC∗etaI
        println(”eta_e = ${”%.2f”.format(etaE)}”)
        val hv = 1.24/wavelength
        Responsivity = etaE/hv
        println(”Responsivity = ${”%.2f”.format(Responsivity)} A/W”)
        i++
        }
    val dB3 = 3.0.pow(0.5)/(2.0∗PI∗tauP)
    println(”3dB Freq = ${”%.2e”.format(dB3)} Hz”)
}


Output:
Is = 2.29e−14 A
Idark = 5.00e−09 A
Vbi = 0.77 V
xp = 2.73 nm
xn = 2728.30 nm
alpha = 3172.01 /cm
Power = .2f W
Responsivity = 0.25 A/W
————————————————————————————————————————————
Iteration = 1
I = 6.25e−08 A
Znmax = 0.08 nm
Zpmax = 64.57 nm
Z = 2.80 um
Ep = 5.20e−07 V/cm
En = 1.08e−04 V/cm
Eav = 2.11e+04 V/cm
eta_c = 0.43 /cm
eta_e = 0.41
Responsivity = 0.21 A/W
————————————————————————————————————————————
Iteration = 2
I = 5.18e−08 A
Znmax = 0.06 nm
```

```
Zpmax = 53.55 nm
Z = 2.78 um
Ep = 4.31e-07 V/cm
En = 8.93e-05 V/cm
Eav = 2.11e+04 V/cm
eta_c = 0.43 /cm
eta_e = 0.41
Responsivity = 0.21 A/W
```

```
Iteration = 3
I = 5.17e-08 A
Znmax = 0.06 nm
Zpmax = 53.42 nm
Z = 2.78 um
Ep = 4.30e-07 V/cm
En = 8.90e-05 V/cm
Eav = 2.11e+04 V/cm
eta_c = 0.43 /cm
eta_e = 0.41
Responsivity = 0.21 A/W
3dB Freq = 1.38e+03 Hz
```

- Look up the standard optical dispersion values for silicon (for example, from refractiveindex.info), and plot the responsivity spectrum of this photodiode between 400nm and 1100nm assuming the same value of photocurrent as before.

Code (will not run on Kotlin Playground - requires local file of silicon dispersion)

```
import java.io.File
import kotlin.math.*
//Andrew Sarangan

fun DoubleArray.interpolate(aVal: Double, bArray: DoubleArray): Double {
    var locA: Int = 0
    this.map { it - aVal }
        .forEachIndexed { i, v ->
            if (i != this.size - 1) {
                if (v * this[i + 1] <= 0.0) {
                    locA = i
                }
            }
        }
    val x1 = aVal - this[locA]
    val x2 = this[locA + 1] - aVal
    return bArray[locA] + x1 / (x1 + x2) * (bArray[locA + 1] - bArray[locA])
}

fun readIndex(lambdaNM: DoubleArray, fileName: String): Pair<DoubleArray,
    DoubleArray> {
    val lines =
        File(fileName).readLines() // This is a 3-column file of wavelength(nm), n,
    k

    val fileLambda = DoubleArray(lines.size)
    val fileN = DoubleArray(lines.size)
    val fileK = DoubleArray(lines.size)
    var i = 0
    for (line in lines) {
        val match = Regex("([\\d\\.-]+)[\\t\\s]+([\\d\\.-]+)[\\t\\s]+([\\d\\.-]+)").
    find(line)
        fileLambda[i] = match!!.groupValues[1].toDouble()
        fileN[i] = match.groupValues[2].toDouble()
        fileK[i] = match.groupValues[3].toDouble()
        i++
    }

    val n = lambdaNM.map { fileLambda.interpolate(it, fileN) }.toDoubleArray()
    val k = lambdaNM.map { fileLambda.interpolate(it, fileK) }.toDoubleArray()
    return Pair<DoubleArray, DoubleArray>(n, k)
}

fun main() {
    val nLambda = 1000 // Number of wavelength points
    val lambda1 = 400.0
    val lambda2 = 1100.0
    val dLambda = (lambda2 - lambda1) / nLambda.toDouble()
    val lambdaNM = DoubleArray(nLambda) { lambda1 + it * dLambda }

    val (n, k) = readIndex(lambdaNM, ”silicon.nk”)
    val alpha = k.zip(lambdaNM) { k, lambdaNM -> 4.0 * PI / (lambdaNM / 1.0e7) * k
    }.toDoubleArray()

    val Z = 2.78e-4
    val d = 1.0e-4
    val etaI = 0.95
    val R = lambdaNM
            .mapIndexed { i, lambda ->
                val etaC = exp(-alpha[i] * d) * (1.0 - exp(-alpha[i] * Z))
                val etaE = etaC * etaI
                etaE * (lambda / 1.0e3) / 1.24
            }
            .toDoubleArray()

    R.forEachIndexed { i, R -> println(”${lambdaNM[i]}\t$R\t${alpha[i]}”) }
}
```

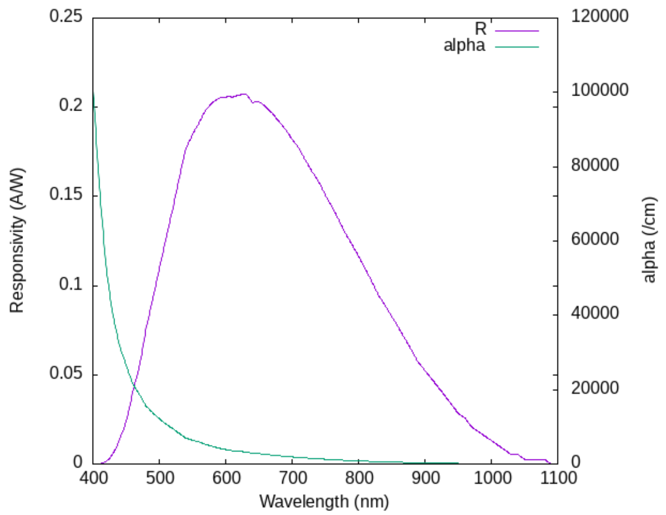

Figure 16: Solution

2. - A silicon photodiode is desired for a medical scanning application at a wavelength of 400 nm. Using the same material parameters as given above, design an appropriate structure that would optimize the responsivity for 400 nm.

    Junction depth is the primary factor affecting the short wavelength response. Though not possible in practice, we can make the junction depth equal to zero to examine how it would affect the responsivity.

- Explain how PIN photodiodes increase collection efficiency compared to PN photodiodes.

Lower doping of the I region enlarges the depletion region, resulting in a larger collection volume.

3. An In$_{0.53}$Ga$_{0.47}$As/InP photodiode has a responsivity of 0.9 A/W at a wavelength of 1350nm. Assuming the internal quantum efficiency is 100%, what is the thickness of the In$_{0.53}$Ga$_{0.47}$As layer?

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val kappa = 0.104
    val wavelengthUM = 1.450
    val alpha = 4.0*PI/(wavelengthUM*1.0e-4)*kappa
    val etaI = 1.0
    val R = 0.9
    val etaC = R/etaI*1.24/wavelengthUM
    println(”eta_c = ${”%.2f”.format(etaC)}”)
    val D = 1.0/alpha * ln(1.0/(1.0 - etaC))
    println(”D = ${”%.2f”.format(D*1.0e7)} nm”)
}

>>eta_c = 0.77
>>D = 1628.93 nm
```

Avalanche Photodiodes

Avalanche photodiodes (also known as APDs) are photodiodes with internal gain. With PN and PIN photodiodes, we saw that the external quantum efficiency (EQE) is always smaller than 1.0. In APDs, the internal gain can produce an EQE larger than 1.0. This is accomplished by biasing the photodiode with a high enough reverse bias voltage to operate it close to its avalanche breakdown regime. Avalanche breakdown is a mechanism where each carrier (electron and hole) accelerate and gain enough energy such that they produce more electron-holes pairs by impact ionization. These are known as secondary electrons and holes. These secondary electrons and holes accelerate and create even more electron-hole pairs. Hence an avalanche of carriers can be created from the first few electron-hole pairs created by the incident photons (which are known as primary electrons and holes).

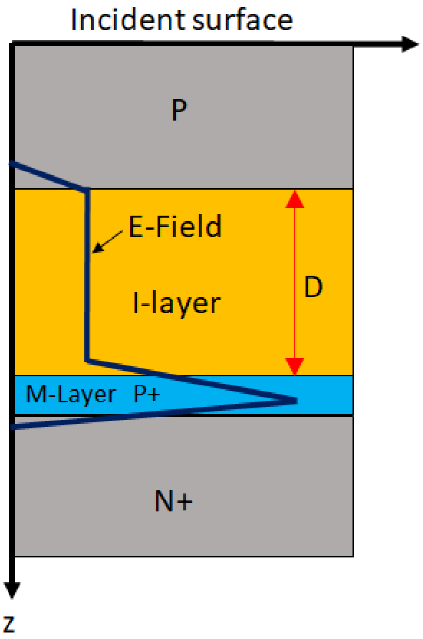

Figure 17: Avalanche photodiode configuration with a thin P+ multiplication region at the N+ interface.

Avalanche multiplication is produced by creating a thin layer (M-layer) inside the PIN photodiode structure with a very large internal bias field. This requires a junction with a narrow space charge width. Since the space charge width is inversely proportional to the doping densities (equations (28) and (29)), we can create such a layer at the junction between an N+ region and a P+ region. This layer acts basically as a multiplier of photogenerated carriers produced in the I-layer.

Electrons and holes have different ionization coefficients α_n and α_p, which vary greatly from one semiconductor material to another. The ratio between the two coefficients has a profound impact on the gain as well as the frequency response of the device.

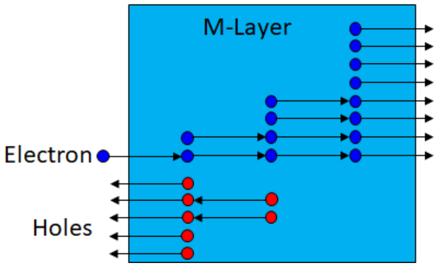

Figure 18: Illustration of carrier multiplication

The photogenerated carriers in the I-layer will enter the M-layer as electrons, while the photogenerated holes will enter the top P-type layer. The electrons entering the M-layer will undergo impact ionization resulting in many more electrons and holes. All of these electrons will exit the M-layer and enter the N+ layer below it, while all of the excess holes generated by impact ionization will enter the I-layer above it.

Referring to Fig 19, we can write the growth of the electron drift current in the M-layer due to impact ionization as

$$\frac{dJ_n(z)}{dz} = \alpha_n J_n(z) + \alpha_p J_p(z)\,. \tag{65}$$

The total electron drift current exiting the M-layer will be the sum of the original electron current injected from the I-layer into the M-layer, plus the excess electron current created by the impact

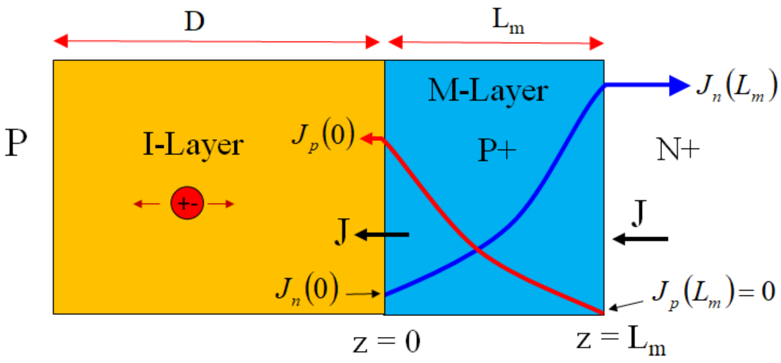

Figure 19: Transit time model of an avalanche photodiode

ionization (which is equal to the excess hole current created by the same impact ionization):

$$J_n\left(L_m\right)=J_n\left(z\right)+J_p\left(z\right). \tag{66}$$

From this, we can write

$$J_p\left(z\right)=J_n\left(L_m\right)-J_n\left(z\right). \tag{67}$$

Equation (67) can be substituted into equation (65) to produce

$$\frac{dJ_n\left(z\right)}{dz}=\left(\alpha_n-\alpha_p\right)J_n\left(z\right)+\alpha_p J_n\left(L_m\right). \tag{68}$$

This differential equation can be solved to give:

$$\begin{aligned}
z\Big|_0^{L_m} &= \frac{1}{\left(\alpha_n-\alpha_p\right)}\ln\left[\left(\alpha_n-\alpha_p\right)J_n\left(z\right)+\alpha_p J_n\left(L_m\right)\right]\Bigg|_{J_n(0)}^{J_n(L_m)} && (69)\\
J_n\left(L_m\right) &= J_n\left(0\right)\left[\frac{\left(\alpha_n-\alpha_p\right)}{\alpha_n e^{-\left(\alpha_n-\alpha_p\right)L_m}-\alpha_p}\right] && (70)\\
&= J_n\left(0\right)\underbrace{\left[\frac{\left(1-\frac{\alpha_p}{\alpha_n}\right)}{e^{-\left(1-\frac{\alpha_p}{\alpha_n}\right)\alpha_n L_m}-\frac{\alpha_p}{\alpha_n}}\right]}_{G} && (71)
\end{aligned}$$

We can associate the terms inside the square brackets as the avalanche gain, G. Additionally, we can also express the gain in terms of the ratio of the ionization coefficients

$$k=\frac{\alpha_p}{\alpha_n}, \tag{72}$$

resulting in

$$G=\frac{(1-k)}{e^{-(1-k)\alpha_n L_m}-k}. \tag{73}$$

Depending on the value of the ratio k, we can make the following observations:

$$\begin{aligned}
k=0 &\quad:\quad G=e^{\alpha_n L_m} && (74)\\
k=\infty &\quad:\quad G=1 && (75)\\
k=1 &\quad:\quad G=\frac{1}{1-\alpha_n L_m}. && (76)
\end{aligned}$$

When $k = 0$, the multiplication factor grows exponentially with the thickness of the M-layer. When $k = \infty$, there will be no gain regardless of the M-layer thickness. This is understandable because $k = \frac{\alpha_p}{\alpha_n} = \infty$ implies that α_n is negligibly small compared to α_p. This results in no gain for electrons, which is the photogenerated carrier type being injected into the M-layer. When $k = 1$, equation (73) becomes indeterminate. However, we can use L'Hospital's rule to find its convergence value as $\frac{1}{1-\alpha_n L_m}$.

It is also possible to have a situation that will produce an indefinitely large gain. This will occur when

$$e^{-(1-k)\alpha_n L_m} = k. \tag{77}$$

This is an unstable condition, and must be avoided. Re-arranging equation (77), we can get

$$\alpha_n L_m = \frac{1}{1-k} \ln\left(\frac{1}{k}\right). \tag{78}$$

As k gets smaller, $\alpha_n L_m$ has to become very large for this instability to occur. In practical device dimensions, $k \to 0$ virtually guarantees that this condition will never occur except in very long devices.

The value of k is a material property, and not something we can control. Materials with $k = 0$, or small values of k are naturally better for APDs because they provide a classical amplification factor with predictable gain values. Additionally, the excess noise factor F of the amplifier (to be discussed later) is directly related to the APD gain and the k value as follows:

$$F = kM + (1-k)\left(2 - \frac{1}{M}\right). \tag{79}$$

A smaller k value would therefore lead to a smaller excess noise factor, which is typically desired. The table below lists the k values for commonly used semiconductors.

Semiconductor	k
Si	0.02-0.05
Ge	0.7-1.0
In$_{0.53}$Ga$_{0.47}$As	0.5-0.7

Silicon has the most optimal value of k. Ge is worse and is rarely used in APDs. In$_{0.53}$Ga$_{0.47}$As is also widely used in 1.3μm & 1.5μm APDs for fiber telecommunication applications.

An important aspect in APDs is the build-up time needed to produce gain. This significantly adds to the transit time, and hence affects the frequency response of the device. In a regular heterostructure PIN photodiode, the transit time through the space charge layer was

$$t_r = \frac{D}{v_p}, \tag{80}$$

where D is the thickness of the I-layer and v_p is the drift velocity of the holes. We consider holes instead of electrons for calculating the transit time because when both carriers move simultaneously (in opposite directions), it is the slowest carrier that will determine the overall transit time. In this case we are assuming the drift velocity of holes is smaller than that of electrons. While that is true in PIN photodiodes, in APDs, the slowest transit process consists of three terms:

$$t_r = \frac{D}{v_n} + \tau_m + \frac{D}{v_p}, \tag{81}$$

where τ_m is the build-up time in the M-layer. Considering a single electron generated at the left edge of the I-layer, the first term $\frac{D}{v_n}$ is the time it takes for it to reach the M-layer. τ_m is the time between when that electron enters the M-layer, and the time when the last hole from ionization exits the M-layer back into the I-layer. Finally, $\frac{D}{v_p}$ is the time required for that last hole to travel through the I-layer and exit to the left. The build-up time τ_m is a bit involved to derive. It can, however, be expressed approximately as

$$\tau_m \approx \frac{kGL_m}{v_n} + \frac{L_m}{v_p}. \tag{82}$$

Because the M-layer is designed to contain a very large electric field, the values of v_n and v_p will be well into the saturation regime. Therefore, we can write this as

$$\tau_m \approx \frac{kGL_m}{v_{n,sat}} + \frac{L_m}{v_{p,sat}}. \tag{83}$$

The internal gain in APDs make them useful in applications where the photocurrent is small. As we will see later, amplifying the signal immediately following detection helps to preserve the quality of the signal during subsequent processing. Compared to photoconductors, the transit time in APDs can be much smaller if they are built as heterostructures, which can lead to high detection speeds.

Example

Consider an APD with an I-layer thickness of 25μm, M-layer thickness of 3μm, and is reverse biased at -5V such that $\alpha_n = 1 \times 10^5$/cm, and $k = 0.1$. The built-in voltage is 0.7V.

The gain can be calculated as

$$G = \frac{(1-k)}{e^{-(1-k)\alpha_n L_m} - k} = 13.8. \tag{84}$$

The electric field in the I-layer will be

$$E = \frac{V_{bi} - V_a}{D} = \frac{5.7}{25 \times 10^{-4}} = 2280\text{V/cm}. \tag{85}$$

The drift velocity of electrons and holes in the I-layer will be

$$v_n = \mu_n E = 1400 \times 2280 = 3.2 \times 10^6 \text{ cm}^2\text{/s} \tag{86}$$
$$v_p = \mu_p E = 450 \times 2280 = 1 \times 10^6 \text{ cm}^2\text{/s}. \tag{87}$$

These are well below the saturation velocities of 1×10^7 cm/s and 7×10^6 cm/s for the electrons and holes, respectively. Therefore, the electron and hole transit time through the space charge layer will be

$$\frac{D}{v_n} = 0.78 \text{ ns} \tag{88}$$
$$\frac{D}{v_p} = 2.5 \text{ ns}. \tag{89}$$

The M-layer build-up time will be

$$\tau_m \approx \frac{kGL_m}{v_{n,sat}} + \frac{L_m}{v_{p,sat}} = 56 \text{ ps}. \tag{90}$$

Finally, the total transit time will be

$$t_r = \frac{D}{v_n} + \tau_m + \frac{D}{v_p} = 3.3 \text{ ns}, \tag{91}$$

resulting in a 3dB modulation frequency of

$$f_{3dB} = \frac{\sqrt{3}}{2\pi t_r} = \frac{\sqrt{3}}{2\pi \times 4.6 \times 10^{-9}} = 84 \text{ MHz}. \tag{92}$$

Geiger Mode APD

The APDs discussed so far are known as linear-mode APDs. The gain is treated as a constant such that the number of electrons collected at the output is a multiple of the photocurrent. Hence they are useful in analog and digital communication applications. APDs can also be operated in another mode, known as Geiger-mode. In this regime, the gain is purposely designed to be extremely large, and close to the unstable value outlined in equation (77). When photons are incident, the large gain will trigger a nearly continuous avalanche of electrons - the current flow never stops once it is triggered (i.e., infinite gain). This is useful in applications of detecting signals above a minimum threshold, usually photon counting applications in certain medical and quantum encryption technologies.

Photoconductors

Photoconductors are made from undoped, or lightly doped semiconductors, with a fairly low carrier concentration. When light is incident, extra electron-hole pairs will be generated. This will result in an increased carrier concentration, and an increased conductance. The increased conductance will last until the last of the photo-generated carriers have been removed from the semiconductor.

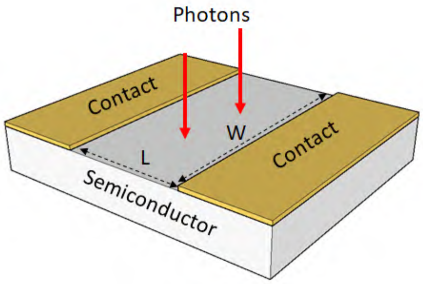

Figure 20: Simple photoconductor

Consider a simple photoconductor configuration as shown in Fig 20. The surface of the material is typically anti-reflection coated to minimize reflection and to maximize absorption. If the incident intensity is I_i W/cm^2, the intensity profile inside the semiconductor will decay with depth as

$$I\left(z\right)=I_{i}e^{-\alpha z} \tag{93}$$

where α is the attenuation coefficient of light in the material. As discussed earlier, this attenuation will be a strong function of wavelength. Short wavelengths will be absorbed very quickly, while long wavelengths will penetrate deeper into the semiconductor, as illustrated in Fig 21.

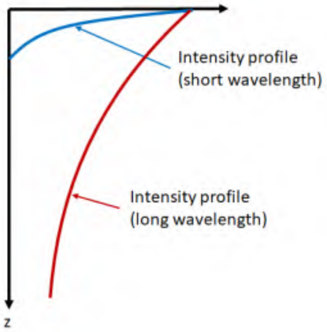

Figure 21: Intensity profile inside the semiconductor for short and long wavelength radiations.

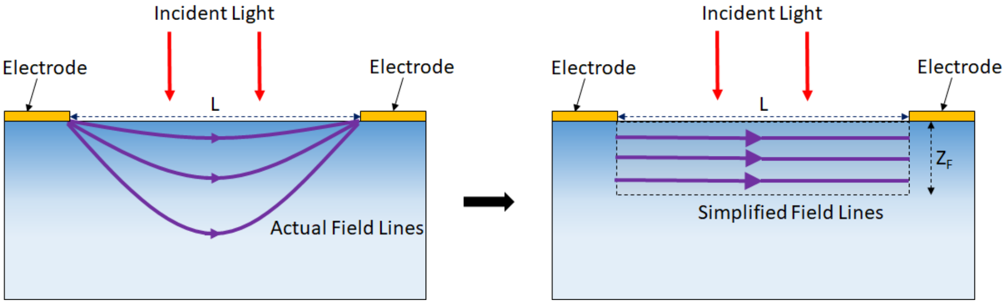

Figure 22: Illustration of field distribution in the photoconductor and the approximate representation of this field

When a voltage is applied at the two electrodes, a two-dimensional electric field distribution will be created as shown in Fig 22. This field distribution can be determined only by solving the Poisson's equation in conjunction with the current continuity equation. This is not a trivial task. However, we can make several major simplifications that can allow us to proceed forward. Because the carrier concentration has its highest value near the surface, and because the electric field strength between the two electrodes is also highest near the surface, the vast majority of current will flow horizontally near the surface. The field value near the surface will be $\frac{\Delta V}{L}$, where

ΔV is the applied voltage difference between the electrodes, and L is the distance between the electrodes. The field will decline as we move deeper into the substrate. The exact field distribution will depend on the permittivity of the material and the electrode spacing, but we can make some approximations. We will assume that the electric field is uniform and parallel with a value of $\frac{\Delta V}{L}$ up to a depth of Z_F inside the semiconductor.

Each exciton created within the depth Z_F will be separated into an electron and hole. This is the collection region of the photoconductor. Excitons created outside this region will not be separated. Of course, in reality the collection region will not be that distinctly separated, but we will make this assumption to make the calculations easier. Therefore, the collection efficiency becomes:

$$\eta_c = \left(1 - e^{-\alpha Z_F}\right). \tag{94}$$

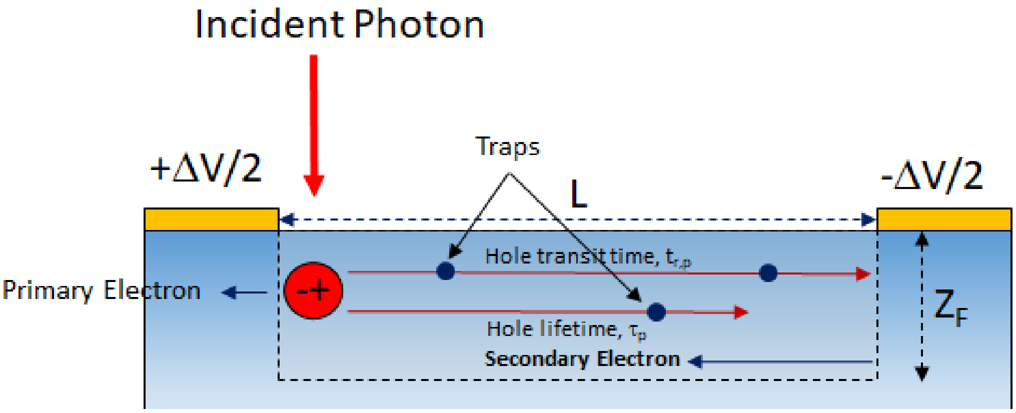

Figure 23: Illustration of primary hole and secondary electron transport

Each exciton will release an electron and a hole. Consider, for example, this exciton to be created at the left edge of the photoconductor. After the exciton is separated, the hole will start drifting towards the right electrode. The electron from the same exciton will quickly exit the semiconductor to the left. However, to maintain charge neutrality, another electron will enter from the right electrode and proceed towards the left electrode. This is known as a secondary electron, because it is not the same electron that was created by the photon. If the secondary electron reaches the left electrode while the hole is still in transit, another secondary electron may be emitted. This process will continue until the hole recombines and disappears, or reaches the right electrode. Therefore, multiple electrons could transit the device before the hole disappears. In other words, multiple electrons (one primary electron and multiple secondary electrons) will be collected at the terminal for each exciton that was created. This results in what is known as the photoconductive gain, which can be written as

$$G = \frac{\tau}{t_{r,n}} \tag{95}$$

where $t_{r,n}$ is the electron transit time defined as

$$t_{r,n} = \frac{L}{\mu_n E} = \frac{L^2}{\mu_n \Delta V}, \tag{96}$$

and τ is the *smaller* of either the hole lifetime or the hole transit time

$$\tau = \tau_p \text{ or } \frac{L^2}{\mu_p \Delta V}. \tag{97}$$

In practice, trap states will be abundant in the semiconductor, especially near the surface, which will effectively reducing the hole mobility, which leads to a larger gain than what the equation predicts.

In this description, we have assumed that $\mu_n\tau_n \gg \mu_p\tau_p$. In other words, we have assumed that the electron can drift to a much larger distance than the hole. This is true in almost all semiconductors, but the opposite condition can also be considered, which will lead to an inverse relationship for G as compared to equation (95).

The external quantum efficiency becomes

$$\eta_e = \eta_i\eta_c G, \tag{98}$$

and the current output will be

$$I = \frac{P_i}{h\nu}\eta_e \tag{99}$$

from which we can get the responsivity

$$\mathcal{R} = \frac{\eta_e}{h\nu}. \tag{100}$$

The dark current (with no illumination) will be due to the intrinsic carrier concentration in the semiconductor. This can be calculated from

$$\begin{aligned} I_D &= qZ_FWn_i\left(\mu_n+\mu_p\right)E & (101)\\ &= qZ_FWn_i\left(\mu_n+\mu_p\right)\frac{\Delta V}{L}. & (102)\end{aligned}$$

Example

Consider a photoconductor made of intrinsic crystalline silicon with an electrode spacing of $L = 500\mu$m, and a width of $W = 5$mm. The refractive index of silicon at a wavelength of $\lambda = 600$ nm is $3.943 - j0.025$. The attenuation coefficient, therefore, is $\alpha = \frac{4\pi}{\lambda} \times 0.025 = 5236$ /cm. The mobilities values are $\mu_n = 1400$cm^2/V/s and $\mu_p = 450$cm^2/V/s, and the minority carrier recombination lifetimes are $\tau_n = \tau_p = 10\mu$s. Assume $\eta_i = 0.9$, with an applied voltage of 5V, and the incident intensity is $I_i = 10$ mW/cm^2. Also assume that the collection distance is $Z_F = 5\mu$m. The intrinsic carrier concentration is 1.5×10^{10} cm^{-3}.

First, lets calculate the dark current value:

$$I_D = qZ_FWn_i\left(\mu_n+\mu_p\right)\frac{\Delta V}{L} = 0.11\mu A. \tag{103}$$

The collection efficiency will be

$$\eta_c = \left(1-e^{-\alpha Z_F}\right) = 0.93. \tag{104}$$

The electron and hole transit times will be

$$\begin{aligned} t_{r,n} &= \frac{L^2}{\mu_n\Delta V} = 0.36\mu\text{s} & (105)\\ t_{r,p} &= \frac{L^2}{\mu_p\Delta V} = 1.1\mu\text{s}. & (106)\end{aligned}$$

Since the lifetime τ_p was given as 10μs, the photoconductive gain will be limited by the hole transit time, resulting in

$$G = \frac{t_{r,p}}{t_{r,n}} = 3.1. \tag{107}$$

The external quantum efficiency becomes

$$\eta_e = \eta_i \eta_c G = 2.6. \tag{108}$$

Finally, the responsivity is

$$\mathcal{R} = \frac{\eta_e}{h\nu} = 1.25\text{A/W}. \tag{109}$$

For the given intensity of $I_i = 10$ mW/cm^2, the incident power will be

$$P_i = I_i W L = 0.25\text{mW}. \tag{110}$$

The resulting output current will be

$$I = I_d + \mathcal{R}P_i = 0.31\text{mA}. \tag{111}$$

Modulation Response of Photoconductors

The gain in photoconductors at least partly comes from a long recombination lifetime τ_p, or a hole transit time $t_{r,p}$, whichever is smaller. Unfortunately, has the effect of reducing the modulation response of photoconductors.

The amplitude of the small signal responsivity becomes:

$$|r| = \frac{\mathcal{R}}{\sqrt{1+(\omega\tau)^2}}, \tag{112}$$

where τ is the hole transit time, or the hole recombination time, whichever is smaller. This is identical to the expression we had for LEDs. Therefore, we can conclude that the 3dB nodulation frequency is

$$f_{3dB} = \frac{\sqrt{3}}{2\pi\tau}. \tag{113}$$

Example (Contd.)

In the previous example, we had $\tau = 1.1\mu$s. Therefore, the 3dB modulation frequency will be 250 kHz. The frequency response plot corresponding to this case is shown in Fig 24.

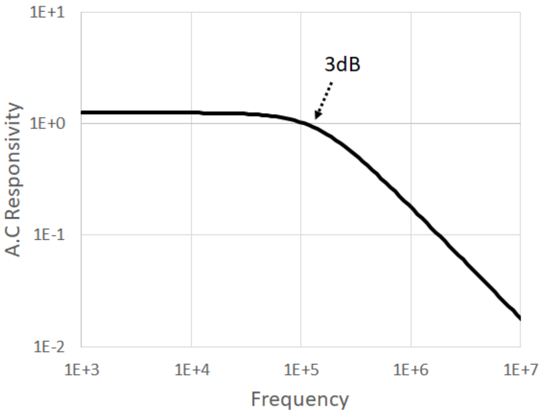

Figure 24: Modulated responsivity of the photoconductor as a function of frequency

Photoconductive Devices

The majority of commercially available photoconductors are made of CdS, CdSe or a mixture of both. The bandgap of CdS is 2.42eV, which corresponds to a wavelength of 512nm. The bandgap of CdSe is 1.74eV, which corresponds to a wavelength of 712nm. Most of the commercial devices use polycrystalline films which results in these films absorbing beyond their band edge. The typical absorption spectrum of CdS photocells is from 400nm to 700nm. For longer wavelengths, PbS ($E_g = 0.37$eV) and PbSe ($E_g = 0.27$eV) are commonly used. These correspond to bandgap wavelengths of 3.35μm, and 4.6μm, respectively.

The devices typically use an inter-digitated contact geometry to maximize the sensitivity by increasing the effective W without increasing the surface area. An example is shown in Fig 25.

For infrared wavelengths, narrow bandgap materials such as PbSe, PbS, InSb and Hg$_{1-x}$Cd$_x$Te are used. However, very narrow bandgap materials will behave like metals at room temperature (excessive intrinsic carrier concentrations) so they must be cryogenically cooled during operation.

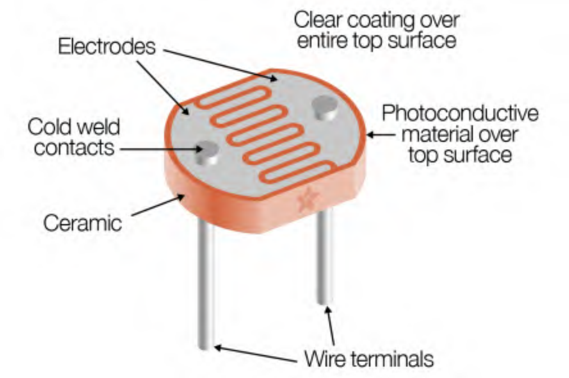
Electrodes
Clear coating over entire top surface
Cold weld contacts
Photoconductive material over top surface
Ceramic
Wire terminals

Figure 25: Typical configuration of a CdS photocell. Source: Adafruit.

One of the drawbacks of photoconductors compared to photodiodes is the higher dark current. As we can see from equation (102), the dark current is dictated by the intrinsic carrier concentration. Ideally, the intrinsic carrier concentration should be as small as possible. Even though silicon has an intrisic carrier concentration of 1.5×10^{10}cm^{-3} at room temperature, in practice, this is extremely difficult to realize. There is always some level of unintentional doping from the crystal growth process, and carrier concentrations are usually a factor of two or three above this theoretical value. For example, in the above example, if the background doping of the silicon was 1×10^{12}cm^{-3} (which is still considered lightly-doped), the dark current will be 100 times higher, or 11μA. Photodiodes, on the other hand, are not susceptible to this effect. Other than for the construction of PIN diodes, an intrinsic semiconductor is not require to create a PN junction. In fact, lower dark currents can be achieved by doping the diode *higher*, not lower.

Thermal Detectors

Thermal detectors are an indirect measurement of radiation. They absorb light and convert that energy into heat. The resulting temperature rise is measured using a thermoresistive element (bolometer) or a pyroelectric element (thermopile). Thermoresistive elements produce a change in resistance, and pyroelectric elements will produce a voltage in response to a temperature change. Unlike photon detectors (photoconductors and photodiodes), thermal detectors are intrinsically slower. Their response time will be determined by the thermal time constant of the system rather than the minority carrier lifetimes.

Bolometers

The sensitivity of bolometers is defined by the change in resistance for a given incident radiation intensity. It is determined by several main factors: (1) thermal resistance between the sensing element and the surrounding, (2) optical absorption of the sensing element, and (3) the temperature coefficient of resistance (TCR) of the sensing element. The response time of the bolometer is determined by the thermal conductivity and the thermal mass of the sensing element.

To maximize sensitivity, the sensing element must have a very large thermal resistance to the ambient environment. This will allow the sensing element to reach a higher temperature for a small incident radiation. For fast response, the thermal mass has to be kept small. For this reason, bolometers are often made as suspended thin films, as shown in Fig 26. The thin film configuration ensures that the thermal mass remains low. The suspension allows the film to have minimal thermal contact with the substrate. The thermal conductivity can be reduced even further if the air surrounding the film is removed to create a vacuum. However, the thin film configuration also makes it difficult to absorb radiation within one pass. Therefore, a reflector is often used to create a second pass. Additionally, the air (or vacuum) gap between suspended film and the substrate can be used to optimize a cavity resonance. Since the film has to meet multiple requirements, it is typically created as a stack consisting of several films of different materials. One film is optimized as the sensor element with a large temperature coefficient of resistance. Another film is designed to provide maximum optical absorption, and another film is designed for best structural support.

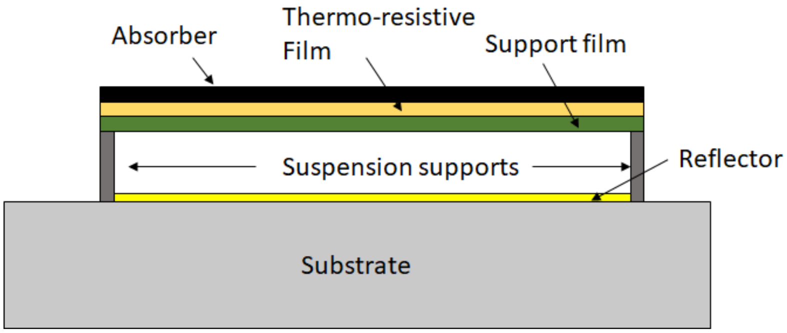

Figure 26: Typical bolometer design

Metal films such as platinum, nickel and titanium have a positive temperature coefficient (PTC). In other words, their resistance increases with increasing temperature. Other materials such as amorphous silicon (a-Si) and vanadium oxide (VO$_x$) are also used, which have a negative temperature coefficient (NTC) of resistance. The support film is usually silicon nitride (Si$_3$N$_4$). This is a mechanically strong film that can also be easily fabricated. The absorber can be a combination of metals or semiconductors. The bottom reflector is usually a metal such as gold.

The bolometer structures are fabricated using MEMS (Micro-Electro-Mechanical Systems) processes. This also allows the bolometers to be made in very small sizes and duplicated many times in an array format to provide imaging capability. An example is shown in Fig 27.

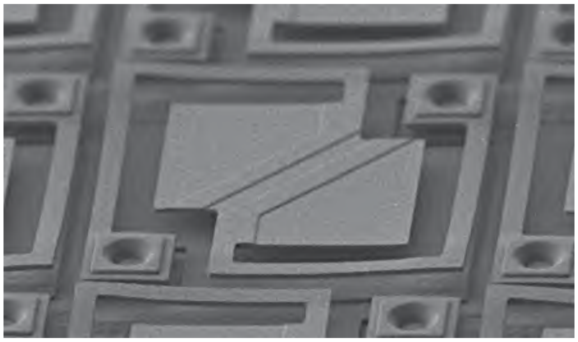

Figure 27: One pixel of a MEMS micro-bolometer array. Source: Hamamatsu

The response of the bolometer can be written as

$$C\frac{dT}{dt} = P_a + P_r - \frac{1}{R_t}(T - T_b) \tag{114}$$

where C is the thermal mass, P_a is the radiative power absorbed by the sensor, R_t is the thermal resistance, P_r is the joule heating from the electrical bias, and T_b is the temperature of the substrate. This is basically an energy balance equation between the power absorbed by the sensor and the power lost to the ambient. Under steady state condition, the solution becomes

$$0 = P_a + P_r - \frac{1}{R_t}(T - T_b) \tag{115}$$

$$T = T_b + (P_a + P_r)\,R_t. \tag{116}$$

The responsivity can be defined as the change in temperature for a change in incident power, resulting in

$$\mathcal{R} = \frac{\Delta T}{\Delta P_a} = R_t. \tag{117}$$

In other words, the sensitivity of the bolometer is equal to its thermal resistance. A bolometer that is highly thermally isolated from its ambient will produce a greater response compared to one that is directly attached to a thermally conductive heat sink. The units of responsivity is in K/W, which is different than in photodiodes and photoconductors. The high thermal resistance is normally accomplished by the suspended structure.

We can also derive an expression for the dynamic response of the bolometer. Assuming the system maintains a constant bias power P_r, we can determine the small signal response of this bolometer. Using a similar technique as we did with other detectors, we can get an expression for the small signal response of the bolometer. By setting the small signal quantities

$$\begin{aligned} T &= T_o + \delta T e^{j\omega t} && (118) \\ P_a &= P_{ao} + \delta P_a e^{j\omega t}, && (119) \end{aligned}$$

we can substitute these into equation (114) to get

$$j\omega C \delta T e^{j\omega t} = \cancel{P_{ao}} + \delta P_a e^{j\omega t} - \cancel{P_r} - \cancel{\frac{1}{R_t}\left(T_o - T_b\right)} + \frac{1}{R_t}\delta T e^{j\omega t}. \qquad (120)$$

The non-varying terms cancel out as indicated by the strikeouts, resulting in

$$\begin{aligned} j\omega C \delta T e^{j\omega t} &= \delta P_a e^{j\omega t} + \frac{1}{R_t}\delta T e^{j\omega t} && (121) \\ \frac{\delta T}{\delta P_a} &= \frac{R_t}{1 + j\omega R_t C} && (122) \\ \left|\frac{\delta T}{\delta P_a}\right| &= \frac{R_t}{\sqrt{1 + (\omega\tau)^2}} && (123) \end{aligned}$$

where the time constant is

$$\tau = R_t C. \qquad (124)$$

Following the same process as before, the 3dB frequency becomes

$$f_{3dB} = \frac{\sqrt{3}}{2\pi\tau}. \qquad (125)$$

The thermal time constant τ is the product of the thermal mass and thermal resistance, which is an identical formula to the RC time constant encountered in electrical circuits consisting of a capacitor and a resistor. A faster time response would require a small thermal mass and a small thermal resistance. The small thermal mass is achieved by having a thin film structure as the sensing element. However, the small thermal resistance is opposite of what we needed for a large responsivity. This conflicting requirement is a design trade-off. The responsivity and the speed has to be balanced depending on the intended application.

The electrical resistance of the sensor is used as a direct measure of the temperature. Assuming a linear model, the resistance of the sensor can be expressed as

$$R = R_o\left(1 + \alpha\left(T - T_o\right)\right) \qquad (126)$$

where α is the temperature coefficient of resistance (TCR), and R_o is the resistance when the temperature is T_o. The value of α can be positive or negative depending on the type of material (negative for metals, positive for semiconductors).

Measuring the resistance of the sensor is not trivial because the measurement itself will heat up the bolometer. This is represented by P_r in equation (114). Furthermore, the thermal resistance R_t and the thermal coefficient of resistance α will also be a functions of temperature, which will create nonlinearities in the measurements. Therefore, instead of measuring the resistance, a better technique is maintain a constant temperature in the bolometer (which corresponds to a constant bolometer temperature) by adjusting the electrical power P_r. The value of P_r required to maintain a

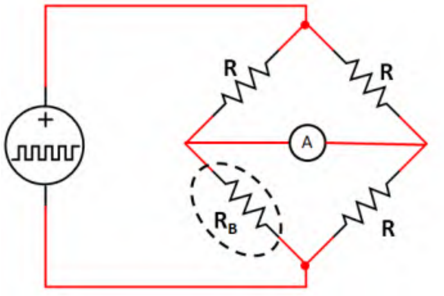

Figure 28: Bridge circuit configuration for sensing the bolometer temperature

constant temperature will then become the measure of the incident radiation being measured. This can be done using a bridge circuit configuration as shown in Fig 28. Three temperature-insensitive fixed resistors are used with the fourth resistor in the bridge as the bolometer. The branches are balanced by adjusting the pulse width of the source until zero current is observed between the two branches. The pulse width can then be used as a measure of the temperature of the sensor.

Radiant Power in Thermal Imaging

The radiant power incident on a single pixel of a thermal camera can be calculated if we know the f/# of the camera. This radiant power is independent of the distance of the object. In other words, a distant object and a close-up object will produce the same power on the pixel. This feature enables us to calibrate the intensity in terms of the temperature of the object, assuming we know the emissivity of the source. This can be derived as follows:

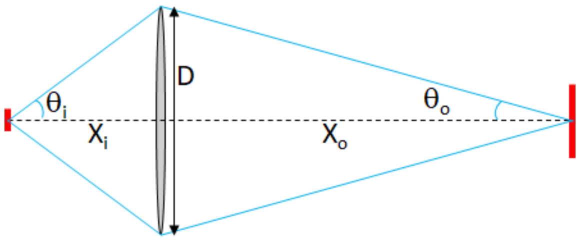

Figure 29: A simple imaging system using a thin lens

Referring to Fig 29, and using the thin lens approximation, we have

$$\frac{1}{X_o} + \frac{1}{X_i} = \frac{1}{f} \tag{127}$$

where X_i is the image distance, X_o is the object distance, and f is the focal length of the lens. If D is the diameter of the aperture, this can also be written as

$$\frac{D}{2X_o} + \frac{D}{2X_i} = \frac{D}{2f} \tag{128}$$

$$\tan\theta_o + \tan\theta_i = \tan\theta_a \tag{129}$$

where θ_a is

$$\tan\theta_a = \frac{D}{2f}. \tag{130}$$

If $f \gg D$, then θ_a will be small, resulting in $\tan\theta_a \approx \sin\theta_a \approx \theta_a$. Therefore,

$$\frac{D}{2f} = \sin\theta_a = NA \tag{131}$$

which is the numerical aperture of the lens. This approximation holds only for small values of numerical aperture.

The magnification can be defined as

$$M = \frac{\tan\theta_i}{\tan\theta_o} = \frac{\tan\theta_a - \tan\theta_o}{\tan\theta_o}. \tag{132}$$

Using the paraxial approximation, this becomes

$$M = \frac{\theta_a - \theta_o}{\theta_o}. \tag{133}$$

In realistic implementations, $X_o \gg X_i$. This results in $\theta_o \ll \theta_o$. Therefore,

$$M = \frac{\theta_a}{\theta_o}. \tag{134}$$

This can also be expressed in terms of the surface area of the object and image, as

$$\frac{A_o}{A_i} = M^2 \tag{135}$$

where A_o is the area of the object and A_i is the area of the image.

Thermal emission from a point source is given by Plank's law,

$$\phi_e^p(\lambda) = \epsilon \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}, \tag{136}$$

where ϵ is the emissivity of the object and $\phi_e(\lambda)$ is in the units of power per unit solid angle per unit surface area per unit wavelength. As derived earlier, the emission from a two-dimensional surface is

$$\phi_e^s(\lambda, \theta) = \phi_e^p(\lambda) \cos\theta. \tag{137}$$

The units of $\phi_e^s(\lambda, \theta)$ is also in power per unit solid angle per unit surface area) Therefore, total power emitted by a surface per unit wavelength whose area is A_o will be

$$\begin{aligned} A_o \Phi_e(\lambda) &= A_o \int_0^{2\pi} \int_0^{\theta_o} \phi_e^p(\lambda)\ \sin\theta \cos\theta \, d\phi \, d\theta && (138) \\ &= 2\pi A_o \int_0^{\theta_o} \phi_e^p(\lambda)\ \sin\theta \cos\theta \, d\theta && (139) \\ &= \pi \phi_e^p(\lambda)\ A_o \sin^2\theta_o. && (140) \end{aligned}$$

Substituting equations (134) and (135) results in

$$\begin{aligned} \Phi_e(\lambda) &= \pi \phi_e^p(\lambda)\ A_i M^2 \sin^2\theta_o && (141) \\ &= \pi \phi_e^p(\lambda)\ A_i \left(\frac{\theta_a}{\theta_o}\right)^2 \sin^2\theta_o. && (142) \end{aligned}$$

Furthermore, using the paraxial approximation, we can set $\sin\theta_o \approx \theta_o$. Along with the definition of numerical aperture, we can get

$$\begin{aligned} \Phi_e(\lambda) &= \pi \phi_e^p(\lambda)\ A_i \sin^2\theta_a && (143) \\ &= \pi \phi_e^p(\lambda)\ A_i (NA)^2. && (144) \end{aligned}$$

Finally, the power received on the image area A_i becomes

$$P_e = \int_{\lambda_1}^{\lambda_2} \pi \phi_e^p(\lambda)\ A_i (NA)^2 \ d\lambda \tag{145}$$

We can see that equation (145) does not contain object or image distances. It depends entirely on the size of the image and the numerical aperture of the lens.

Example

We will assume that the IR camera has a germanium lens with an f/# of 16. The transparency range of germanium is from 1.7μm to 10μm, which effectively blocks nearly all solar illumination from reaching the camera. Given a pixel size of 5μm$\times 5\mu$m, we can calculate the power reaching the pixel when the object temperature is $50°$C.

The numerical aperture of the lens is

$$NA = \sin\theta_a = \frac{D}{2f} = \frac{1}{2}\times\frac{1}{16} = 0.03125\ . \tag{146}$$

The power can be calculated as

$$P_e = \int_{1.7\mu\text{m}}^{10\mu\text{m}} \pi\frac{2hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda kT}-1}\ (5\mu\text{m}\times 5\mu\text{m})\times 0.03125^2\ d\lambda = 4.8\text{pW}. \tag{147}$$

Next, if the thermal resistance R_t of the microbolometer is 1×10^{10} K/W, the $50°$C object will raise the pixel temperature by 48mK.

Homework 8

1. Consider an avalanche photodiode (APD) which has an ionization ratio of $k = 1$.
    - Show that the expression for gain will converge to $\frac{1}{1-\alpha_n L_m}$.
    - Explain how this can lead to an unstable gain.
    - Explain how this aspect can be utilized as a Geiger mode APD.
2. A microbolometer thermal camera has a germanium lens (AR coated), and it is being used to measure the temperature of objects.
    - Explain why ordinary lenses made out of silica glass cannot be used in this camera.

      > Silica glass is not transparent at infrared wavelengths beyond 3um.

    - Assuming the camera has an aperture stop of f/5.6, and the pixel size on the sensor is 50μm x 50μm, calculate the radiative power received on one pixel from 35°C, 100°C and 300 °C ideal blackbody sources.

      Assuming a range of 1.7um to 10um:

      Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val f = 5.6
    val angle = atan(1.0/f/2.0)
    val Area = (50.0e-6).pow(2)
    val h = 6.62607015e-34
    val c = 3.0e8
    val k = 1.38064852e-23
    val dlambda = 0.01

    for (Tc in listOf<Double>(35.0, 100.0, 300.0)){
        var Es = 0.0
        val Tk = Tc+273.0
        var lambda = 1.7
        while (lambda < 10.0){
            val phi = 2.0*h*c.pow(2)/((lambda*1.0e-6).pow(5)) /
                    (exp(h*c/(lambda*1.0e-6*k*Tk))-1.0)*dlambda*1.0e-6
            Es += phi*PI*sin(angle).pow(2)*Area
            lambda += dlambda
        }
        println("T = ${Tc}C;  Pe = ${"%.2e".format(Es)} W")
    }
}

>>T = 35.0C;  Pe = 2.93e-09 W
>>T = 100.0C;  Pe = 9.31e-09 W
>>T = 300.0C;  Pe = 8.61e-08 W
```

Image Sensors

Image sensors consist of a two-dimensional array of photodetectors to measure the distribution of light intensity. The unique challenge in image sensors, as compared to discrete photodetectors, is due to the complexity involved in integrating a very large number of photodetectors onto a single semiconductor chip and collecting all the signals and converting them into a usable form. Some of the performance metrics of image sensors are pixel count, pixel size, linearity, dynamic range and noise.

In the following section, a brief description of the major image sensor technologies is discussed. In all of these approaches, the electrical signaling, photodetector biasing and multiplexing is done using silicon electronics. This makes silicon the preferred material for making photodetectors. That way, all of the components can be made on the same chip at the same time. However, silicon has limitations, especially with respect to its detection range. Silicon cannot detect photons with energy smaller than its bandgap - which corresponds to 1.1μm in wavelength. Therefore, infrared imaging at 1.5μm or 3μm requires a fundamentally different approach.

CMOS Image Sensors

CMOS stands for Complementary Metal Oxide Semiconductor. This is a manufacturing process used for producing digital electronic chips. Photodetectors are analog components, and it is not easy to integrate high quality photodetector into a CMOS platform. However, over the past decade, new CMOS technologies have emerged that include good quality photodetectors, resulting in the widespread availability of CMOS image sensor. The integrated circuit is fabricated just like any other digital electronic circuit. It contains a two-dimensional array of PN junction photodiodes, along with an amplifier and a biasing circuit for each photodetector. Each pixel also contains a color filter (red, green or blue) and a microlens to increase the fill-factor of each pixel.

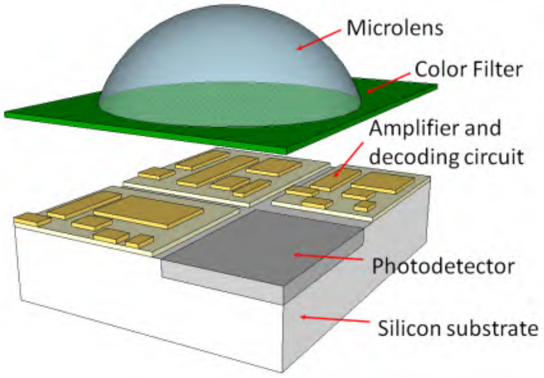

Figure 30: CMOS active pixel sensor.

In the early days of electronic image sensors, CCD's (charged coupled devices) were the preferred technology. However, CMOS image sensors have come a long way in the past 10 years, and have surpassed the performance of CCDs. These can be found in nearly all consumer electronic cameras. Pixel sizes are on the order of 1μm, and pixel counts of 8 million is not uncommon.

CCD Image Sensors

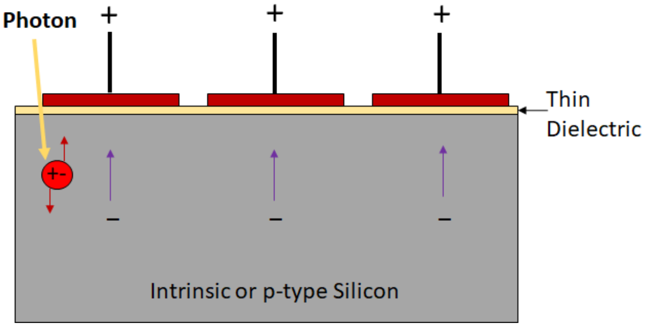

Figure 31: Charge Coupled Device (CCD)

Charge coupled devices are metal-oxide semiconductor (MOS) capacitors. They are made on a silicon substrate with a thin dielectric spacer (usually silicon dioxide), and a transparent polysilicon electrode. Assuming a lightly p-doped substrate, a positive voltage is applied to all the electrodes. This will repel all the holes under the electrodes and attract the minority carrier electrons. However, since the density of electrons in a p-type substrate is very small, they would need to drift from deep within the substrate to reach the surface. This process will take some time, especially if the field is small. Before the capacitor reaches its equilibrium charge, any photons absorbed in the vicinity of the electrode will produce a more immediate source of electrons, which will quickly increase the negative charges under the electrode. Therefore, the process can be described as a slow rate of charging due to minority carrier drift and thermal generation of electrons in the absence of photons, and a faster rate of charging depending on the photon flux.

One of the distinguishing features of CCD image sensors compared to CMOS is the lack of any pixel-level amplifiers or decoding circuitry. Charges accumulated in the capacitor are read out by shifting the charges from one capacitor to the adjacent capacitor. They are read one at a time by a conversion circuit located at the end of each row. This type of charge movement makes it possible for the entire sensing area to consist of just capacitors with no other electronic components. This gives CCDs a very high fill factor and very high sensitivity. Although CMOS sensors have virtually overtaken CCDs in consumer applications, CCD sensors are still used in some applications that require high sensitivity, such as astronomy and spectroscopy.

Compared to CMOS sensors, CCDs have greater sensitivity and greater fill factors because almost the entire area of the pixel is occupied by the optical sensor. Therefore, no microlenses are used on CCDs.

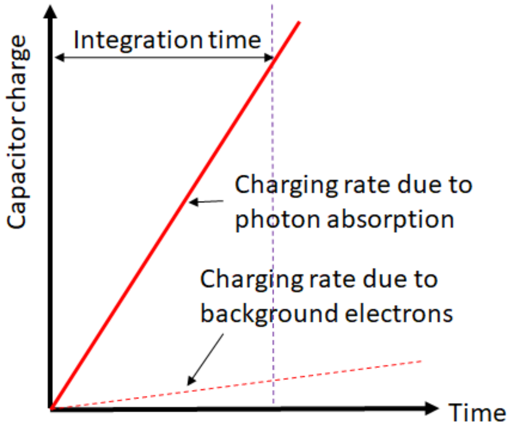

Figure 32: Charging rate of a CCD due to background electrons and photon flux

Infrared Image Sensors

Silicon can only detect up to 1.1μm in wavelength. While this is suitable for visible cameras, there are many applications that require image sensing in the infrared spectrum. $1.3-2.0\mu$m wavelength range (known as short wave infrared - SWIR) is important in laser imaging. A phenomenon known as night glow produces an emission in the upper atmosphere that has a peak near 1.7μm, making this wavelength range suitable for night vision. The atmosphere has a transparent window in the $3-5\mu$m spectral range (known as mid-infrared - MWIR), as well as in the $8-12\mu$m range (long-wave infared - LWIR). These are important bands for thermal imaging, where black body radiation from objects near room temperature can be detected.

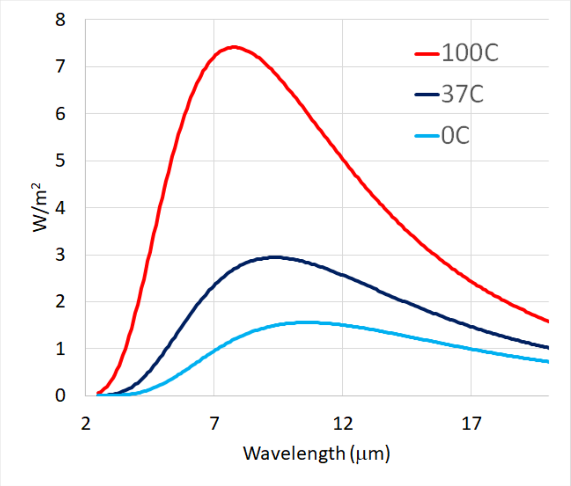

Figure 33: Black body emission from a $0°$C, $37°$C and $100°$C sources.

For example, Fig 33 shows the black body emission spectra from $0°$C, $37°$C (human body temperature), $100°$C sources. The human body temperature peaks at around a wavelength of 9μm, and the $100°$C objects peaks at 7.5μm. Clearly, image sensors operating in this regime will be able to utilize these emissions to produce images that do not rely on reflected sunlight or artificial light.

Although there are many semiconductors that can detect photons in these wavelength ranges, the fact that the photodetectors have to interface with silicon for biasing the detectors and collecting the signals is a major technological bottleneck. This requirement has lead to the hybridized detector configuration where silicon is used for electronics and another material is used as the photodetector.

Fig 34 shows a configuration commonly used in image sensors using photodetector materials other than silicon. The photodetector array is constructed using a backside-illuminated thinned-substrate geometry. Example materials include In$_{0.53}$Ga$_{0.47}$As for the $1.3-1.7\mu$m spectral range, InSb $3-5\mu$m range, and MCT for $3-12\mu$m range. The electric circuitry, including amplifiers and biasing circuits for each photodetector pixel is made on a silicon CMOS platform using an identical array configuration. The two chips are then hybridized together using indium solder to make the electrical connections. This process is known as flip-chip bonding. This configuration has allowed non-silicon photodetectors to exist on a silicon platform. As discussed earlier, for wavelengths longer than 3μm, the semiconductors have to be cooled to very low

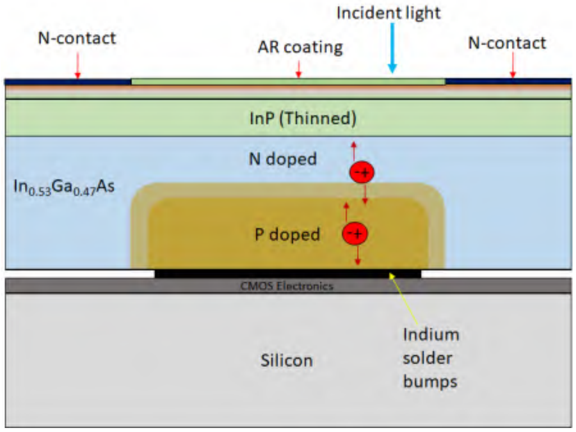

Figure 34: Backside illuminated In$_{0.53}$Ga$_{0.47}$As photodiode hybridized to a silicon readout integrated circuit (ROIC)

temperatures. This introduces numerous challenges, such as matching the thermal expansion coefficients between silicon and the detector material. Nevertheless, this is currently the most widely used configuration for infrared imaging.

An alternate, lower-cost option for imaging infrared has also emerged using microbolometers as the sensors. Because these do not rely on a semiconductor bandgap, they can be operated at room temperature. However, they are much slower than photodetectors because their operating speeds are limited by the thermal time constant $R_t C$ outlined in equation (124). They are also significantly more noisier. Known as ”uncooled thermal imagers”, these sensors cost significantly less than cooled photon detectors, and are becoming widely available in consumer applications.

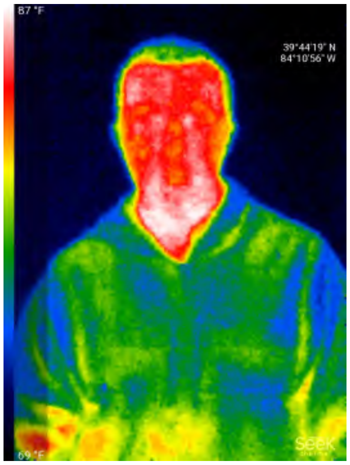

Figure 35: Image from an uncooled microbolometer thermal camera

Materials used in Image Sensors

While the purpose of the detector material is to absorb the photons and convert them to an electrical current, the optical components used for focusing the image have to be transparent to the incident radiation. For example, in the visible spectrum, silicon is commonly used as the photodetector material, and silica glass is used for making the optical components. The bandgap of silicon is 1.1eV, so it absorbs wavelengths shorter than 1.1μm. Although silica glass is not a semiconductor, it is transparent for wavelengths longer than about 350nm. For wavelengths shorter than 350nm, fused silica or quartz can be used. However, as we saw in our earlier discussions, the collection efficiency of silicon photodetectors will decline at very short and very long wavelengths. Therefore, the useful range of most silicon photodetectors with a silica glass optics is about 400nm to 1000nm.

Most commercial visible cameras restrict the spectral range of silicon photodetectors to the visible spectrum (400nm to 700nm). This avoids artifacts from radiation that may be present in the near infra-red (700nm to 1100nm). This can be accomplished with a short-pass filter with a cutoff wavelength at 700nm.

Infrared cameras use a variety of other materials. Short-wave IR (SWIR) cameras use $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ as the detector material, which has a bandgap of 0.74eV. Therefore, it is suitable for detecting wavelengths in the range of 1.0μm $- 1.7\mu$m. The reason for this specific stoichiometry has to do with lattice matching. $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ has the same crystal structure and lattice constant as InP, which is a common substrate. This allows $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ to be produced more easily than other stoichiometries of $\text{In}_{1-x}\text{Ga}_x\text{As}$.

Mid-wave IR (MWIR) cameras use InSb or $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$. The bandgap of InSb at 77K is 0.23eV, which corresponds to a wavelength of 5.4μm. Therefore, it can be used to detect 3μm $- 5\mu$m, which is the MWIR band. InSb is also produced as substrates, so lattice-matching to another material is not necessary.

$\text{Hg}_{1-x}\text{Cd}_x\text{Te}$ is more difficult because it is not lattice matched to any common substrates. In most implementations, silicon is used as the substrate with CdTe as an intermediate layer to buffer the $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$. Because none of the layers are lattice-matched, the quality of the resulting $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$ will be lower compared to other materials, resulting in high dark currents and lower responsivities. The advantage of $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$, however, is that it can span a large range of bandgaps, from -0.3eV (which is actually a metal) to 1.6eV, depending on the value of x. The bandgap of Hg$_{0.7}$Cd$_{0.3}$Te is 0.24eV at 77K, and is suited for MWIR.

Hg$_{0.79}$Cd$_{0.21}$Te has a bandgap of 0.1eV at 77K. This can be used in the 8μm-12μm band, known as the long-wave IR (LWIR) band. This band also contains the peak emission wavelength of moderately warm objects in the 0C to 100C range, making it an important band for thermal imaging. But the high cost of Hg$_{1-x}$Cd$_x$Te and the requirement for cryogenic cooling has been a barrier for widespread commercial applications. As a result, microbolometers have emerged as the competition to Hg$_{0.7}$Cd$_{0.3}$Te in the LWIR. These are uncooled detectors, and rely on direct temperature sending rather than bandgap detection. The most common material used for temperature sensing is VO$_x$, which is an oxide of vanadium that has a high temperature coefficient of resistance (TCR). They are manufactured as suspended membranes as was already discussed in Fig 27.

Solar Cells

Photovoltaic Mode

The photodiodes that we examined so far were for the purpose of detecting signals, not for generating power. A photodiode can also produce power under illumination. This is known as the photovoltaic mode of operation. Whether a photodiode consumes power or produces power depends on its operating point in the I-V curve.

As stated earlier, the I-V curve of a photodiode under illumination is

$$I = \left(I_s + \frac{|V_a|}{R_{\text{shunt}}}\right)\left(e^{V_a/V_t} - 1\right) - I_{ph}, \tag{148}$$

where I_{ph} is the photocurrent, and R_{shunt} is the shunt resistance to account for additional recombination processes as discussed earlier. The I-V curve is shown in Fig 36. There are four quadrants on this plot. When the product $I \times V$ is positive, it represents power being consumed by the photodiode. This happens in the lower left and upper right quadrants. $I \times V$ would be negative in the upper left and lower right quadrants, but only the lower right quadrant contains a portion of the I-V curve. This is the photovoltaic operational region of the photodiode. The magnitude of $I \times V$ in this quadrant will depend on the exact operating point of the photodiode. This will depend on the illumination intensity and the load attached to the photodiode. The desired operating point is where the $I \times V$ is a maximum (with a negative sign), and corresponds to the largest rectangle contained under the I-V curve as indicated in Fig 36. This operating point can be calculated if all the parameters of the diode are known.

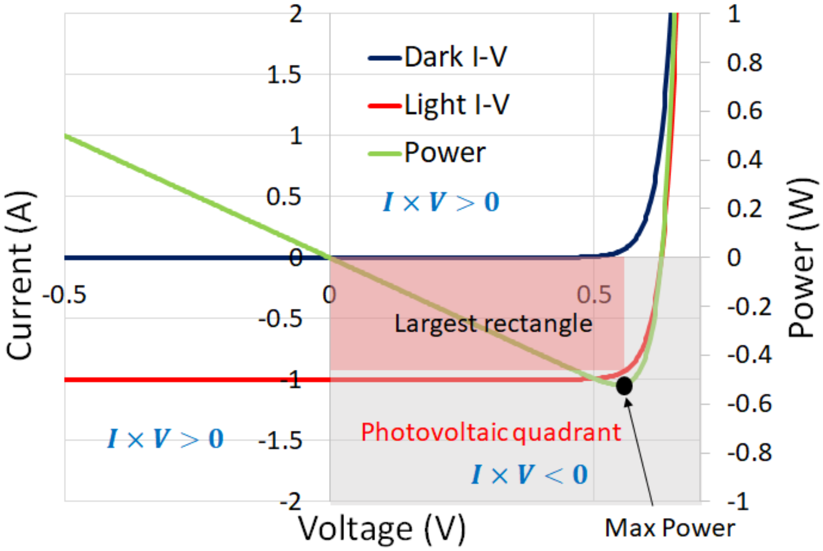

Figure 36: Photovoltaic quadrant of a photodiode, and the point of maximum power generation

Using equation (148), we can write the expression for power as

$$P = IV = V_a\left[\left(I_s + \frac{|V_a|}{R_{\text{shunt}}}\right)\left(e^{V_a/V_t} - 1\right) - I_{ph}\right]. \tag{149}$$

Fig 36 also shows the calculated value of $I \times V$, where we can see it reaches a peak negative value.

Solar Irradiance

The efficiency of solar cells is defined as the electrical power produced divided by the electromagnetic power incident on the surface of the photodetector. The incident power is a function of the solar illumination and angle of incidence as well as atmospheric absorption. Therefore, it is a highly variable quantity. However, the solar illumination reaching the earth's atmosphere is a very predictable quantity based on the blackbody radiation theory. This radiation intensity is designated as AM0 radiation (zero atmospheric absorption), and has a value of approximately 1355 W/m^2. AM1 is designated as one atmosphere worth of absorption, which occurs at normal incidence on the equator. AM2 would be two atmospheres worth of absorption, which would occur at an inclination of 60-deg. AM1.5 is used as the standard average illumination in the northern hemisphere, which is about 1000 W/m^2.

AM0 intensity can be calculated by treating the sun as a black body source with a surface temperature of 5778 K. The emission spectrum from a point source is characterized by Plank's law

$$\phi_e^p(\lambda) = \epsilon \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}. \tag{150}$$

given in power per unit area per unit solid angle per unit wavelength. As discussed previously, spectral radiance of a surface will have an additional $\cos\theta$ term due to Lambert's projection:

$$\phi_e^s(\lambda, \theta) = \cos\theta \, \phi_e^p(\lambda). \tag{151}$$

To find the emission from a unit surface area of the sun, we have to integrate over the hemisphere angles.

$$\begin{aligned} \phi_e^s(\lambda) &= \int_0^{2\pi} \int_0^{\pi/2} \phi_e^s(\lambda, \theta) \sin\theta \, d\theta \, d\phi & (152) \\ &= \int_0^{2\pi} \int_0^{\pi/2} \phi_e^p(\lambda) \cos\theta \sin\theta \, d\theta \, d\phi & (153) \\ &= 2\pi \phi_e^p(\lambda) \left. \frac{\sin^2\theta}{2} \right|_0^{\pi/2} & (154) \\ &= \pi \phi_e^p(\lambda). & (155) \end{aligned}$$

The total power (over all wavelengths) per unit area emitted by the surface becomes:

$$\Phi_e = \int_0^\infty \phi_e^s(\lambda) \, d\lambda. \tag{156}$$

Setting $T = 5778$K, we can calculate the power emitted by a unit surface area of the sun:

$$\Phi_e = \int_0^\infty \pi\epsilon \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1} \, d\lambda = 63.1 \text{ MW/m}^2 \tag{157}$$

One square meter of the sun projects over a much larger distance by the time the light reaches earth. We can calculate this based on the area of the sun and the distance from the sun. The surface area of the sun is $4\pi R_s^2$ where R_s is the radius of the sun (which is 432,170 miles). The receiving surface area at earth's average orbital distance is $4\pi D^2$ where D is the distance between the earth and sun (92.96 million miles). Therefore, we can find the intensity reaching the earth's atmosphere by evaluating:

$$\Phi_r = \Phi_e \frac{R_s^2}{D^2}, \tag{158}$$

where Φ_r represents the power received (hence the subscript r).

$$\Phi_r = 1355 \text{ W/m}^2. \tag{159}$$

This is the solar energy value used for calculating illumination on solar cells on near-earth orbiting satellites. On the surface of the earth, the power will be much lower due to atmospheric absorption. As stated before, average values for AM1.5 is around 1000 W/m^2, but this can vary greatly depending on atmospheric conditions.

Solar Cell Efficiency

We noted earlier that the spectral responsivity of a photodiode depends on the thickness of the I-layer and the junction depth:

$$\mathcal{R} = \eta_e \ \frac{q}{h\nu}. \tag{160}$$

For the purpose of this calculation, we will assume a perfect photodiode with an infinitely thick I-layer and a nearly zero junction depth, and a 100% internal quantum efficiency. This allows both the short and long wavelength regimes to be absorbed, and results in a collection efficiency of 100% for all wavelengths up to the bandgap wavelength to give a responsivity value of

$$\begin{aligned} \mathcal{R} &= \frac{q}{h\nu} \ \text{ for } h\nu > E_g \qquad (161)\\ &= 0 \ \text{ for } h\nu < E_g. \qquad (162)\end{aligned}$$

Next, we multiply this spectral responsivity by the solar irradiance spectrum to get the resulting current per unit area of the photodetector:

$$\begin{aligned} J &= \frac{R_s^2}{D^2}\int_0^{\lambda_g} \phi_e^s(\lambda)\,\mathcal{R}d\lambda \qquad (163)\\ &= \frac{R_s^2}{D^2}\int_0^{\lambda_g} \phi_e^s(\lambda)\left(\frac{q\lambda}{hc}\right)d\lambda \qquad (164)\\ &= \pi\frac{R_s^2}{D^2}\int_0^{\lambda_g}\left(\frac{2qc}{\lambda^4}\frac{1}{e^{hc/\lambda kT}-1}\right)d\lambda \qquad (165)\end{aligned}$$

For silicon, using $E_g = 1.1$eV results in a bandgap wavelength of $\lambda_g = 1.127\mu$m. Substituting this into equation (165) results in a photocurrent of 518 A/m^2. In other words, AM0 illumination on a one square meter of a perfect silicon solar cell will produce a current of 518 Amps.

This current can be used in the diode equation (148) to find the bias point that produces maximum power. In order to perform this calculation, we need I_s. We derived the expression for I_s earlier as

$$I_s = qA\left(\frac{D_n}{L_n}\frac{n_i^2}{N_A} + \frac{D_p}{L_p}\frac{n_i^2}{N_D}\right). \tag{166}$$

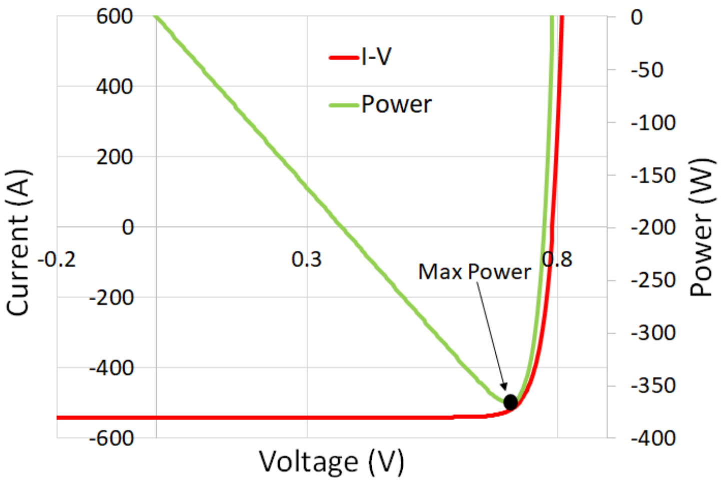

Figure 37: I-V and power curve of a solar cell with the maximum current of 518 Amps at AM0 illumination

This, along with R_{shunt}, can be used to plot the power vs voltage curve, and the find the maximum power value.

Assuming silicon, we can use representative values for the material parameters and n- and p-type doping. Using $N_A = 10^{18}$ cm^{-3}, $N_D = 10^{18}$ cm^{-3}, we can get $L_n = 20\mu$m, $L_p = 10\mu$m, along with $\tau_n = 8\mu$s, $\tau_p = 2\mu$s, $D_n = 0.5$ cm^2/s, $D_p = 0.5$ cm^2/s, $n_i = 1.5 \times 10^{10}$ cm^{-3}. Substituting these into equation (166) results in $I_s = 0.2$nA for a photodiode with an area of 1m^2. For now, lets assume $R_{\text{shunt}} \to \infty$. We can plot the I-V curve of the photodiode and find the point of maximum power. This is shown in Fig 37. The maximum power value works out to 324 W. Since the incident AM0 power is 1355W, the solar cell efficiency is $324/1355 = 24$%.

The diode saturation current I_s and R_{shunt} have a significant effect on the maximum power value that can be obtained from a solar cell. We can expand equation (149) and group the terms separately to allow for a better interpretation:

$$P = \underbrace{V_a I_s \left(e^{V_a/V_t} - 1\right)}_{\text{Ideal diode dissipation}} + \underbrace{\frac{V_a^2}{R_{\text{shunt}}} \left(e^{V_a/V_t} - 1\right)}_{\text{Nonideal recombination dissipation}} - \underbrace{V_a I_{ph}}_{\text{Generated power}} \tag{167}$$

The power generated by the photocurrent is the last term, $V_a I_{ph}$. Part of this power is expended on producing the diffusion currents in the diode. This is the first term, $V_a I_s \left(e^{V_a/V_t} - 1\right)$, and it is an inevitable component even in a perfect diode. However, we can minimize this power loss by making the value of I_s small. The second term is the power lost to the shunt resistor. This really consists of several non-ideality factors from Auger recombinations, Shockley-Read-Hall recombinations and leakage currents. The remaining power is what is dissipated in the external circuit as the useful power.

Increasing the efficiency of solar cells requires the following considerations:

- From equation (167), we can see that increasing the photocurrent I_{ph} will increase the generated power. This means, with all other factors remaining fixed, for a given illumination spectrum, using a smaller bandgap material will allow more of the longer wavelengths to be absorbed.

- Reducing I_s will reduce the internal power dissipation in the diode. From equation (166), I_s (per unit area) is determined by doping of the PN junction as well as the intrinsic carrier density n_i. The diffusion coefficients D_n and D_p as well as L_n and L_p also depend on the doping values. In general, the lifetimes τ_n and τ_p will decrease as doping values are increased, resulting in a both D_n/L_n and D_p/L_p becoming larger. Therefore, increasing doping to reduce I_s is not a straightforward effect. Additionally, as we saw earlier, the intrinsic carrier density n_i is inversely related to the bandgap of the material (see equation (168)). This is opposite of the statement we made earlier for increasing the photocurrent. Therefore, this effect partially negates the benefits of increasing the photocurrent by choosing a smaller bandgap material. Additionally, temperature also plays a role in n_i. Reducing temperature will decrease n_i, but active cooling of solar cells consumes energy, so it is not always a practical solution.
- A larger V_a will generate higher power. This is the voltage across the diode at the point of maximum power. It is a numerically computed value, but it depends strongly on I_s and V_t. A larger I_s will increase V_a, but as we stated previously, a larger I_s also increases the internal power consumption.
- Increasing the shunt resistance (or decreasing the leakage currents) will improve the efficiency of solar cells. These leakage currents get worse at high doping levels, which is opposite of the requirement for achieving a low I_s.

As we can see, there are several competing factors, and we cannot increase the efficiency of solar cells by simply changing one factor.

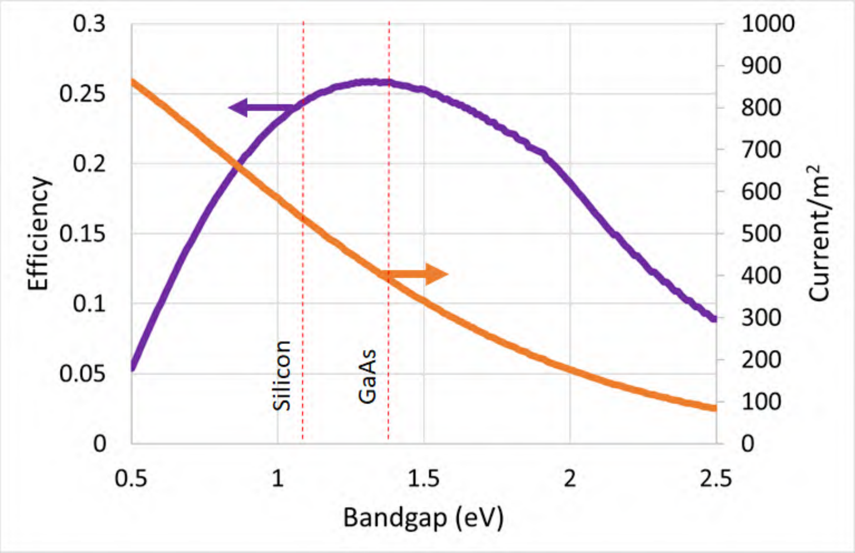

Figure 38: Solar cell efficiency vs bandgap of the material, assuming all other material parameters remain the same as silicon.

We can examine what would happen hypothetically if the bandgap of the material is changed. For this analysis, we will keep the same diffusion coefficients as in the above silicon solar cell example, but modify just the bandgap. Besides the cut-off wavelength of absorption, the largest change in the diode from a different bandgap will be in the intrinsic carrier concentration. Earlier we saw that the intrinsic carrier concentration was

$$n_i = \frac{4\sqrt{2}}{h^3} \left(\pi kT\right)^{3/2} \left(m_c m_v\right)^{3/4} e^{-E_g/2kT}. \tag{168}$$

At room temperature, this was about 1.5×10^{10}cm^{-3} for silicon. Though it is somewhat fictitious, we can evaluate n_i by allowing the silicon bandgap to change, and examine how the solar cell efficiency changes. This is shown in Fig 38. Also shown on the right hand axis is the generated photocurrent per square meter from solar radiation. We can see that the bandgap of silicon is fairly close to the optimum value required to harvest the maximum energy from solar illumination, making it not only highly efficiency, but also inexpensive.

Solar Cell Structures

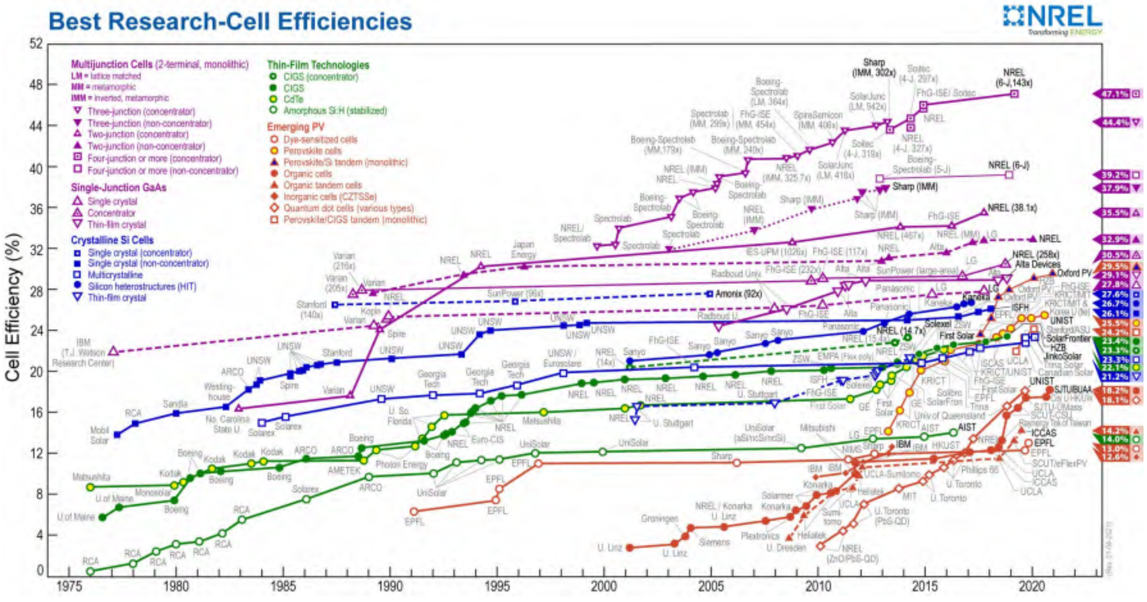

Figure 39: Best Research-Cell Efficiency Chart. Source: NREL

To capture the maximum possible energy, solar cells have to be made in large panel sizes. Instead of being a few microns or millimeters, they have to be on the order of meters. This makes the fabrication process vastly different than other types of photodetectors. Even inexpensive materials like silicon become unfavorable at such size scales. Therefore, alternative techniques are being explored. The use of solar concentrators is one such method. Concentrators utilize reflectors to focus light onto a smaller solar cell. This makes it possible to use a high efficiency solar cell without having to make them in large sizes. However, this will require more complicated hardware and installations.

Cost, durability and manufacturing cost are some of the largest driving factors in solar cells more than efficiency. Thin film solar cells (on glass or other substrates) are more economical compared bulk crystalline silicon. As a result, a significant portion of the solar market is comprised of thin film solar cells, using materials such as CdTe, amorphous silicon, and copper indium gallium selenide (CIGS). Even though their efficiencies are lower than silicon, they are less expensive to manufacture. Solar cells based on organic films are even less expensive than inorganic films, but their efficiencies are even lower. All of these areas are currently undergoing significant development.

For very specialized applications such as spacecraft and autonomous systems, it is possible to achieve efficiencies higher than the values depicted in Fig 38. This is done using a technique known as multi-junction diode configuration. By creating a photodiode with junctions consisting of multiple junctions each made from a different bandgap material, it is possible to achieve values as high as 45%. The multi-junctions are stacked in such a way that short wavelengths are absorbed by the first junction, and the longer wavelengths by the lower junctions. By separating the the solar spectrum into separate segments for each junction, it is possible to achieve a greater overall efficiency.

Anti-reflection coatings are an important aspect of solar cells. The native reflection from silicon is nearly 40%, so applying a coating will significantly increase the efficiency. However, all thin film coatings will exhibit angular sensitivity. In other words, it is not possible to design a thin film anti-reflection coating that will work for all incidence angles. In solar cells, this becomes an important limitation because the angle difference between sunrise and sunset can be 180-degrees. The solution commonly used is to rotate the solar panels to always provide a normal incidence, but this adds to the mechanical complexity of an installation. Alternative methods are being explored to provide angular-insensitive anti-reflection coatings, such as moth-eye structures.

Solar Cell Arrays and Load Balancing

As evident from our discussion so far, a solar cell will provide optimum power only when it is operated near its peak-power bias point. The voltage supplied by the solar cell at this peak power will be in the range of $0.6-0.7$V. The exact voltage will depend on the diode characteristics as well as the illumination conditions. Since most electrical loads require much higher voltages than $0.6-0.7$V, solar cells are typically connected in series to drive such loads.

Most solar panels are made up of cells arranged in a two-dimensional configuration. For example, a 1 square meter panel could consist of $10,000$ cells of 1cm $\times 1$cm in size. The cells are then connected in series and parallel depending on the specific load it is designed to drive. For example, connecting 100 cells in series will increase the operating voltage 100-fold. Connecting 100 of those units in parallel will increase the current 100-fold. Therefore, it is possible to achieve any combination of voltage and power as required to drive the load.

However, loads are never static, nor is the illumination on the solar panel. Therefore, it becomes necessary to make real-time adjustment to the load to compensate for these effects. This is done by the load balancer. It is basically a DC/DC converter that varies the effective impedance seen by the solar panel to always maintain operation at the peak power bias point.

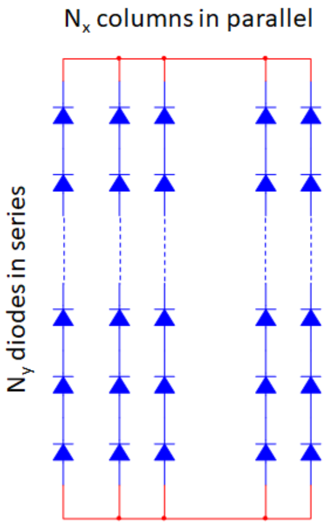

Figure 40: A two-dimensional grid of solar cells to create the voltage and current required to drive a load

Homework 9

1. Calculate the maximum photocurrent density that can be produced in a perfect GaAs solar cell (in A/m^2) on a satellite in orbit around Mars. If $I_s = 10^{-8}$A/m^2, and $R_{\text{shunt}} = \infty$, calculate the peak power. Then, calculate how much larger the GaAs panels have to be compared to the ones around earth's orbit to get the same power.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val h = 6.62607015e-34     //Plank's constant
    val c = 3.0e8          //Speed of light
    val k = 1.38064852e-23     //Boltzmann constant
    val q = 1.602e-19      //Electronic charge
    val T = 5778.0             //Temperature of the sun
    val Rs = 432170.0        //Radius of sun (miles)
    //val D = 92.96e6            //Distance to earth (miles)
    val D = 134.4e6          //Distance to mars (miles)
    //val Eg = 1.1               //Bandgap of silicon
    val Eg = 1.42            //Bandgap of GaAs
    val lambda1 = 0.1        //Integration start (um)
    val lambda2 = 20.0       //Integration end (um)
    val dlambda = 0.001      //(um)

    val lambda = DoubleArray(((lambda2-lambda1)/dlambda).toInt()){lambda1+it*dlambda
    }
    val P = lambda.map{lambda ->
            PI*(Rs/D).pow(2)*2.0*h*c.pow(2)/((lambda*1.0e-6).pow(5))/(exp(h*c/(
    lambda*1.0e-6*k*T))-1.0)*dlambda*1.0e-6
            }.toDoubleArray()
    val I = lambda.map{lambda ->
        if (lambda < 1.24/Eg){
            PI*(Rs/D).pow(2)*2.0*q*c/((lambda*1.0e-6).pow(4))/(exp(h*c/(lambda*1.0e
    -6*k*T))-1.0)*dlambda*1.0e-6
        }
        else{0.0}
        }.toDoubleArray()

    println(”Incident radiative power = ${”%.2f”.format(P.sum())} W”)
    println(”Photocurrent = ${”%.2f”.format(I.sum())} A”)
}

>>Incident radiative power = 652.51 W
>>Photocurrent = 188.44 A
```

   The incident radiation is 652/1355 = 48% that of earth. Therefore, we would need roughly double the size for each solar panel to produce the same power as on earth.

2. Lasers can be used for beaming power straight to drones. Consider a drone with a downward-facing silicon photovoltaic cells and a 10W 950nm ground-based laser system. Assume the entire laser beam is incident on the photovoltaic cell on the drone. Assuming an external quantum efficiency of 100%, calculate the photocurrent generated in the cells. If the photovoltaic cell has $I_s = 0.1$nA with $R_{\text{shunt}} = \infty$, calculate the energy conversion efficiency. Explain why it is possible to get higher conversion efficiency compared to solar illumination.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val wavelength = 0.950
    val Responsivity = wavelength/1.24
    val Pin = 10.0
    val Iph = Pin*Responsivity
    println(”Iph = ${”%.2f”.format(Iph)} A”)

    val Is = 0.1e-9
    val Vt = 0.026

    fun IV(Va:Double):Double{
        val I = Is*(exp(Va/Vt)-1.0) - Iph
        val P = I*Va
        return P
    }

    val Pmax = abs(DoubleArray(1000){it*0.0015}.map{IV(it)}.min())
    println(”Pmax = ${”%.2f”.format(Pmax)} W”)
    println(”Efficiency = ${”%.2f”.format(Pmax/Pin*100)}%”)
}

>>Iph = 7.66 A
>>Pmax = 4.18 W
>>Efficiency = 41.78%
```

Noise

Noise is the unavoidable random portion of any measurement. In photonic systems, noise arises due to the statistical nature of photons and electrons. Typically, a measurement is performed on a time-varying signal by integrating the signal for repeating intervals of time, known as the integration time. This becomes the smallest increment of time for that system. The number of electrons collected during each integration time, Q_n, represents the measurement. This is illustrated in Fig 1.

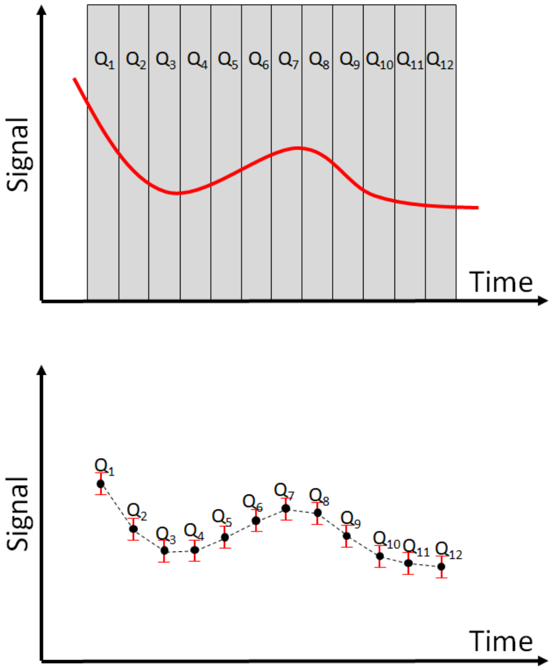

Figure 1: Measurement on a time-varying signal with a fixed integration time

Due to randomness in every system, the number of electrons collected within each integration window will have an expectation value (mean) with a standard deviation. The signal-to-noise ratio is defined as the square of the ratio between the expectation value and the standard deviation. That is:

$$\text{SNR} = \left(\frac{\bar{n}}{\sigma}\right)^2 = \frac{\bar{n}^2}{\sigma^2} \tag{1}$$

where $\bar{n}$ is the expectation value and σ is the standard deviation, defined by:

$$\bar{n} = \frac{1}{N}\sum_0^N Q_n \tag{2}$$

$$\sigma^2 = \frac{1}{N}\sum_0^N (Q_n - \bar{n})^2 . \tag{3}$$

Quantum Shot Noise

The arrival rate of electrons and photons are mathematically modelled as a Poisson's distribution. This makes the math easier because the standard deviation of a Poisson's distribution is equal to the square root of the mean. Therefore,

$$\sigma_s = \sqrt{\bar{n}_s}. \tag{4}$$

As a result, the SNR becomes:

$$\text{SNR} = \frac{\bar{n}_s^2}{\sigma_s^2} = \frac{\bar{n}_s^2}{\bar{n}_s} = \bar{n}_s. \tag{5}$$

σ_s is also known as the quantum shot noise because it arises from the discrete nature of the electrons and photons. In the absence of any additional noise, this becomes the fundamental limit of SNR.

We can also express $\bar{n}_s$ and σ_s in terms of incident power and integration time. If the incident power is P, and the integration time is T_i, and the external quantum efficiency of the detector is η_e, the average number of electrons collected during each integration time will be

$$\bar{n}_s = \frac{\eta_e P T_i}{h\nu}. \tag{6}$$

We can also express the integration time T_i as a sampling rate B, where

$$B = \frac{1}{T_i}. \tag{7}$$

Therefore, the mean and variance become

$$\bar{n}_s = \frac{\eta_e P}{Bh\nu} \tag{8}$$

$$\sigma_s = \sqrt{\frac{\eta_e P}{Bh\nu}} \tag{9}$$

resulting in

$$\text{SNR} = \frac{\bar{n}_s^2}{\sigma_s^2} = \frac{\eta_e P}{Bh\nu}. \tag{10}$$

From this expression we can see that the SNR can be improved by increasing the incident power and the external quantum efficiency, or by decreasing the sampling rate (which is the same as increasing the integration time). Photon energy also plays a role because for a given power there will be more photons if the wavelength of those photons is longer.

Shot noise is also known as fundamental noise, because it is the ultimate limit of what can be achieved. A signal quality simply cannot improve beyond the shot noise limit.

Thermal (Johnson) Noise

The random motion of electrons due to thermal energy is not captured by the shot noise model. Shot noise only contains the noise due to the quantized nature of electrons. It can be thought of as a discretization noise similar to the noise in a digital system.

Electrons are constantly in motion with an average thermal energy kT. Even without an applied bias, free electrons will always be in motion due to this energy. However, they will be aligned in random directions, so the average current along any specific direction will be zero. But the instantaneous current will not be zero. As a result, during an integration period, the number of electrons collected can be slightly higher, or slightly lower, than the signal electrons.

The mean and deviation of the thermal noise current can be derived (not shown here) as:

$$\bar{I}_t = 0 \tag{11}$$

$$\sigma_{I_t} = \sqrt{\frac{4kTB}{R}} \tag{12}$$

where R is the equivalent resistance of the circuit. The number of electrons collected per integration time becomes:

$$\bar{n}_t = 0 \tag{13}$$

$$\sigma_t = \sigma_{I_t}\frac{1}{Bq} = \sqrt{\frac{4kT}{q^2RB}} \tag{14}$$

The fact that $\bar{n}_t$ is zero should not be surprising. This is simply because no net current can be generated due to the random thermal motion in a resistor. The standard deviation is the spread around this mean value. Clearly, it gets larger at higher temperatures. It also gets larger in smaller resistors, as well as with smaller sampling rate. It may seem odd that σ_t gets better (i.e smaller) with increasing sampling rate. But this behavior is identical to the shot noise deviation we discussed in equation (9). When this deviation is considered in combination with the mean signal value (which also declines linearly with increasing B, instead of as the square root of B), the net result is a decline in the SNR with increasing B, which is consistent with expectations.

Dark Current Noise

All photodetectors will allow some current to flow even without any optical illumination. This is known as the dark current, I_D. In the case of a reverse-biased photodiode, this will be the reverse saturation current of the diode I_s. In the case of photoconductors, dark current will be due to the conductivity of the intrinsic semiconductor. It may seem that dark current is a simple offset which can be subtracted out, effectively making it zero. While that is truel, the shot noise from the dark current cannot be subtracted because it is random in nature. Assuming dark current subtraction, we can get:

$$\bar{n}_D = 0 \tag{15}$$

$$\sigma_D = \sqrt{\frac{I_D}{qB}}. \tag{16}$$

Other Sources of Noise

Depending on the type of photodetector, there could be other sources of noise. For example, carrier generation and recombinaton rates can produce a noise known as G-R noise in photoconductors. There is also another type of noise that increases with decreasing frequency, known as 1/f noise.

Total Noise

In terms of the shot noise, thermal noise and dark current noise sources, we can write the mean and variance of the whole system as

$$\bar{n} = \bar{n}_s + \bar{n}_t + \bar{n}_D = \bar{n}_s \tag{17}$$

$$\sigma^2 = \sigma_s^2 + \sigma_t^2 + \sigma_D^2. \tag{18}$$

As a result, the expression for SNR becomes:

$$\text{SNR} = \frac{\bar{n}_s^2}{\sigma_s^2 + \sigma_t^2 + \sigma_D^2}. \tag{19}$$

Example

Consider a 1μW optical signal at a wavelength of 0.6μm wavelength incident on a silicon photodetector with an external quantum efficiency of $\eta_e = 0.8$. The detection circuitry is at room temperature (300K), and has an equivalent resistance of 150Ω. The sampling rate is 10MHz. We can find the total SNR as follows:

First find the average number of electrons collected per integration time:

$$\bar{n}_s = \frac{\eta_e P}{Bh\nu} = 2.41 \times 10^5. \tag{20}$$

The shot noise deviation is equal to the square root of the mean number of detected electrons.

$$\sigma_s = 491. \tag{21}$$

Next, find the thermal noise quantities:

$$\bar{n}_t = 0 \tag{22}$$

$$\sigma_t = \sqrt{\frac{4kT}{q^2RB}} = 2.07 \times 10^4. \tag{23}$$

Clearly, in this example, thermal noise is far greater than shot noise.

We can calculate the overall SNR as

$$\text{SNR} = \frac{\bar{n}_s^2}{\sigma_s^2 + \sigma_t^2} = 135. \tag{24}$$

SNR is often expressed in dB, which is $10 \log \text{SNR}$. Therefore,

$$\text{SNR} = 21.3 \text{ dB}. \tag{25}$$

Noise Equivalent Power (NEP)

Noise-equivalent power, or NEP, is the incident power that will produce a SNR value equal to 1.0. This is also a useful performance metric because it indicates the minimum detectable power. Signals with SNR < 1.0 cannot be detected because it will be buried inside the noise. In fact, SNR $= 2$ is often considered the lowest detectable limit in practice.

NEP can be calculated by setting

$$\frac{\bar{n}_s^2}{\sigma_s^2 + \sigma_t^2 + \sigma_D^2} = 1. \tag{26}$$

Substituting the expressions for shot, thermal and dark current noise terms results in

$$\frac{\left(\frac{\eta_e P}{Bh\nu}\right)^2}{\frac{\eta_e P}{Bh\nu} + \frac{4kT}{q^2RB} + \frac{I_D}{qB}} = 1. \tag{27}$$

NEP can also be interpreted as the smallest *difference* in signal that can be detected.

Example (continued)

We can also calculate the NEP by solving for P in equation (27). This requires a nonlinear numerical solver. From this, we can get $P = 86$nW, which is the minimum detectable power. The given power of 1μW is clearly far greater than this value.

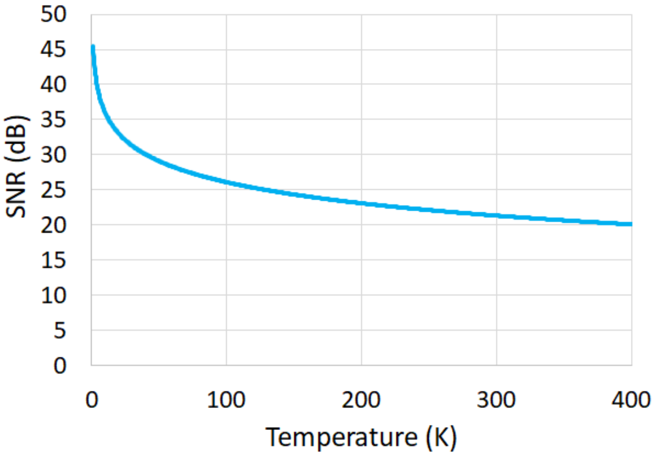

Figure 2: SNR vs temperature for P=1μW.

Looking at the deviation values from this example, we can see that the system performance is dominated by the thermal noise, because $\sigma_s \approx 500$ and $\sigma_t \approx 20,000$. Therefore, it is possible to improve the performance by reducing the temperature. For example, if the reduce the temperature to 100K, the SNR will increase from 21.3dB to 26.1dB, which is a more than twice the original SNR. The plot of SNR vs temperature is shown in Fig 2.

It is also possible for the system to switch from being thermal noise limited to shot noise limited, or vice versa. This can happen at a certain temperature, or at a certain input power level. We can find the power level at which the system will have equal contribution of noise from the shot-noise and thermal noise (at 300K). This power can be calculated by setting the shot noise deviation equal to the thermal noise deviation and solving for the P. That is:

$$\sqrt{\frac{\eta_e P}{Bh\nu}} = \sqrt{\frac{4kT}{q^2 RB}} \tag{28}$$

$$P = \frac{4kTh\nu}{q^2\eta_e R} = 1.78 \text{ mW}. \tag{29}$$

Therefore, when operating well above this power level, temperature changes will not have a significant effect on the SNR. The plot of the noise variances vs temperature is shown in Fig 3, as well as the SNR on the right hand scale.

Fig 3 shows the shot noise and thermal noise deviations and SNR as a function of incident power, for $T = 300$K. The NEP is also identified on the plot.

Detectivity

Earlier we saw that NEP was a measure of the minimum detectable signal. Optical receivers normally consist of a photodetector and bias circuit, and thermal noise often becomes the limiting case. A smaller value for NEP implies that the receiver has a high sensitivity. Detectivity is defined as the inverse of the NEP, such as

$$D = \frac{1}{NEP}. \tag{30}$$

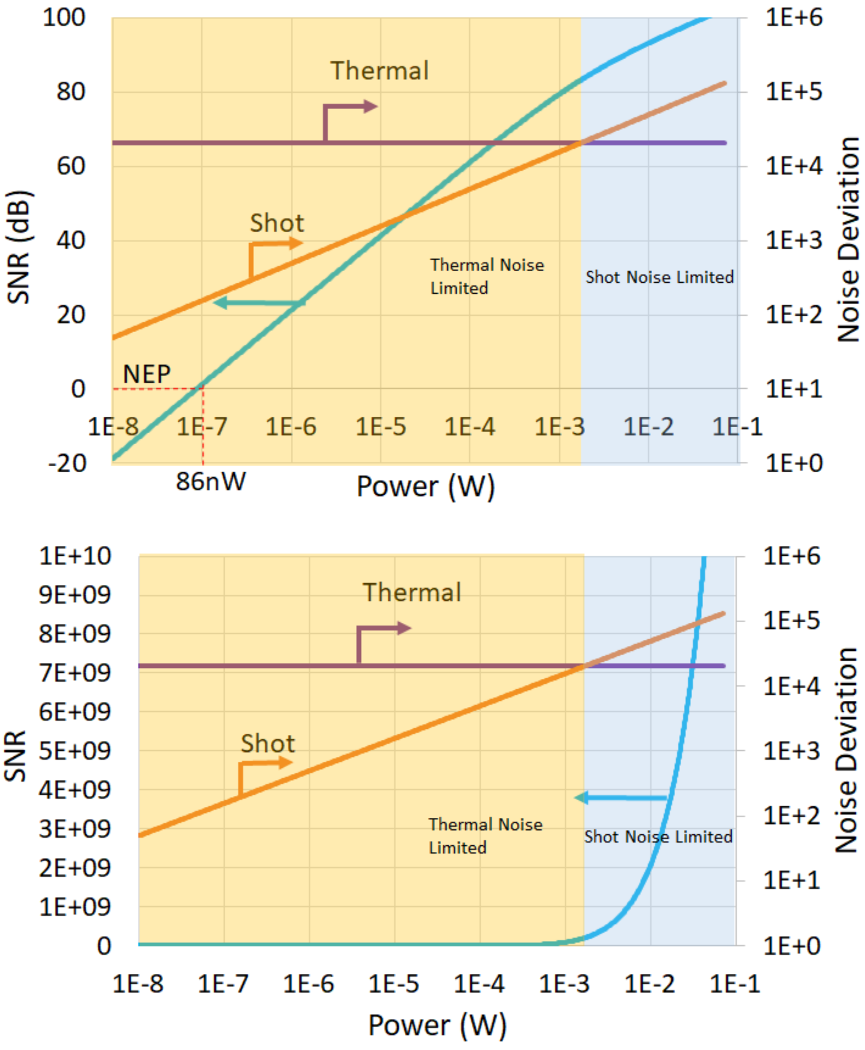

Figure 3: SNR, thermal and shot noise deviations vs input power for T=300K.

Noise Equivalent Temperature Difference (NETD)

Whereas NEPD was the smallest difference in incident power that that can be detected by the photodetector, in thermal detectors, we can also express this as a the temperature difference of the emitting source. In other words, NETD is the smallest difference in temperature that the detector can sense, assuming a blackbody radiation.

Amplifier Noise

Amplifiers are commonly used to increase the strength of signals. However, an amplifier will amplify the signal and the noise equally, resulting in no improvement. In other words, louder does not necessary mean better. In fact, since amplifiers are never perfect, they will add their own noise to the signal, which actually makes the amplified output worse than the input. However, an amplifier can improve the SNR in some circumstances. It depends on if the thermal noise is added before or after the amplifier.

Amplifier noise is accounted for by a dimensionless parameter F known as the excess noise factor. If an input signal has a mean $\bar{n}$ and a standard deviation σ, the output signal from the amplifier will have a mean $M\bar{n}$ and a standard deviation of $\sqrt{F}M\sigma$, where M is the amplifier gain. $F = 1$ represents a noise-less amplifier, but in all real amplifiers F will be larger than 1.

Thermal noise added before the amplifier

Consider the case where an amplifier is inserted into the system at the end of the detection

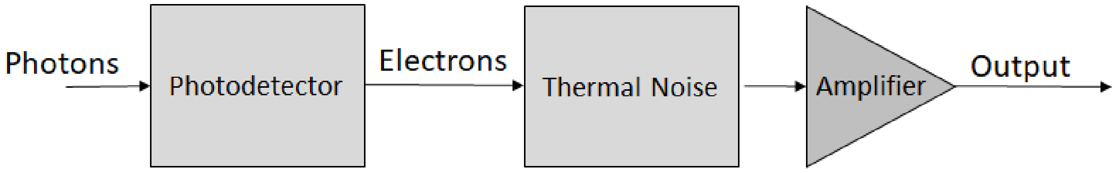

Figure 4: Illustration of thermal noise being added before the amplification

line. The photons are detected, converted into electrons, passed through some circuitry with thermal noise, and then amplified at the end with an amplifier with gain G and an excess noise factor F. In this case, we can write the mean and deviation of the output signal as:

$$\bar{n} = G\bar{n}_s \tag{31}$$

$$\sigma = \sqrt{F}G\sqrt{\sigma_s^2 + \sigma_t^2}. \tag{32}$$

We should note that when there are multiple noise sources present, the variances add, not the deviations.

We can consider two cases: with the amplifier, and without the amplifier. The SNR value with the amplifier will be:

$$\text{SNR}|_{\text{with amplifier}} = \frac{(G\bar{n}_s)^2}{FG^2(\sigma_s^2 + \sigma_t^2)} \tag{33}$$

$$= \frac{\bar{n}_s^2}{F(\sigma_s^2 + \sigma_t^2)}. \tag{34}$$

Without the amplifier, the SNR would be

$$\text{SNR}|_{\text{without amplifier}} = \frac{\bar{n}_s^2}{\sigma_s^2 + \sigma_t^2}. \tag{35}$$

Since $F > 1$, we can see that using the amplifier had only made the SNR worse. In other words, an amplifier at the final stage only makes the signal quality worse.

Thermal noise added after the amplifier

Next, consider the case where the amplifier is inserted early on in the sequence before any thermal noise gets added. In other words, the signal is amplified as soon as it is detected, then

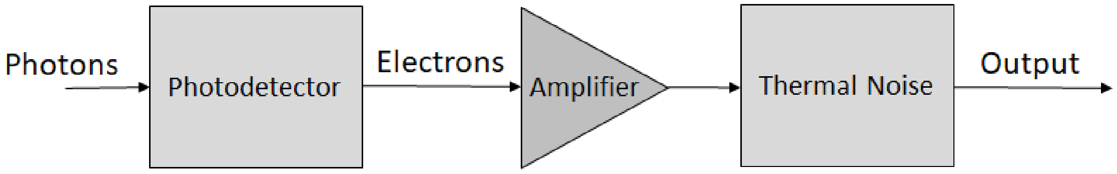

Figure 5: Illustration of thermal noise being added after the amplification

passed through the remaining circuitry. In this case, we can write the mean and deviations of the signal as:

$$\bar{n} = G\bar{n}_s \tag{36}$$

$$\sigma = \sqrt{\left(\sqrt{F}G\sigma_s\right)^2 + \sigma_t^2} \tag{37}$$

The resulting SNR values will be

$$\text{SNR}|_{\text{with amplifier}} = \frac{G^2\bar{n}_s^2}{FG^2\sigma_s^2 + \sigma_t^2} \tag{38}$$

$$= \frac{\bar{n}_s^2}{F\sigma_s^2 + \sigma_t^2/G^2}. \tag{39}$$

We can see that the effect of thermal noise has been reduced by a factor of G^2, except for the additional factor F. Without the amplifier, the SNR would have been:

$$\text{SNR}|_{\text{without amplifier}} = \frac{\bar{n}_s^2}{\sigma_s^2 + \sigma_t^2}. \tag{40}$$

We can conclude that an amplifier can be used to reduce the effects of noise only if it is inserted before that noise is added to the signal. This means the gain has to occur as close to the photodetector as possible. This is the fundamental reason for why APDs perform significantly better than other types of amplified photodetectors. The signal in an APD is amplified in the adjacent layer, before it even leaves the semiconductor material.

The strategy of amplifying the signal before adding noise applies to any types of noise, not just thermal noise. For example, if a cable has to be routed through a high noise environment, adding an amplifier prior to the point where it enters the high noise environment will reduce the impact of the noise by a factor of the gain.

Example

Consider a 1.5μm APD with $\eta_e = 0.5$, $G = 25$, and $F = 2$. The input optical power is 1μW. The sampling rate is 500MHz. The detected signal is fed to a circuit that has an equivalent resistance of 50Ω. Assuming everything is at room temperature, find the NEP of the system, using the APD as well as another equivalent regular photodiode (without gain).

$$\bar{n}_s = \frac{\eta_e P}{Bh\nu} = 7802 \tag{41}$$

$$\sigma_s = \sqrt{\frac{\eta_e P}{Bh\nu}} = 88 \tag{42}$$

$$\sigma_t = \sqrt{\frac{4kT}{q^2 RB}} = 5081 \tag{43}$$

$$\text{SNR}|_{\text{with APD}} = \frac{\bar{n}_s^2}{F\sigma_s^2 + \sigma_t^2/G^2} = 759 = 28.8 \text{ dB} \tag{44}$$

$$\text{SNR}|_{\text{with PD}} = \frac{\bar{n}_s^2}{\sigma_s^2 + \sigma_t^2} = 2.35 = 3.7 \text{ dB.} \tag{45}$$

Clearly we can see from this example that amplifying the signal immediately after detection has boosted the quality from 2.35 (barely detectable) to 759 (high quality).

Dynamic Range

Dynamic range of a sensor is the ratio between the minimum detectable signal and the maximum signal before the sensor becomes saturated. The minimum theoretically detectable signal is the NEP, even though in practice it will be several times that value. The maximum signal also has a limit. During the integration time, electrons are collected in a capacitor causing a linear increase in voltage with charge. However, the circuit will not be able to process any voltages higher than some maximum value, which is typically the supply voltage of the system. As a result, the signal will become saturated at high input power levels. If we represent this incident power as P_{sat}, the dynamic range becomes

$$\text{DR} = \frac{P_{\text{sat}}}{NEP}. \tag{46}$$

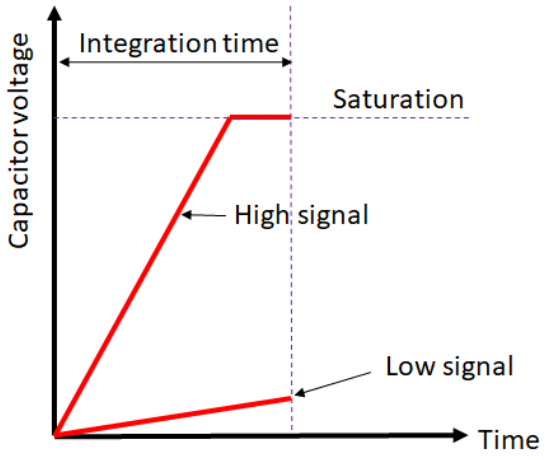

Figure 6: Illustration of photodiode saturation

Since all signals are eventually discretized into a digital representation, the dynamic range becomes the determining factor for how many discretization levels are needed to represent the signal (i.e., number of bits). Since it is not possible to detect a signal below NEP, there is no point in discretizing at finer intervals than the NEP. Setting the discretization increment to NEP, the number of discrete levels between NEP and P_{sat} becomes equal to the dynamic range DR. Therefore, the number of binary bits (bit-width) required to represent the signal can be written as

$$\text{Bit Width} = \log_2 (\text{DR}) \,. \tag{47}$$

For example, if the dynamic range is 4000, a 12-bit system should be adequate to represent the signal. This is illustrated in Fig 7.

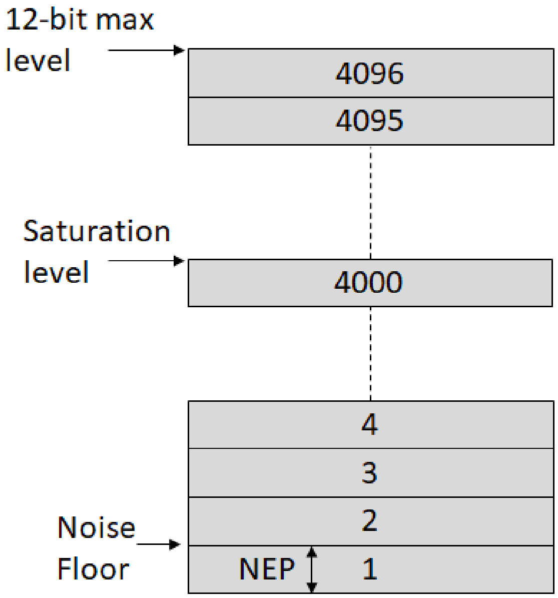

Figure 7: Representation of a 12-bit digital system

For comparison, the dynamic range of the human eye excluding the effects of the pupil and dark adaptation is about 14 bit-widths. Even the best cameras have only 12 bit-width of dynamic range. Consumer cameras have much lower dynamic range, of $8 - 10$ bit-widths worth.

Homework 10

1. A 1.55μm optical communication receiver is being used to detect incoming photons using a 100MHz sampling rate. For noise calculation purposes, the equivalent electrical impedance of the receiver can be assumed to be 75 Ω. The system must achieve a NEP better than 10nW. Several options are being considered:

   (a) A PIN photodiode with an external quantum efficiency (EQE) $\eta_e = 0.9$, and a negligible dark current. This is the cheapest option.

   (b) An APD with an $\eta_e = 0.5$, $G = 25$, $F = 2.5$, and a dark current of 1nA. But the cost is much higher than a PIN.

   (c) A thermoelectric cooling system to bring the temperature down to -50°C. This will add a moderate cost to the system.

   - Find the lowest cost option to achieve the NEP objective (i.e., PIN, PIN+cooling, APD or APD+cooling). Assume all factors other than the thermal noise remain the same when the temperature is lowered.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun main() {
    val k = 1.38064852e-23
    val q = 1.602e-19
    val hv = 1.24/1.55*q
    val B = 100.0e6
    val R = 75.0
    val P = 10.0e-9

    fun SNR(P:Double,etaE:Double,T:Double,G:Double,F:Double,Id:Double){
        val ns = etaE*P/(B*hv)
        val sigmaNs = ns.pow(0.5)
        val sigmaT = (4.0*k*T/(q.pow(2)*R*B)).pow(0.5)
        val sigmaD = (Id/(q*B)).pow(0.5)
        val snr = ns.pow(2) / (F*sigmaNs.pow(2) + sigmaT.pow(2)/G.pow(2) + sigmaD.
    pow(2))
        println(”ns = ${”%.2f”.format(ns)}”)
        println(”sigma_ns = ${”%.2f”.format(sigmaNs)}”)
        println(”sigma_t = ${”%.2f”.format(sigmaT)}”)
        println(”sigma_d = ${”%.2f”.format(sigmaD)}”)
        println(”SNR = ${”%.4f”.format(snr)} = ${”%.2f”.format(log(snr,base=10.0))}
    dB”)
        println()
}

    //PIN at 300K
    var etaE = 0.9
    var G = 1.0
    var F = 1.0
    var Id = 0.0
    var T = 300.0
    println(”PIN at $T K”)
    SNR(P,etaE,T,G,F,Id)

    //PIN at -50C
    T = 273.0-50.0
    println(”PIN at $T K”)
    SNR(P,etaE,T,G,F,Id)

    //APD at 300K
    etaE = 0.5
    Id = 1.0e-9
    T = 300.0
    G = 25.0
    F = 2.5
    println("APD at $T K")
    SNR(P,etaE,T,G,F,Id)

    //APD at -50C
    T = 273.0-50.0
    println("APD at $T K")
    SNR(P,etaE,T,G,F,Id)
}
--------------------------------------------------------
PIN at 300.0 K
ns = 702.25
sigma_ns = 26.50
sigma_t = 9277.67
sigma_d = 0.00
SNR = 0.0057 = -2.24 dB

PIN at 223.0 K
ns = 702.25
sigma_ns = 26.50
sigma_t = 7998.91
sigma_d = 0.00
SNR = 0.0077 = -2.11 dB

APD at 300.0 K
ns = 390.14
sigma_ns = 19.75
sigma_t = 9277.67
sigma_d = 7.90
SNR = 1.0969 = 0.04 dB

APD at 223.0 K
ns = 390.14
sigma_ns = 19.75
sigma_t = 7998.91
sigma_d = 7.90
SNR = 1.4719 = 0.17 dB
```

The APD at 300K is barely detectable, but APD at -50C is the only system that can produce an acceptable SNR.

Optical Diffraction Gratings

A grating is a periodic structure etched into a reflective or transmissive surface. The periodic nature of the interface can significantly modify the way electromagnetic waves interact with each other. A planar surface will produce a reflected wave at the same angle as the incident wave, but when the interface is periodic, reflection (or transmission) can occur at other angles. Most importantly, these other angles are a strong function of wavelength, which allows a grating to be used to produce a strongly wavelength-sensitive behavior. For example, a grating is used in spectrometers to disperse the light (i.e., send different wavelengths in different directions). Gratings can also be used as a wavelength-selective mirror, which we examined earlier in DBR and DFB lasers. Gratings can also be used to couple light into or out of a waveguide.

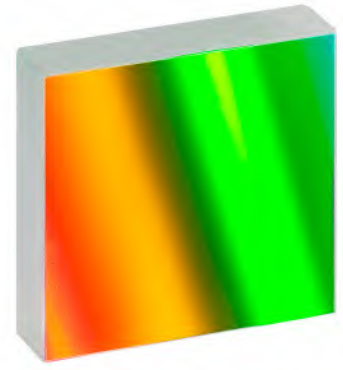

Figure 1: Reflection grating. Source: Thorlabs

This chapter is only a brief summary of the main principles of diffraction gratings. A comprehensive treatment is quite complex and requires coupled wave theory, which we will not pursue here.

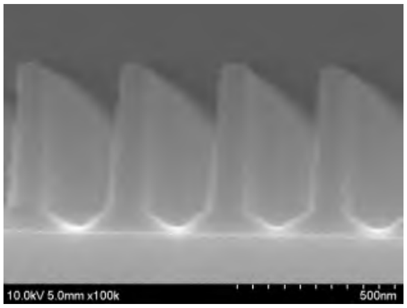

Figure 2: Close-up image of a grating.

Consider a plane wave incident on a planar optical interface. We can easily derive the relationship between the incident, reflected and transmitted waves. The relationship comes from the requirement that the phase fronts must be matched at the optical interface. In other words, the transverse k-vector (the k-vector that is parallel to the optical interface) must be equal on both sides of the interface. This is illustrated in Fig 3. The interface can be viewed as a coupling agent between the incident, reflected and transmitted waves.

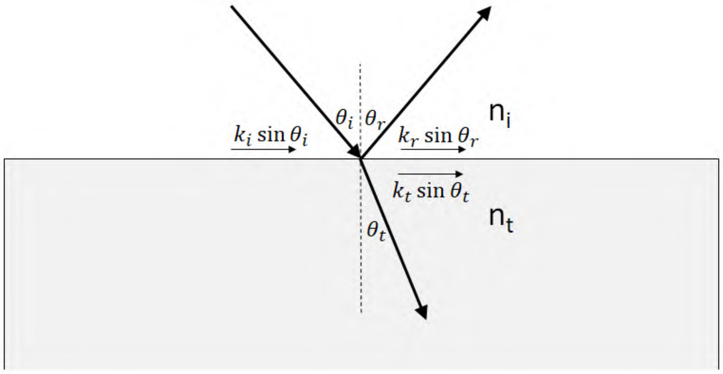

Figure 3: Fresnel reflection and transmission at a planar optical interface

For the reflected wave, this coupling can be mathematically expressed as

$$\begin{aligned} k_i \sin\theta_i &= k_r \sin\theta_r && (1)\\ k_0 n_i \sin\theta_i &= k_0 n_i \sin\theta_r, && (2) \end{aligned}$$

which results in

$$\theta_i = \theta_r. \tag{3}$$

For the transmitted wave, we have

$$\begin{aligned} k_i \sin\theta_i &= k_t \sin\theta_t && (4)\\ k_0 n_i \sin\theta_i &= k_0 n_t \sin\theta_t, && (5) \end{aligned}$$

which results in

$$n_i \sin\theta_i = n_t \sin\theta_t. \tag{6}$$

The intensity of the reflection and transmission is dictated by the field continuity relations. In this description, we will focus only on the direction of the beam and not on its intensity. A coupled-wave analysis is required for the the beam intensity in the presence of a grating.

A grating can be viewed as a standing wave at the optical interface. Instead of the interface being planar, in this case the interface has a periodic surface height modulation. This is illustrated in Fig 4. Alternatively, it can also be a periodic refractive index modulation. The grating is assumed to have a period of Λ with a corresponding wave vector of

$$K = \frac{2\pi}{\Lambda}. \tag{7}$$

Just like a plane optical interface, a grating interface can be viewed as a coupling agent between the incident, reflected and transmitted waves. However, in this case the relationship between the incident, reflected and transmitted waves will be different than the result derived for a plane interface.

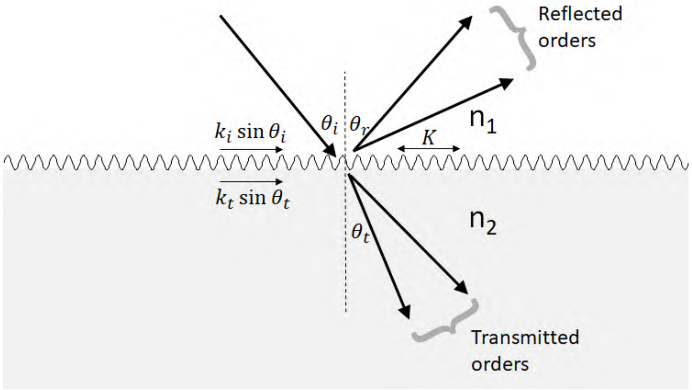

Figure 4: Wave interaction across an interface that has a grating

As stated before, a detailed analysis requires coupled-wave theory. But we can summarize the main results, at least with respect to the direction of the diffracted waves, in terms of the k-vectors of optical waves and the K-vector of the grating. Each reflected wave will have a transverse k-vector that will satisfy the following relationship:

$$k_i \sin\theta_i \pm qK = k_i \sin\theta_r \tag{8}$$

where q is an integer that refers to the order of the diffraction. $q = 1$ is referred to as the first order diffraction, and $q = 2$ is the second order diffractions, etc.. This equation is almost identical to the one before except for the presence of the grating vector K. We should also be able to identify that equation (8) collapses to equation (1) when $q = 0$. In other words, zeroth order reflection is the same as the ordinary reflection of the beam from a plane (non-corrugated) interface.

For the transmitted beam, we an similarly write

$$k_i \sin\theta_i \pm qK = k_t \sin\theta_t. \tag{9}$$

The grating pitch Λ is an important parameter that determines all of the reflected or transmitted angles except for the zeroth order ($q = 0$) beam. This is often quoted in product literature as the groove density. For example, a groove density of 1200/mm will have a period of $\Lambda = 833$ nm.

Whether a grating will act as a reflector, as a transmitter or as both depends entirely on the substrate refractive index n_t. If the substrate is highly reflective, such as a metal, there will be no transmitted orders. For highly transparent substrate, most of the light will be transmitted through the substrate (and the amount of reflection will depend on the refractive index contrast between n_i and n_t).

The intensity of each reflected or transmitted order, of course, is not part of our analysis. This will be determined by the refractive index of the grating as well as the height and shape of the modulation. Through careful design, it is possible to increase the power in one specific reflection (or transmission) order. A more detailed analysis requires the use of Rogorous Coupled Wave Analysis (RCWA).

Reflection Grating

In the case of a reflection grating, the incident and reflected media will be the same. As a result, equation (8) will become:

$$n_i k_0 \sin\theta_i \pm qK = n_i k_0 \sin\theta_r, \tag{10}$$

from which we can get the beam angle as

$$\begin{aligned} \sin\theta_r &= \sin\theta_i \pm \frac{qK}{n_i k_0} && (11) \\ &= \sin\theta_i \pm \frac{q\lambda}{n_i \Lambda}. && (12) \end{aligned}$$

We can see that there will be two beams corresponding to the ± signs. One is referred to as the $+q$ order, and the other is $-q$ order. However, both orders may not always be present. In some cases, one of the orders will result in a value for $\sin\theta_r$ larger than 1. This corresponds to an evanescent field, and will not produce a propagating beam in the far field.

Transmission Grating

For transmission, the incident and reflected media will be different. We can use equation (9) to get:

$$n_i k_0 \sin\theta_i \pm qK = n_t k_0 \sin\theta_t \tag{13}$$

$$n_t \sin\theta_t = n_i \sin\theta_i \pm \frac{q\lambda}{n_i\Lambda}. \tag{14}$$

Example

If a 632nm HeNe wavelength is incident on a reflection grating with a groove density of 1200/mm, at an angle of 45-degrees incidence, we can calculate the reflection angles corresponding to the different orders by using

$$\sin\theta_r = \sin\theta_i \pm \frac{q\lambda}{\Lambda}. \tag{15}$$

Order	$\sin\theta_r$	θ_r
0	0.707	+45
+1	1.46	-
-1	-0.05	-2.9
+2	2.22	-
-2	-0.81	-54.0
+3	2.98	-
-3	-1.56	-

Table 1: Diffraction orders for $\theta_i = 45$-deg, $\lambda = 632$nm with 1200/mm groove density

Therefore, only two orders will exist, because all other orders produce $\sin\theta_r > 1$. These reflection orders are shown in Fig 5.

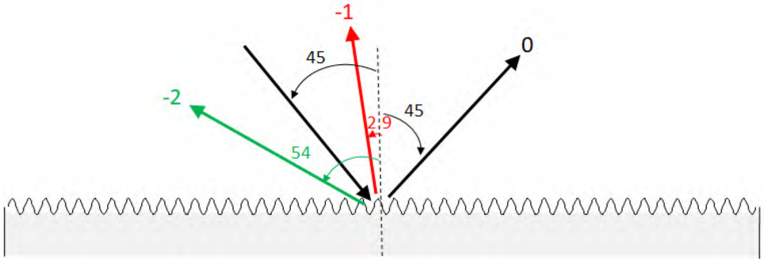

Figure 5: Diffraction orders from a 632nm light incident at 45-deg on a diffraction grating with 1200/mm groove density

Had this been a transmission grating, the diffractions angles calculated above will be exactly the same except they will be on the transmission side, assuming we are measuring the angles on the air side and not inside the substrate.

Grating Scattering Diagram

The form of equations (8) and (9) makes it possible to visualize the various diffraction orders on a scattering diagram. The diagram corresponding to the previous example is shown in Fig 6. We should be able to verify that this is a diagrammatic representation of equations (8) and (9). The radius of the upper semi-circle is equal to the wave vector $n_i k_0$. The $+1$ and -3 orders fall outside the circle, representing a field that propagates evanescently, so it will exist only in the near field.

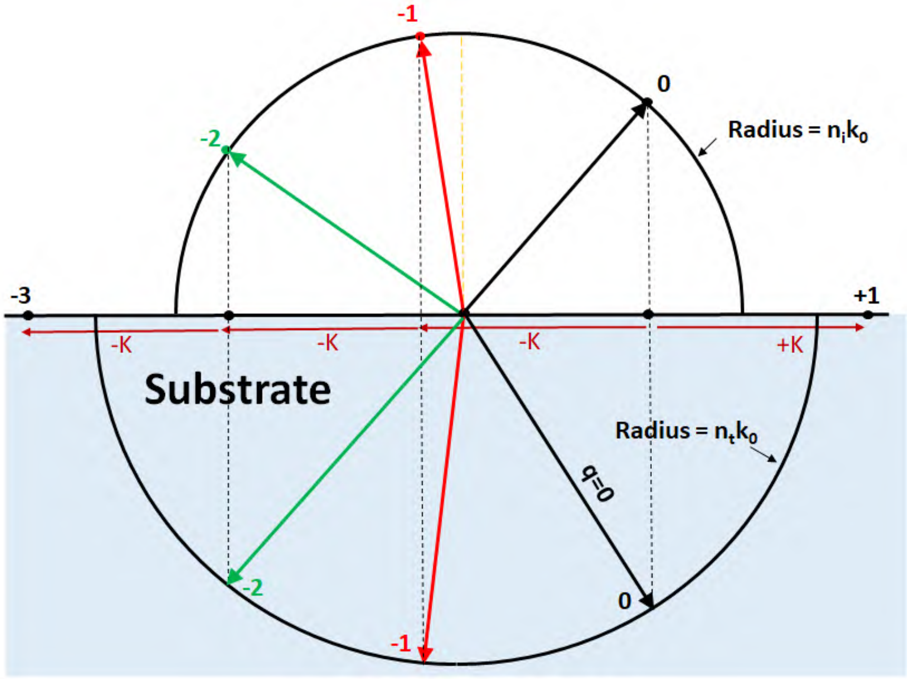

Figure 6: Representation of the diffraction modes from the previous example on a scattering diagram

The lower half of the circle is for transmission into the substrate. In the case of a transmission grating, diffracted modes will exist in the lower part of the circle. The radius of the lower circle is larger to account for the fact that the higher substrate index compared to the incident medium. We should be able to identify that it is possible, in some cases, to have a transmitted order without a reflected order. For example, if the substrate index were larger, it is possible for the $+1$ order to fall inside the radius of the lower circle, resulting in a transmitted diffraction.

CCD Spectrometers

CCD spectrometers are widely used to measure the spectral content of an optical signal. Their operation is primarily based on a diffraction grating. We can understand their basic operation using the theory discussed above.

Fig 7 shows the simplified building blocks of a typical CCD spectrometer. A small slit is used as a spatial filter of the incoming light beam. A collimating optic and a focusing optic are used to bring the beam to a focus on the CCD array. A diffraction grating is inserted in the path of the collimated beam to spatially separate the light spectrum. The separation of wavelengths will be determined by the groove density of the grating. This allows different wavelengths to come to a focus at different positions on the CCD array (we are making the assumption that there is no significant change in the focus distance due to the beam tilt). Each pixel on the CCD can be mapped for a specific wavelength, allowing a full display the entire spectrum in real time.

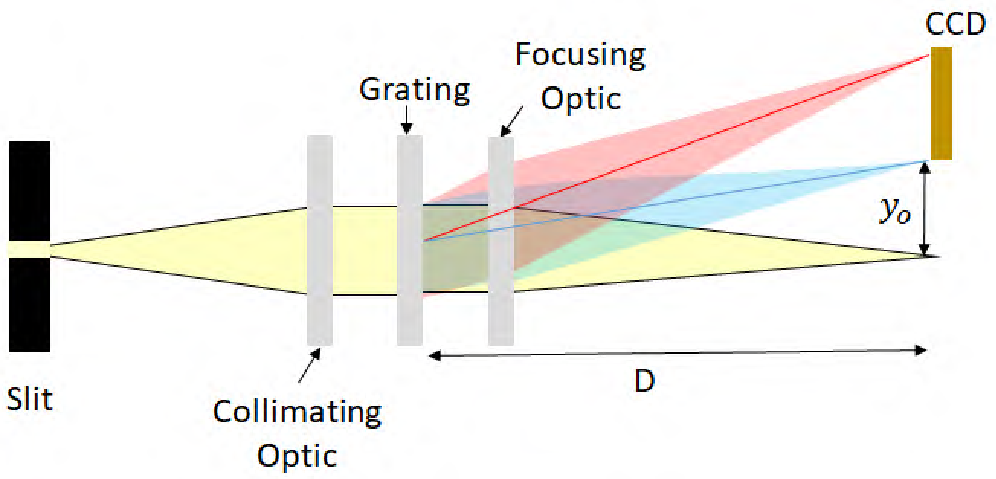

Figure 7: Simplified representation of the CCD spectrometer

Using equation (15), and referring to Fig 7, we can show that the vertical displacement of the beam focus on the CCD will be

$$
\begin{aligned}
y_\lambda &= D\tan\theta_t && (16)\\
&= D\frac{\sin\theta_t}{\sqrt{1-\sin^2\theta_t}} && (17)\\
&= D\frac{\sin\theta_i \pm q\lambda/\Lambda}{\sqrt{1-(\sin\theta_i \pm q\lambda/\Lambda)^2}}. && (18)
\end{aligned}
$$

If the pixel spacing on the CCD is Δ, the pixel number that corresponds to each y_λ can be written as

$$N_\lambda = \frac{1}{\Delta}\left(y_\lambda - y_o\right) \tag{19}$$

where N_λ is the pixel number on the linear CCD array, and y_o is the vertical translation distance of the CCD (the distance of the first pixel from the zeroth order focus point) as depicted in Fig 7.

For the case of a normally incident beam, assuming $+1$ order, equation (18) becomes

$$y_\lambda = D\frac{\lambda/\Lambda}{\sqrt{1-\lambda/\Lambda^2}} = \frac{D\lambda}{\sqrt{\Lambda^2-\lambda^2}}. \tag{20}$$

From this, we can get

$$N_\lambda = \frac{1}{\Delta}\left(\frac{D\lambda}{\sqrt{\Lambda^2-\lambda^2}} - y_o\right). \tag{21}$$

In practice, the pixel number that corresponds to each wavelength (equation (21)) is measured rather than calculated. During the calibration process the pixel number corresponding to known wavelength sources are measured, and then fitted to polynomial function to map all the pixels to a different wavelength.

Also note that we have applied equation (8) to this scenario even though the incident wave on the grating is not strictly a plane wave. This approximation is valid if the f/# of the focusing optic is large. In general, the effects of non-planar waves can be accounted for in the model by expressing the nonplanar wave as a sum of many plane waves using Fourier expansion. This allows us to calculate the interaction of each plane wave with the grating separately. Finally, an inverse Fourier transform is performed to combine all their effects together.

Example

Consider a 4096-pixel CCD spectrometer with a pixel pitch of 7μm, a grating groove density of 1200/mm ($\Lambda = 833$nm) with a grating-to-CCD distance of $D = 5$cm. We'll assume normal incidence on the grating. Referring to Fig 7, the CCD can be vertically translated to produce different spectral ranges. For instance, if the wavelength incident on the first pixel is 400nm, we can calculate the translation distance as

$$y_o = y_\lambda = \frac{D\lambda}{\sqrt{\Lambda^2 - \lambda^2}} = \frac{50 \times 400}{\sqrt{833.3^2 - 400^2}} = 27.3 \text{ mm}. \tag{22}$$

We can then calculate the wavelength on the last pixel by re-arranging equation (21) as

$$\lambda = \frac{\Lambda\left(N_\lambda \Delta + y_o\right)}{\sqrt{\left(N_\lambda \Delta + y_o\right)^2 + D^2}} = 621.7 \text{ nm}. \tag{23}$$

Therefore, the spectral range of this spectrometer is $400 - 621$nm. The average spectral resolution is

$$\overline{\Delta\lambda} = \frac{621 - 400}{4096} = 0.054 \text{ nm}. \tag{24}$$

It should be noted that this is an average resolution. The resolution will actually be a function of wavelength, an expression for which can be obtained by taking the derivative of equation (21).

Second-order ($q = 2$) Beam

So far we have ignored the effects of second order beam. In equation (18), the wavelength appears as $q\lambda$. That means, the $+1$ diffraction at a wavelength λ will be identical to the $+2$ diffraction at a wavelength $\lambda/2$. This is an important problem in in spectrometers. For example, the $+1$-order beam from 800nm wavelength will illuminate the same pixel as the $+2$-order beam from 400nm. One way to avoid this problem is by limiting the input wavelengths to less than a factor of two, such as $400 - 800$nm. This can be done, for example, with a long-pass filter on the entrance slit. Alternatively, or in addition to the above long-pass filter, it is also possible to use a linearly varying long-pass filter directly in front of the CCD array. For example, the 400nm beam incident at the same location as the 800nm beam will be blocked by this long-pass filter, allowing a greater spectral range than a factor of two. This is shown in Fig 8.

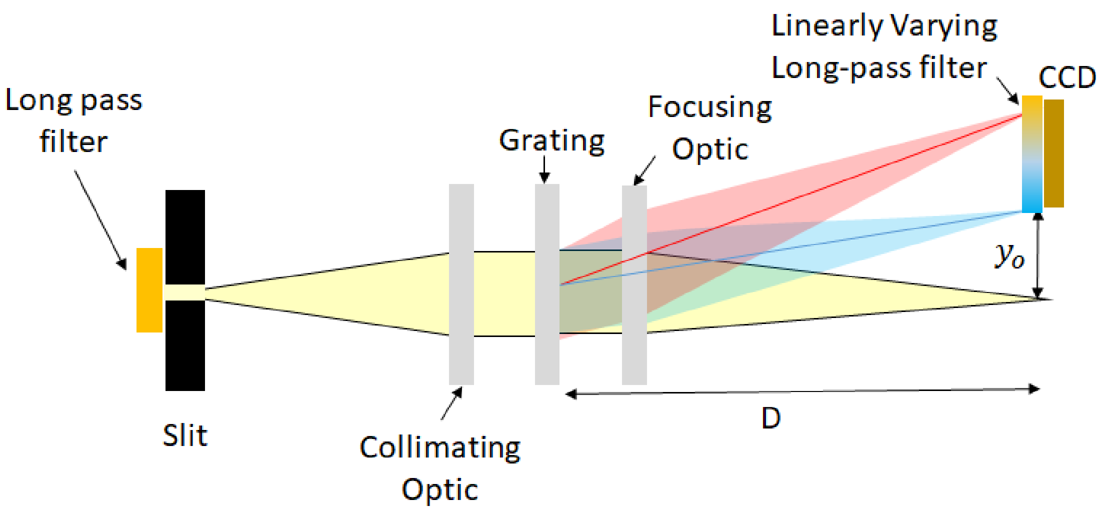

Figure 8: CCD spectrometer with long-pass filters to eliminate second order diffraction

Responsivity Correction

The measured raw signal from each pixel will not necessarily correspond to the actual intensity of the light incident on that pixel. This is because the responsivity function $\mathcal{R}$ is a strong function of wavelength. If the application requires accurate quantification of the signal intensity, then each pixel must be calibrated for their responsivity. Just like the spectral calibration, this is also done by fitting the intensity from a source with a known intensity and then fitting those data points to a polynomial.

Practical Implementation

Practical implementation of CCD spectrometers use reflective optics rather than refractive optics. Folding the beam using two or more mirrors helps to reduce the amount of space required for the system. An example is shown in Fig 9. Not shown in the figure are the CCD array and the varying long-pass filter (which is attached to the CCD), and the circuit board. The CCD sensor is part of a circuit board that collects and processes the signals, and sends it to a computer via a USB connection. Portable CCD spectrometers have become extremely common and inexpensive.

Waveguide Gratings

Gratings can also be incorporated into optical waveguides to accomplish useful functions, such as reflection, filtering and coupling. The grating can be part of the waveguide core, or cladding. The exact location affects the strength of the interaction with the grating, not the diffraction orders. Gratings in used in Distributed Bragg Reflector (DBR) lasers and Distributed Feedback (DFB) lasers to produce feedback. They are also used in Fiber Bragg Gratings (FBG) in fiber

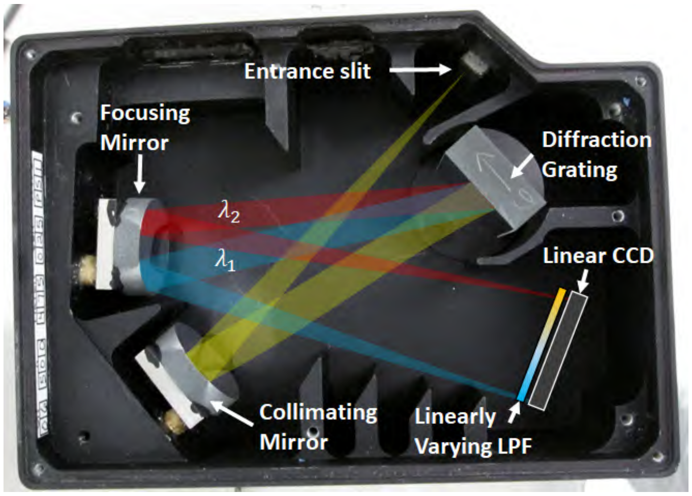

Figure 9: Internal construction of a CCD spectrometer

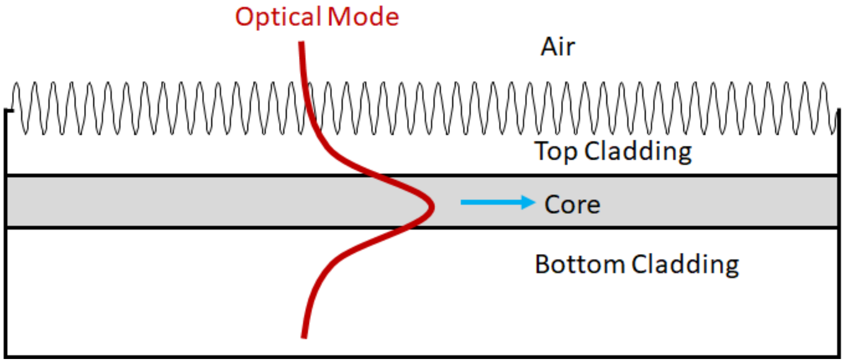

Figure 10: Waveguide with a grating etched into the top cladding

sensors and fiber lasers.

Radiation Mode In-Coupling

Without a grating, it is impossible to couple light into the waveguide from outside the cladding. This is a fundamental mathematical limitation that arises from the principle of orthogonality. Compared to edge-coupling, surface-coupling is a convenient method to couple light into a waveguide. This can be accomplished with the use of an appropriate grating.

Referring back to the scattering diagram concept that was introduced in Fig 6, the guided mode can be placed on the lower half of the circle. Since the mode propagation is parallel to the surface, it will have a reflection angle $\theta_t = \pi/2$. Furthermore, the propagation constant β will typically be slightly higher than the substrate medium. This is because the waveguide core index (which is not indicated in the scattering diagram) is higher than the clad. Therefore, the position of this guided mode will exist slightly outside the lower radius, as shown in Fig 11.

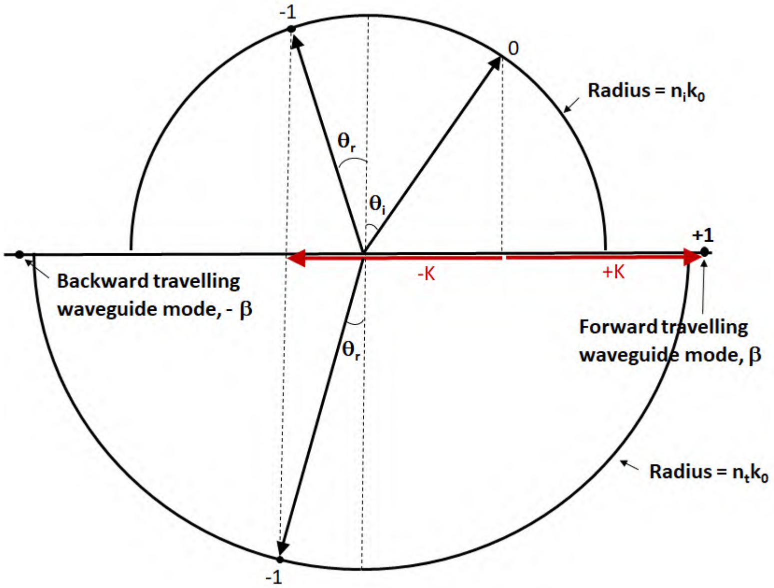

Figure 11: Scattering diagram for coupling into a waveguide

To couple an incident beam at an angle θ_i into the waveguide, we can modify equation (9) as

$$k_0 n_i \sin\theta_i \pm qK = k_0 n_t \sin\theta_t \quad (25)$$

$$k_0 n_i \sin\theta_i \pm qK = \beta. \quad (26)$$

Therefore, we don't really need to know the substrate index, just the effective index of the guided mode. Assuming $n_i = 1.0$ and $\beta = k_0 n_{\text{eff}}$, this can also be written in terms of the free space wavelength as

$$\Lambda = \frac{\pm q\lambda}{n_{\text{eff}} - \sin\theta_i}. \quad (27)$$

If we limit to $+1$ order diffraction, this becomes

$$\Lambda = \frac{\lambda}{n_{\text{eff}} - \sin\theta_i}. \quad (28)$$

Therefore, if we know the angle of incidence, we can calculate the required grating period. Alternatively, if we know the grating period, we can calculate the angle of incidence.

Example

If a waveguide mode has an effective index of 1.8, we can find the required grating period to couple 632 nm light at an incident angle of 45-degrees. Assuming $+1$ order diffraction:

$$\Lambda = \frac{\lambda}{n_{\text{eff}} - \sin\theta_i} = \frac{632}{1.8 - \sin(45)} = 578.3 \text{ nm}. \quad (29)$$

We can also note that the grating will produce -1 order. Using equation (12), we can calculate the radiation angle to be

$$\sin\theta_d = \sin\theta_i - \frac{\lambda}{\Lambda} = -0.38 \tag{30}$$

$$\theta_d = -22.6 \text{ deg.} \tag{31}$$

Radiation Mode Out-Coupling

It is also possible to do the reverse of the previous example, i.e., couple a guided mode to a radiation field. This makes it convenient to extract power out of the waveguide without relying on end-facets or edge-coupling. The same grating that couples light into a waveguide will also couple it out of the waveguide. Therefore the equations are almost identical to the in-coupling case. We can write this as

$$k_0 n_t \sin\theta_t \pm qK = k_0 n_i \sin\theta_r \tag{32}$$

$$\beta \pm qK = k_0 n_i \sin\theta_r. \tag{33}$$

Compared to equation (26), we can see that the $k_0 n_1$ and β have switched places. This is because the input wave is the guided mode, and the output wave is the radiation wave. This results in

$$\Lambda = \frac{\pm q\lambda}{\sin\theta_r - n_{\text{eff}}}. \tag{34}$$

Notice that compared to equation (28), n_{eff} and $\sin\theta_r$ have switched places.

Example

For the same waveguide as the previous example, with an effective index of 1.8, we can find the required grating period to out-couple 632 nm normal to the surface. In this case, the -1-order will diffract the forward-traveling waveguide mode to a surface-normal radiation field ($\theta_r = 0$). This results in

$$\Lambda = \frac{-\lambda}{\sin\theta_r - n_{\text{eff}}} = \frac{632}{1.8 - \sin(0)} = 351.1 \text{ nm}. \tag{35}$$

This is shown diagrammatically in Fig 12.

Bragg Reflector

While the previous examples were about coupling between a radiation mode and a guided mode, it is also possible to use a grating to couple between two guided modes. One example of this is the Distributed Bragg Reflector (DBR), where a forward traveling waveguide mode is coupled to a backward traveling waveguide mode.

In this case, the incident wave is the forward-travelling guided mode, and the diffracted wave is the backward-travelling guided mode. Therefore,

$$n_t k_0 \sin\theta_i \pm qK = n_t k_0 \sin\theta_t. \tag{36}$$

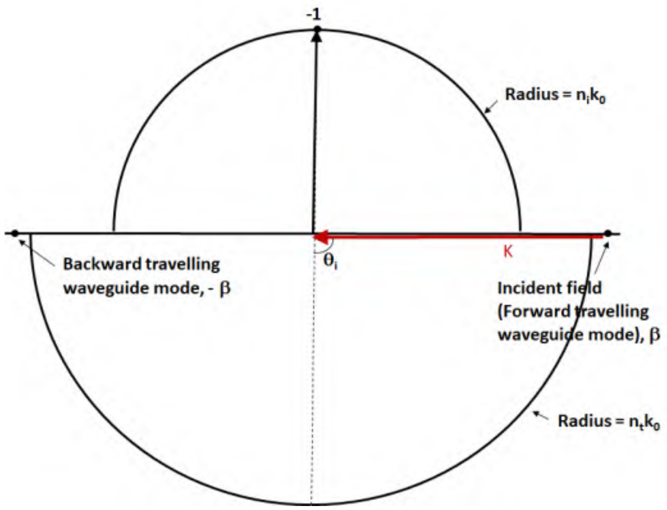
-1
Radius = n_ik₀
Backward travelling waveguide mode, - β
θ_i
K
Incident field (Forward travelling waveguide mode), β
Radius = n_tk₀

Figure 12: Scattering diagram for coupling out of a waveguide at 0-deg

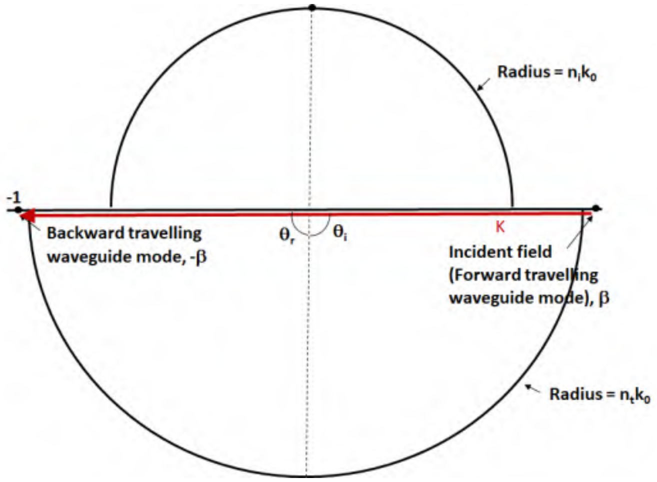
Radius = n_ik₀
-1
Backward travelling waveguide mode, -β
θ_r
θ_i
K
Incident field (Forward travelling waveguide mode), β
Radius = n_tk₀

Figure 13: Scattering diagram of a DBR reflector

In this case, the incident wave will have an angle $\pi/2$ and the diffracted wave will have $-\pi/2$, resulting in

$$\begin{aligned} n_t k_0 \sin\theta_i &= \beta \qquad (37) \\ n_t k_0 \sin\theta_r &= -\beta. \qquad (38) \end{aligned}$$

Therefore,

$$\beta \pm qK = -\beta, \qquad (39)$$

from which we can get

$$\pm qK = -2\beta. \qquad (40)$$

Assuming first order diffraction, we would need $q = -1$, resulting in

$$\begin{aligned} K &= 2\beta \qquad (41) \\ \Lambda &= \frac{\lambda}{2n_{\text{eff}}}. \qquad (42) \end{aligned}$$

In other words, the required grating period is half the wavelength (in the medium) of the guided mode.

Example

Assuming an effective index of 3.0 and a wavelength of 1550 nm, the required grating period that will produce a DBR reflection will be

$$\Lambda = \frac{\lambda}{2n_{\text{eff}}} = \frac{1550}{2 \times 3} = 258.3 \text{ nm}. \tag{43}$$

Homework 11

1. A spectrometer is being designed for operation between 450nm and 950nm, using a 8192-pixel CCD with a pixel pitch of 3μm. The distance between the grating and the CCD is 10cm, and the beam is designed to be normally incident on the grating. Calculate the required groove density of the grating.

   Run this code

```
import kotlin.math.*
//Andrew Sarangan

fun NewtonRaphson(f: (input:Double) -> Double, initialX:Double):Double{
    val dx = 1.0e-10*initialX
    var x1 = initialX
    var fprime:Double
    var x2:Double
    var diff:Double
    do {
      fprime = (f(x1) - f(x1-dx))/dx
      x2 = x1 - f(x1)/fprime
        diff = abs((x2-x1)/x1)
        x1 = x2
    } while (diff > 1.0e-12)
    return x1
}

fun main() {
    //Use equation (21) for 450nm and 950nm wavelengths.
    //At 450nm, N will be equal to 1 (first pixel)
    //At 950nm, N will be equal to 8192 (last pixel)
    //We have two equations and two unknowns (grating period and y0)
    //Substract one equation from the other to eliminate y0
    //Now we have one equation with one unknown (grating period)
    //This is the function f(P) in the following code.

    fun f(Lambda:Double) = (1.0/3000.0)*(950.0*10.0e7/(Lambda.pow(2)-950.0.pow(2)).
    pow(0.5) - 450.0*10.0e7/(Lambda.pow(2)-450.0.pow(2)).pow(0.5)) - 8191.0

    val Lambda = NewtonRaphson(::f,1000.0)
    println(”${”%.2f”.format(Lambda)} nm”)
}

>>2358.34 nm
```

Liquid Crystal Devices

Nematic Liquid Crystals

A liquid crystal (LC) is a state of matter which shares some of the properties of a crystalline solid as well as that of a liquid. Specifically, it can have a long-range order like in a solid crystal, but it can also flow like a liquid. This state exists within a temperature range above the melting temperature of the material but below an upper temperature limit above which it turns into an isotropic liquid. Just like in solid crystals, there is a large number of liquid crystals with different orientations and properties. The most common type is known as the nematic liquid crystal. The molecules of nematic crystals are rod-shaped and align themselves parallel to each other to form a crystalline structure as shown in Fig 1. Hence, nemaic liquid crystals will exhibit uniaxial anisotropic optical properties. In other words, the dielectric constant (and hence the refractive index) will be different when the electric field is parallel to the rods as compared to when the electric field is in either of the two transverse directions. The birefringence is defined as the difference between these two refractive index values, and will be denoted as Δn.

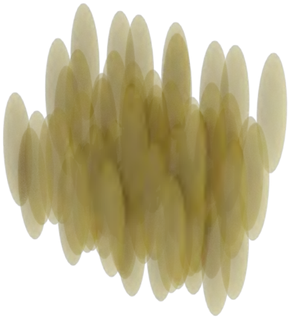

Figure 1: Nematic Liquid Crystals.

When a nematic liquid crystal is placed on a flat substrate, such as glass, the molecules typically orient themselves parallel to the substrate because this produces the lowest energy state of the liquid crystal. However, this can result in a random in-plane orientations. In order to force the molecules to align along a specific direction on the surface, one approach is to etch fine grooves on the substrate. This changes the symmetry of the plane. Molecules that align parallel to the grooves will have a lower free energy than those that align across the grooves. This forces the majority of nematic liquid crystal molecules to align parallel to the grooves. However, the asymmetry created by this technique is quite small, and it is difficult to ensure a large degree of alignment. Instead, a more effective technique is to utilize a separate alignment layer. Typically, this alignment layer is an organic film such as a polyamide. By buffing this polyamide layer with a finely textured cloth, we can create not only physical grooves, but the polyamid molecules will also re-orient themselves parallel to the rubbing direction. The molecular interaction between the liquid crystal molecules and the polyamide molecules enables a larger asymmetry in free energy to be created. This is by far the most commonly used method to align the liquid crystals on substrate.

What makes liquid crystals interesting is the fact that their molecules can move in response to an external electric field. When an electric field is applied, it will alter the electrostatic energy of the liquid crystal. However, since the liquid crystal is highly ansiotropic, this change in energy

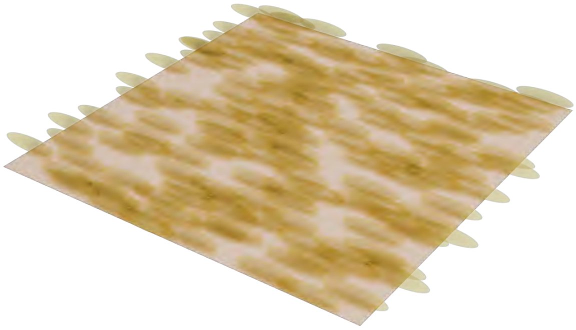

Figure 2: Attachment of nematic liquid crystals to a surface with an alignment layer

will depend on the direction of the electric field with respect to the molecular orientation. In a nematic liquid crystal, the lowest electrostatic energy is achieved when the rods are parallel to the applied electric field. Unlike in a solid crystal, the molecules in a liquid crystal will be able to re-orient themselves in a new direction the minimize total free energy of the system. In other words, we have two competing effects that determine the orientation of the liquid crystal molecules: The first is the polyamide alignment layer, which forces the liquid crystal molecules to be parallel to the polyamide molecules. The second effect is the applied electric field. The resulting final orientation will be determined by both of these effects.

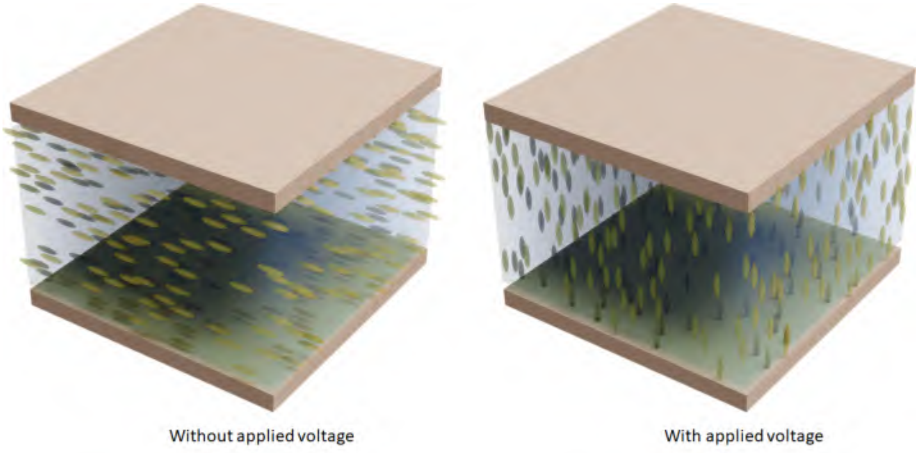

Figure 3: A nematic liquid crystal cell (with parallel alignment layers), with and without an applied field.

A liquid crystal cell consists of liquid crystal molecules sandwiched between two parallel substrates with electrodes on their inner surfaces (with appropriate alignment layers on each), as shown in Fig 3. When there is no field applied, all of the molecules in the liquid crystal will be parallel to the surface, and the incident light will experience two different refractive index values depending on its polarization. With a large voltage applied, the molecules will turn in the direction of the applied voltage. This results in the birefringence of the liquid crystal becoming zero. For intermediate voltages, the birefringence will have a smaller value. Therefore, nematic liquid crystals allow the user to control the birefringence of a liquid crystal cell.

A simple liquid crystal light modulator

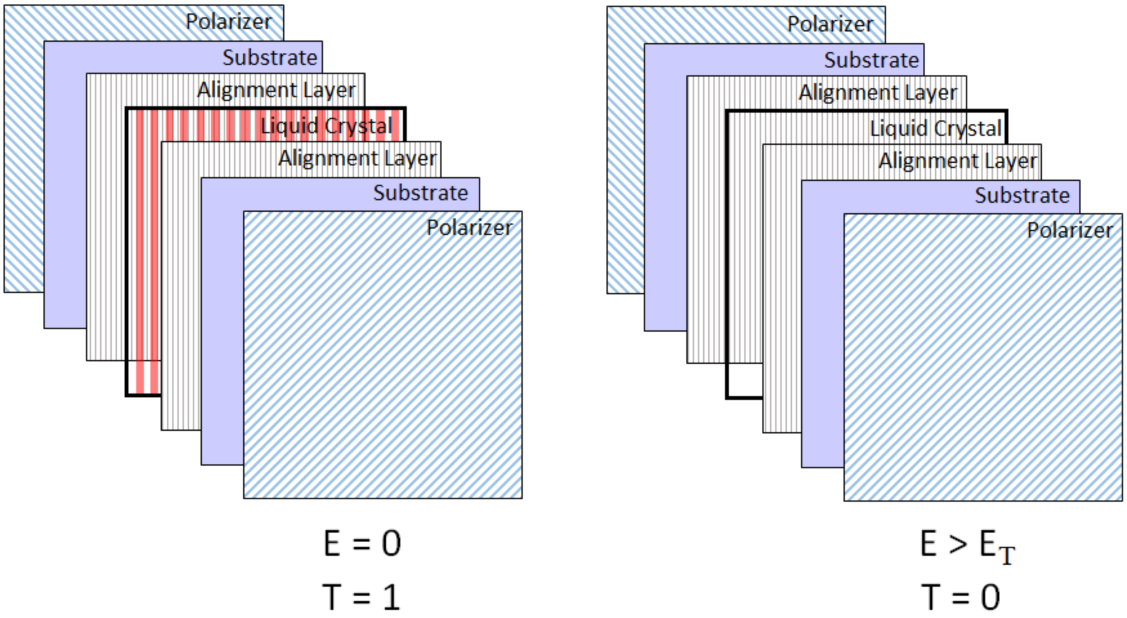

Figure 4: A nematic liquid crystal light modulating cell, with and without an applied field.

The simplest liquid crystal light modulator consists of a cell as shown in Fig 3 with two crossed polarizers at the incident and exit ends of the cell. If the incident polarizer is at 45° with respect to the alignment layer, the incident wave will be split into two orthogonal components traveling at different phase velocities through the liquid crystal. The phase difference between the two polarizations will be $k_0\Delta nz$ where z is the propagation direction. When the total phase difference becomes equal to π, the polarization of the incident wave would have rotated by 90°. Since the exit polarizer is rotated by 90° compared to the incident polarizer, it will result in a maximum transmission through this structure. Mathematically the transmission can be written as

$$T = \sin^2\left(\frac{k_0\Delta nt}{2}\right) \tag{1}$$

where t is the thickness of the liquid crystal cell. The birefringence value Δn will be a function of voltage, having its maximum value at zero field, and declining to zero as the field reaches a value of E_T. Increasing the field beyond this value will not change anything.

We can verify that the transmission will be zero when Δn is zero. Furthermore, we can also verify that the transmission is independent of wavelength. In other words, all wavelengths will have zero transmission. On the other hand, when Δn is at its maximum value Δn_{max} (which corresponds to zero applied field), the transmission value would be higher, but it will not be the same for all wavelengths. If we want the transmission at a wavelength of 550nm to be 1.0, assuming $\Delta n_{max} = 0.1$, the required cell thickness can be obtained by solving for

$$\frac{k_0\Delta n_{max}t}{2} = m\pi + \frac{\pi}{2} \tag{2}$$

where m is an integer. Assuming $m = 0$, the smallest cell thickness can be calculated as

$$t = 2.75\mu\text{m}. \tag{3}$$

This cell thickness would ensure that the 550nm transmission will decline from 1.0 to zero as the voltage in the cell in increased from zero to some sufficiently high value to fully polarize the liquid crystal molecules. The performance will be different at other wavelengths. Clearly, we cannot satisfy $\frac{k_0 \Delta n_{max} t}{2} = \frac{\pi}{2}$ at all wavelengths simultaneously. The plot of transmission versus Δn is shown in Fig 5. We can see that the transmission for 550nm wavelength declines monotonically from 1.0 to zero, but 400nm starts at a lower value, increases and then decreases to zero. Similarly, 700nm starts at a value smaller than 1.0 and declines towards zero. As a result, this liquid crystal cell will not be able to modulate all wavelengths equally. When a voltage is applied, the transmission spectrum will undergo a complex change, eventually settling down at a transmission of zero for all wavelengths.

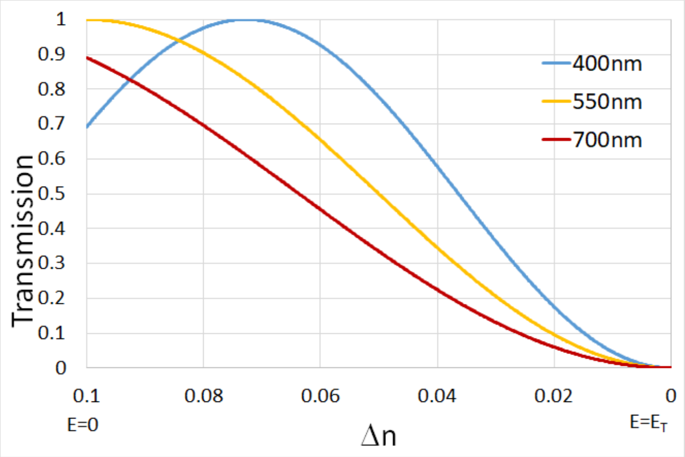

Figure 5: Transmission vs birefringence value of a parallel nematic liquid crystal cell, using $m = 0$, for $\lambda = 400$nm, 550nm and 700nm.

Using larger values of m would actually make this spectral sensitivity worse. For example, Fig 6 shows the transmission plot for $m = 1$ (corresponding to a cell thickness of 8.25μm). We can see that the oscillation pattern becomes more frequent as the cell is made thicker. As a result, this type of cell makes it difficult to achieve achromatic light modulation.

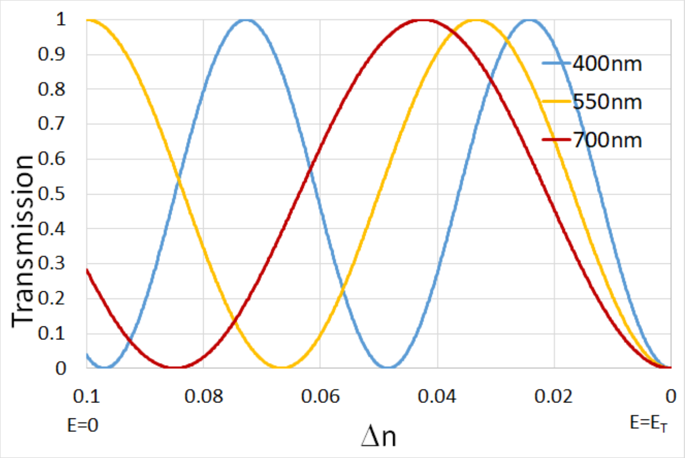

Figure 6: Same as Fig 5, except $m = 1$.

Twisted Nematic Liquid Crystal Light Modulator

In the previous case, the incident light was at 45° to the liquid crystal axis, and the polarization rotation was as a result of the phase velocity difference between the orthogonal polarizations. Naturally, the phase retardation between the two polarizations will be a function of wavelength, resulting in a modulation response that had significant spectral sensitivity. We can reduce this spectral sensitivity by allowing the light to be incident parallel to the liquid crystal axis, and rotating the liquid crystal axis instead. This can be accomplished by rotating the alignment layers by 90°. As a result, the molecules attached to the alignment layers will be rotated by 90° from the incident face to the exit face. This is illustrated in Fig 7.

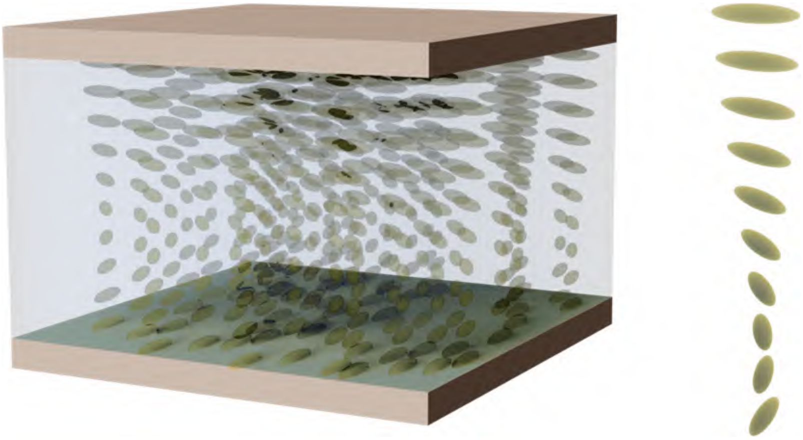

Figure 7: A twisted nematic liquid crystal cell with rotated alignment layers.

The construction of a light modulating cell is shown in Fig 8.

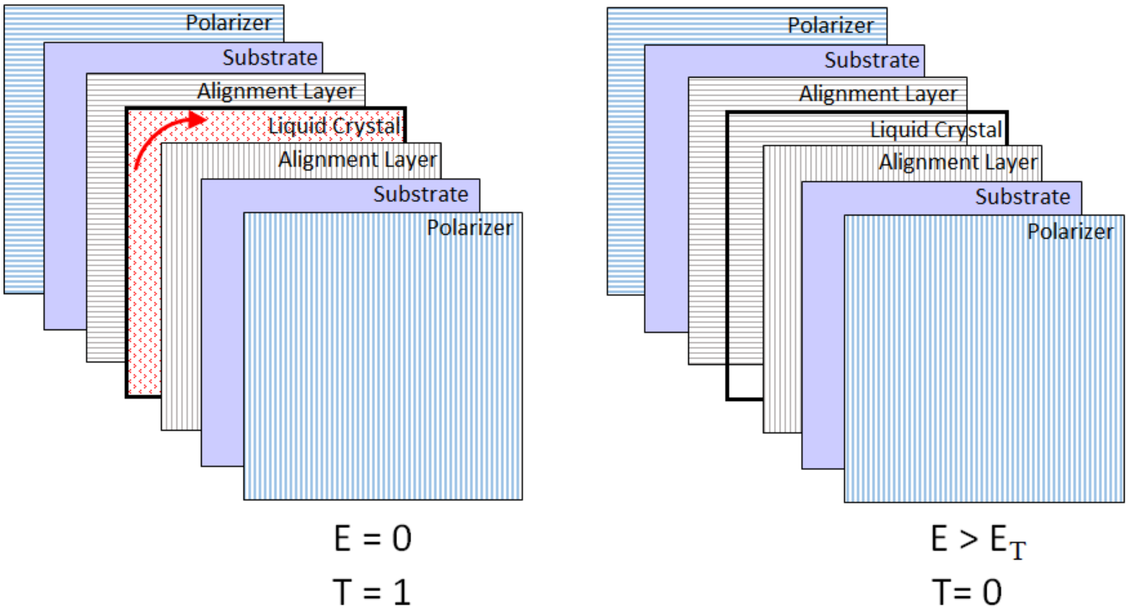

Figure 8: A twisted nematic liquid crystal light modulating cell, with and without an applied field.

The analysis of light propagation through a twisted nematic crystal is somewhat more involved than the parallel case, but we can summarize the result as

$$T = \cos^2 X + \left(\frac{k_0 \Delta n t}{2X}\right)^2 \sin^2 X, \tag{4}$$

where X is defined as

$$X = \sqrt{(\pi/2)^2 + (k_0 \Delta n t/2)^2}. \tag{5}$$

When $\Delta n = 0$ (at $E > E_T$), we can verify that $T = 0$ regardless of wavelength. This is similar to the untwisted liquid crystal modulator. When Δn is at Δn_{max}, the we can solve for the $T = 1$ condition by setting $X = m\pi$. This results in

$$\begin{aligned} (\pi/2)^2 + (k_0 \Delta n t/2)^2 &= (m\pi)^2 && (6) \\ t &= \frac{\lambda}{\Delta n}\sqrt{m^2 - \frac{1}{4}}. && (7) \end{aligned}$$

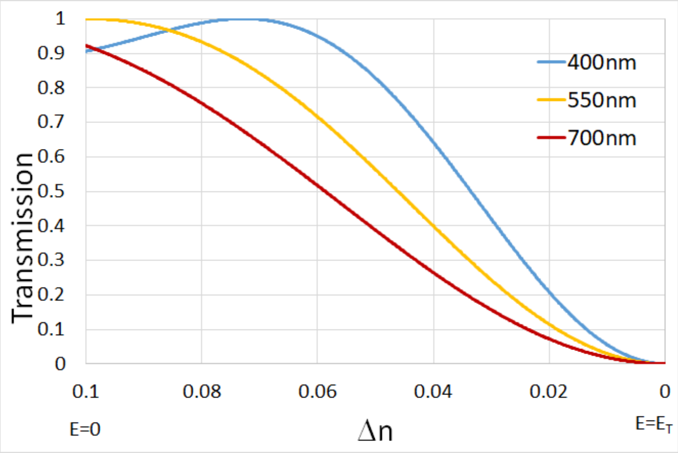

Figure 9: Transmission vs birefringence value of a twisted nematic liquid crystal cell, using $m = 1$ for $\lambda = 400$nm, 550nm and 700nm.

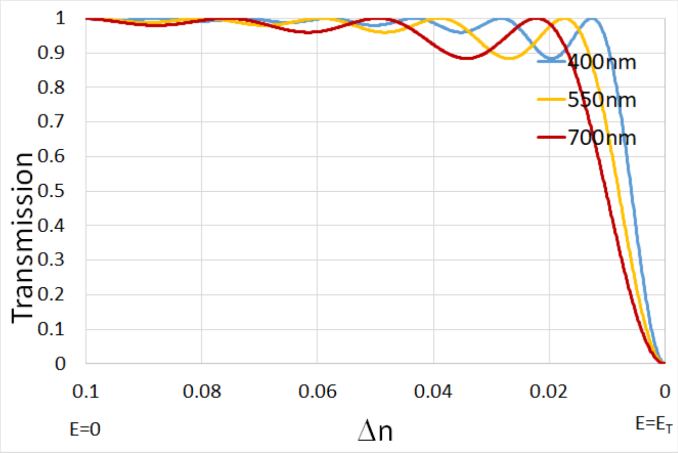

Figure 10: Same as Fig 9, except $m = 5$.

Assuming a wavelength of 550nm, $\Delta n_{max} = 0.1$, and $m = 1$, we can get $t = 4.7\mu$m. The transmission vs Δn is shown in Fig 9. Although the performance may not appear to be particularly better than the parallel nematic cell, when m becomes larger, we can observe a significant difference. The transmission curve for $m = 5$ (which corresponds to a cell thickness of 27.3μm) is shown in Fig 10. We can clearly see that the transmission remains high for all wavelengths for a relatively long range of Δn, and then abruptly falls to zero at Δn approaches zero.

Of course, we could have placed the incident and exit polarizers parallel to each other instead of being rotated by $90°$. This would result in zero transmission when the field is zero, and a high transmission when the field is greater than E_T. The first case is known as normally-white (NW), and the second case is known as normally-dark (ND).

Twisted Nematic Liquid Crystal Displays

Transmissive liquid crystals are widely used in computer displays and televisions. A light source is placed behind the entire panel, and is modulated on or off at each pixel location. The material used for the electrodes is a mixture of indium oxide (In$_2$O$_3$) and tin oxide (SnO$_2$), also known as ITO (Indium-Tin-Oxide). This is a conductive material that is also optically transparent. The twisted nematic liquid crystal is sandwiched between these two substrates. The thickness of the liquid crystal is controlled by placing glass spheres of precisely the right diameter to act as rigid spacers. To enable each pixel to be individually controlled, row and column electrodes are patterned (in separate ITO layers) on one of the substrates, and a ground plane is deposited on the other substrate. At the intersection of each row and column a thin film transistor (TFT) is placed to allow activation of that liquid crystal pixel. A TFT is a thin film version of a crystalline semiconductor transistor. It is produced by deposited amorphous silicon on the panels (rather than on bulk crystalline silicon) and then patterning them into transistor configurations. The row signal is connected to the drain terminal of the TFT, and the column is connected to the gate. This configuration is known as active matrix display because each pixel can be individually modulated by applying the voltage on the appropriate row and column. Crossed polarizers are placed on the outer faces of this configuration. Color can be introduced to each pixel by placing an absorptive color filter to only allow a narrow band of light to pass through (red, green or blue). This configuration is shown in Fig 11.

Fig 12 shows a microscope image of an LCD screen displaying three horizontal colored lines. The white background is produced by illuminating all three primary color pixels. Black pixels have no light output from any pixel. The red line is produced by illuminating just the red pixels. Other color combinations (such as cyan) are produced by controlling the amplitude of each color pixel.

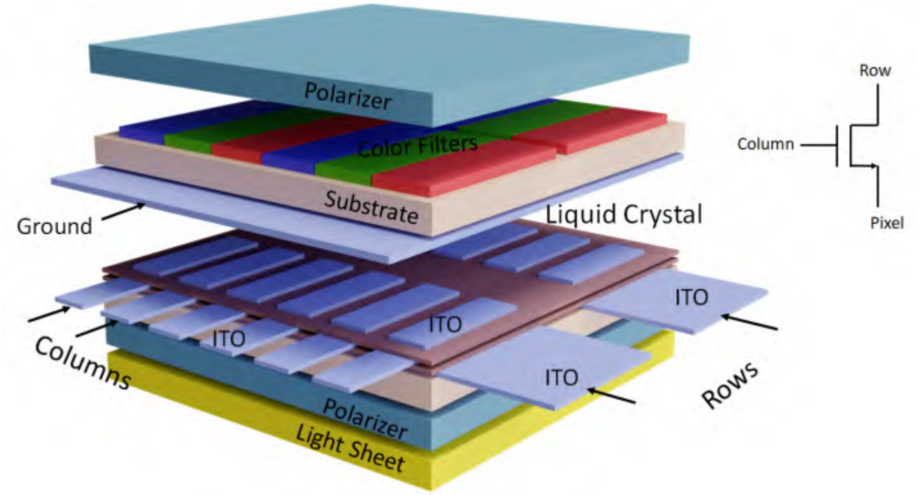

Figure 11: Transmissive liquid crystal configuration using thin film transistors

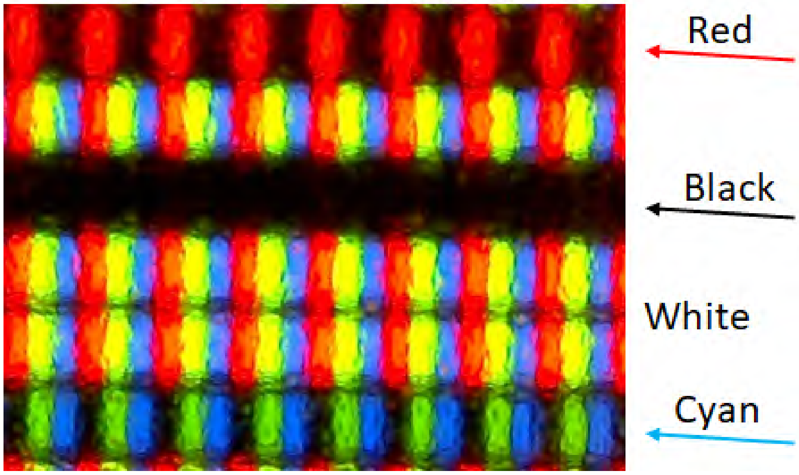

Figure 12: Microscope image of an LCD screen with three horizontal lines

In-plane switching (IPS)

In this configuration, the electrodes are placed side-by-side instead of across the faces of the liquid crystal cell. This produces an electric field that is parallel to the substrate's surface. In the absence of a field, the liquid crystal will exhibit the same behavior as the twisted nematic cell. With the field turned on, the molecules will align parallel to the field (effectively untwisting), resulting in a case similar to the parallel (untwisted) liquid crystal cell.

The polarizers across the cell are placed in the same orientation and in parallel to the incident alignment layer. When the applied field is zero, and if the twisted nematic cell thickness is correctly chosen to provide 90° polarization rotation, we will have zero transmission. When a field higher than E_T is applied, the molecules will untwist, and will be parallel to the incident (and exit) polarizers. Therefore the transmission will rise to 1.0. Therefore, this will be a normally-dark (ND) configuration. The transmission function can be written as:

$$T_{E=0} = \cos^2 X + \left(\frac{k_0 \Delta n_{max} t}{2X}\right)^2 \sin^2 X = 0 \tag{8}$$

$$T_{E=E_{max}} = 1.0 \tag{9}$$

With the field applied, even though the majority of the molecules will align parallel to the applied field, the ones closest to the substrate will still be rotated to align with the alignment layer. However, its contribution to the overall phase will be minimal since this region will be just a few monolayers thick.

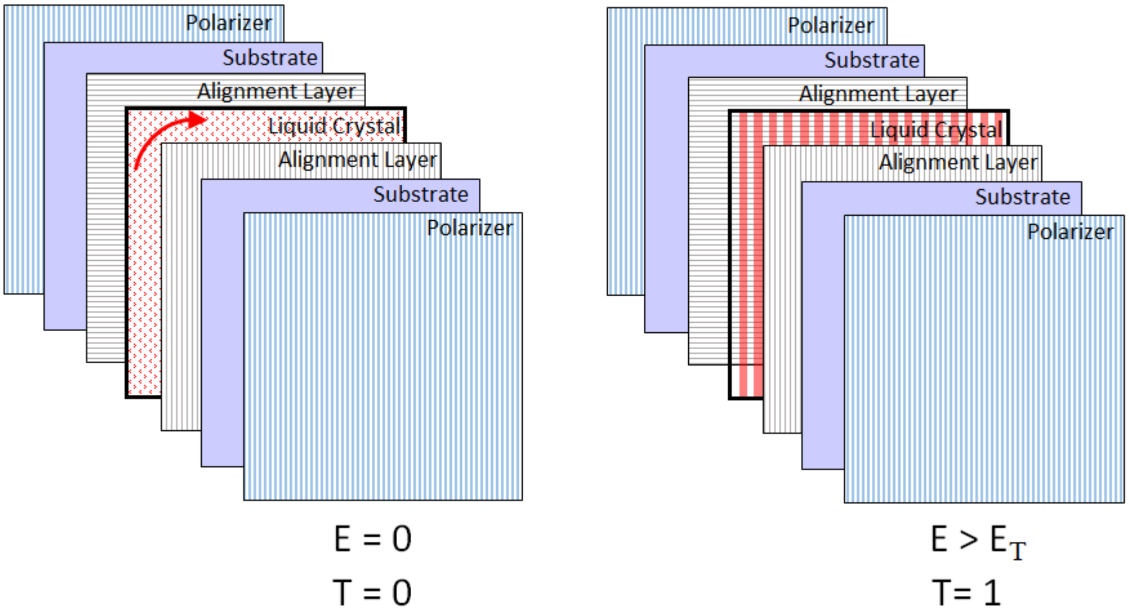

Figure 13: A twisted nematic in-plane liquid crystal light modulating cell, with and without an applied field.

Reflective Liquid Crystal Displays