351-2: Introductory Physics of Materials

David Barton

April 6, 2025

Contents

In the first quarter of this course, we developed a fundamental understanding of the underlying physics of materials. Notably, we developed quantum theory and solutions to the Schrd̈inger equation. The notion of wavefunction and physically measurable properties of materials was developed, as well as the uncertainty principle. Specific examples including particles in potential wells, quantum tunneling, and models for molecular orbitals became the foundation for understanding the energy levels of electrons near atoms.

Starting from this, the introduction to solid state physics presented an introduction to bonding in solid-state systems. Ending with energy bands from free electron models, the Density of states, and Fermi-Dirac statistics led to a deep understanding of how electrons can occupy energy states in materials and how the structure of atoms can dictate the band structure of materials, which in turns enables the understanding of metallic transport and semiconductors.

This course builds extensively on this understanding to understand electronic transport in semiconductors and semiconductor devices. This will culminate in the understanding of how transistors, LEDs, and solar cells can be designed an operate. We will additionally develop an understanding of the properties of dielectrics, manifesting their properties in frequency-dependent permittivity and optical properties of materials. Finally, a discussion of thermal properties of materials will be presented.

This course assumes a working understanding of the content introduced in the first semester of solid state physics and will use aspects of most of that course in this one. Reviewing old course content is strongly advised.

Most of this course reader follow through the text by Kasap. Unless otherwise noted, all images are based on this textbook. The relevant sections of the text are re-arranged slightly in this presentation in order to

1 Semiconductor Physics introduction

1.1 Review of 351-1

In class we reviewed some critical aspects of electronic properties of materials developed in 351-1. Notably, we found in 351-1 that the electronic band structure helps to determine the electronic properties of a material. This analysis shows that electronic conduction occurs within an electronic band of a material, and the energy within a band where you expect 50% occupation of an electron is defined as the fermi level. Therefore, we can classify material transport in a few different ways:

1) A metallic conductor is a material such that the fermi level is within the band of a material - that is, the band is not filled. 2) a semiconductor is a material where the fermi energy is between two bands, with the lower one (the valence band) is nominally fully filled and the upper band (the conduction band) is nominally void of electrons. Here, the gap between these bands is relatively small, of order 1-

2 eV. 3) An insulator has the same structure as an insulator, but the band gap is much larger, over 3 eV. The carrier concentration is significantly lower in an insulator (several orders of magnitude), and engineering the carrier concentration with dopants is typically rather challenging.

1.2 Effective mass, density of states, and carrier concentrations

We additionally found that the electron dispersion relationship can be modeled very similarly to the semiclassical free electron model, but substituting in an effective mass term that takes into account the crystalling potential. This was justified based on the behavior near the gamma point of the electronic band structure. With this, we find that the dispersion relation can be written as:

$$E = \frac{\hbar^2 k^2}{m_e^*} \tag{1.1}$$

where E is energy, $\hbar$ is the reduced Planck constant, k is the crystal momentum and m_e^* is the effective mass (note the *).

We can find from this that the crystal potential modifies the effective mass of an electron particle through the curvature of the band by considering various derivatives of this function:

$$\frac{dE}{dk} = \frac{\hbar^2 E}{m_e^*} \rightarrow \frac{d^2E}{dk^2} = \frac{\hbar^2}{m_e^*} \rightarrow m_e^* = \hbar^2 \left(\frac{d^2E}{dk^2}\right)^{-1} \tag{1.2}$$

Such that the band structure of a material impacts how the electrons can move and conduct, which leads to this description that the electron has a different mass than that of a free electron.

We additionally identified that this dispersion relationship leads to an approximate density of states that follows the following relationship:

$$g(E) \propto E^{1/2} \tag{1.3}$$

Finally, we noted that electrons are fermions and therefore follow the fermi-dirac distribution:

$$f = \frac{1}{\exp\left(\frac{E-E_F}{k_B T}\right) + 1} \tag{1.4}$$

Where f is the occupation value, E is the electron energy, E_F is the fermi energy, k_B is Boltzmann's constant, and T is absolute temperature. This equation shows that thermal excitations can increase the energy of an electron above the fermi energy. Note that the fermi level of a material is essentially the chemical potential for electrons, which will become useful as we describe metal-semiconductor junctions. This is shown schematically for a few temperatures in figure 1.1

Putting all of the parts together we identified that can determine the energy distribution of electrons in a material can be determined through the product $g(E)f(E)$,

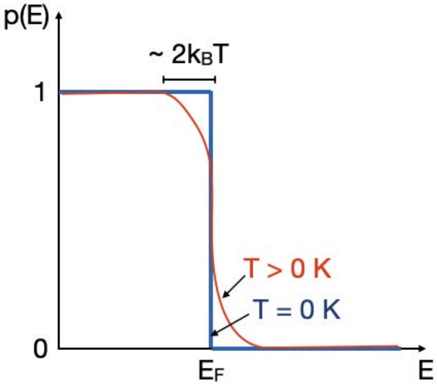


Figure 1.1: Example of Fermi-Dirac distribution at 0K and at a finite temperature. The distribution shows that thermal energy can create a population of electrons with higher energies. Image from 351-1 course reader.

and integrating this function can give the electron concentration:

$$n = \int_0^\infty f(E)g(E)dE \tag{1.5}$$

For a metal, this will modify the energy distribution in a straightforward manner. This is shown in figure 1.2:

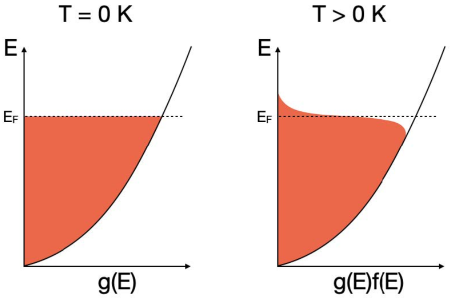


Figure 1.2: electron concentration and energy visualization. The product of Fermi-Dirac distribution and the density of states gives the energy diagram indicated above. Image from 351-1 course reader

For a semiconductor we can visualize the energy distribution within the conduction band and compute the total carrier concentration within the conduction band, as well as the total carrier concentration of holes in the valence band:

$$n_{cond} = \int_{E_C}^\infty f(E)g(E)dE \tag{1.6}$$

$$p_{val} = \int_0^{E_V} f(E)g(E)dE \tag{1.7}$$

where n_{cond} is the number density of electrons in the conduction band, p_{val} is the number density of holes in the valence band, E_V is the valence band energy and E_C is the conduction band energy (relative to 0).

We additionally know that the law of mass action sets a relationship between the concentration of electrons (n) and holes (p) through:

$$np = n_i^2 \tag{1.8}$$

Regardless of whether the semiconductor is an intrinsic or extrinsic semiconductor. For example, if we dope the system with substantially more n-type carriers that are activated at room temperature - that is, doping with $N_d >> n_i$, then we can determine that the electron concentration is $n = N_d$. Through the law of mass action we can therefore additionally conclude that $p = n_i^2/n = n_i^2/N_d$. The energy distribution of electrons and holes in a semiconductor is shown schematically in figure 1.3:

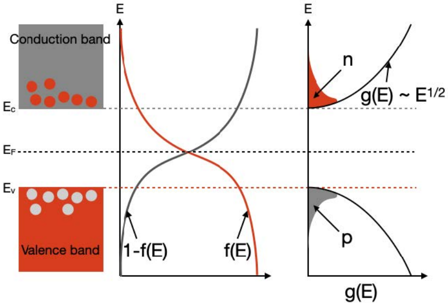


Figure 1.3: Visualization of the band structure of a semiconductor with associated energy levels of electrons and holes, defined through the product of the density of states and fermi-dirac distribution.

Because the Fermi energy E_f is related to the probability that a state is occupied by an electron, we additionally know that doping a semiconductor will raise or lower the fermi level within the band gap, depending on the carrier type. We can relate these through a few different factors. First, using the effective density of states at the valence band edge N_V and conduction band edge N_C, we can relate the carrier concentration to the fermi level through:

$$p = N_V \exp\left(-\frac{E_F - E_V}{k_B T}\right) \tag{1.9}$$

and

$$n = N_C \exp\left(-\frac{E_C - E_F}{k_B T}\right) \tag{1.10}$$

where

$$N_C = 2\left(\frac{2\pi m_e^* k_B T}{h^2}\right)^{3/2} \tag{1.11}$$

and

$$N_V = 2\left(\frac{2\pi m_h^* k_B T}{h^2}\right)^{3/2} \tag{1.12}$$

Here, we see that the band structure of the semiconductor determines the effective density of states through the definition of the effective mass for electrons or holes. Then, using this as a pre-factor, we multiply this by a Boltzmann factor for the energy difference between the fermi energy and the energy level of the valence or conduction band. This is analogous or an extension of the notion that the carrier concentrations are related to the density of states and the Fermi-Dirac distribution. The carrier concentration, band gaps, and fermi levels for semiconductors are are intrinsic, p-type, and n-type, are shown in figure 1.4:

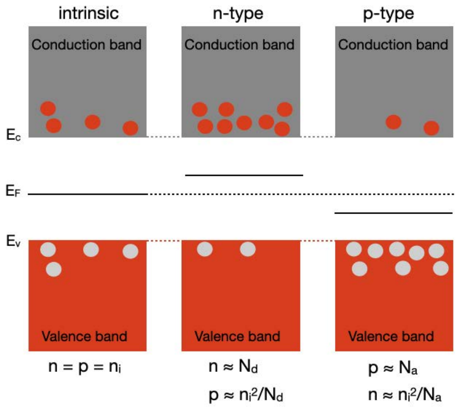


Figure 1.4: Fermi level and visualization of carriers when intrinsic, p-type, and n-type. The fermi level increases when more electron donors are doped into the crystal, whereas more holes leads to a reduction in the fermi leve.

Note that, with the law of mass action, that we can therefore directly determine the intrinsic carrier concentration by combining these equations together to form:

$$np = n_i^2 = N_C N_V \exp\left(-\frac{E_g}{k_B T}\right) \tag{1.13}$$

Where the energy levels for the fermi level disappear and the energy difference between the conduction band edge and the valence band edge is equivalent to the band gap energy E_g. This additionally shows that, for a particular temperature, that the carrier concentration should be higher for materials with a smaller band gap. This makes sense, as the energy gap is smaller such that the boltzmann factor becomes (much) larger due to its exponential dependence with inverse temperature.

1.3 Electronic conductivity consideration

In contrast to a metal, the carrier concentration of a semiconductor is modified strongly as the temperature is increased. In 351-1 we determined that there are a few things to consider. First, dopants themselves need to be thermally activated, as they form shallow donor or acceptor states within the band structure. Thermal energy is required to ionize a dopant (for example, for As -> As+ + e-). This means that there can be a temperature dependence to the carrier concentration of electrons and holes. As temperature increases, the intrinsic concentration of generated electrons and holes additionally goes up through the fermi-dirac distribution. This means that there are 3 regimes to consider:

1) Ionization regime: Low temperature. As temperature increases, more donors or acceptors will be thermally activated until the saturation energy, when nearly all carriers have been activated.
2) Medium temperature (near room temperature, typically): All donors or acceptors have been ionized and activated, so concentration of carriers is nearly constant.
3) High temperature: The temperature is high enough such that the intrinsic generation of electrons and holes is now larger than the doping concentration, and increases with temperature. The material now acts as if it was an intrinsic semiconductor.

These ranges are shown schematically in figure 1.5:

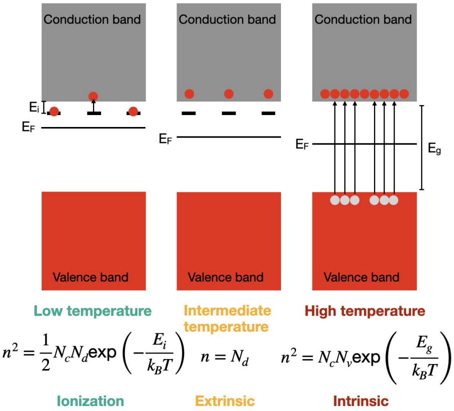


Figure 1.5: Different temperature regimes where carrier concentration changes. Figure from 351-1

The temperature dependence, because it has a Boltzmann factor involved, is more easily described by looking at an Arrhenius-type plot, with the x axis as inverse temperature and the y-axis as the natural log of carrier concentration. This is shown in figure 1.9

As can be seen, the carrier concentration is dramatically different depending on the temperature, which implies that the conductivity of this material will also vary substantially. However, we need to discuss the relevant mechanisms that can limit the overall conductivity. Here, we note that electrons and holes both give rise to the conductivity of a material, such that their mobilities μ_e and μ_h must also be understood. The conductivity σ is expressed as:

$$\sigma = en\mu_e + ep\mu_h \tag{1.14}$$

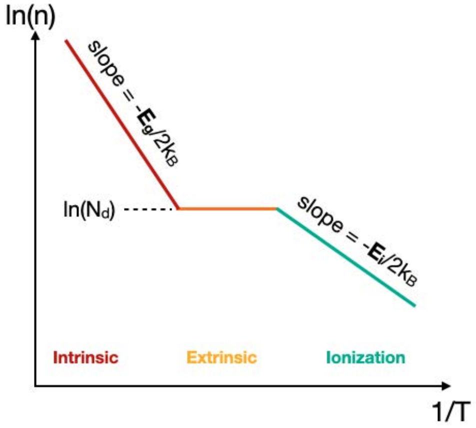


Figure 1.6: Arrhenius plot showing the relevant regions for the carrier concentration. Figure from 351-1.

where

$$\mu_e = \frac{e\tau_e}{m_e^*} \tag{1.15}$$

and

$$\mu_h = \frac{e\tau_h}{m_h^*} \tag{1.16}$$

Here, we are treating the particles semi-classically such that we can consider the scattering times τ_e and τ_h as being related to the time it would take for a spherical particle to scatter off of a lattice site, defect, grain boundary, etc. This can be expressed as:

$$\tau = \frac{1}{Sv_{th}N_S} \tag{1.17}$$

where S is the cross-sectional area of the particle, v_{th} is the thermal velocity, and N_S is the number density of scattering objects (lattice sites, defects, etc.).

There are two key ranges to be discussed. First, we find that at high temperature, the dominant scattering mechanism is scattering off of lattice sites. This makes sense, as thermal energy will nominally increase the effective size of the atom due to thermal vibrations. The average thermal energy we can treat as a semi-classical particle where $1/2m_e^*v_{th}^2 = 3/2k_BT$. The size of the vibrating atoms at lattice sites is given as $S = \pi a^2$, with the effective radius of the atom a increases with temperature as $a^2 \propto T$. This, combined with the fact that the thermal velocity $v_{th} \propto T^{1/2}$ leads to a lattice scattering-limited mobility μ_L that scales as $\mu_L \propto T^{-3/2}$.

Conversely, at low temperatures the dominant scattering mechanism is different. Here, the electrons and lattice sites are relatively sluggish due to the lower temperature. Here, we need to consider the Coulombic interaction between a charged particle (say, the free electron), and a charged dopant site. If the electron gets too close to the dopant site, it will scatter because of the coulombic attraction or repulsion (depending on the charge balance). The criterion for this to occur is when the kinetic energy of the electron matches the potential energy of the coulombic interaction, such that $KE = |PE|$. The potential energy is given by:

$$|PE| = \frac{e^2}{4\pi\epsilon_0\epsilon_r r} \tag{1.18}$$

(here, ϵ_0 is the permittivity of free space and ϵ_r is the relative permittivity of the medium, and r is the distance between the electron and the charged ion core. We can then set the kinetic energy $KE = 3/2k_BT$ equal to this to find the critical radius r_c such that:

$$r_c = \frac{e^2}{6\pi\epsilon_0\epsilon_r k_B T} \tag{1.19}$$

This sets the scattering cross section S such that:

$$S = \frac{\pi e^4}{(6\pi\epsilon_0\epsilon_r T)^2} \propto T^{-2} \tag{1.20}$$

Because the thermal velocity $v_{th} \propto T^{1/2}$, this means that the impurity scattering limited mobility μ_I still depends on temperature, but with the opposite scaling: $\mu_I \propto T^{3/2}$.

These phenomena dominate in different temperature regimes, but in intermediate temperatures they will add in parallel. This means that we can use Matthiessen's rule such that:

$$\frac{1}{\mu_e} = \frac{1}{\mu_I} + \frac{1}{\mu_L} \tag{1.21}$$

Such that the net mobility of the electron μ_e depends both on scattering off of impurity sites and lattice scattering sites. Putting all of this together, we find that the temperature dependence of the conductivity of a doped semiconductor takes a complicated form, as indicated in figure 1.7:

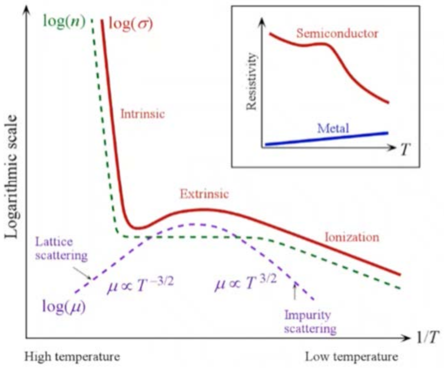


Figure 1.7: Arrhenius plot showing the relevant regions for the conductivity of a doped semiconductor. The non-monotonic mobility dependencies are also displayed. Figure from 351-1.

What is clear is that the physics of semiconductors is rich. One thing to note is that there is a fundamental difference between the temperature dependence of a metal and that of a semiconductor. The resistivity ρ of a material is the inverse of the conductivity. For a metal, the resistivity goes up, as the lattice scattering mechanism we discussed is the predominant scattering mechanism that will decrease the mobility of the electrons in the metal. For a semiconductor, the strong temperature dependence of the carrier concentration means the opposite - in general, we expect the resistivity of a semiconductor to go down when we increase the temperature.

1.4 Absorption and Photoconductivity

We have considered the equilibrium concentration of electrons and holes in semiconductors. However, there are several ways in which the carrier concentration can be modified. We will consider here the role of optical absorption on the creation of carriers and arrive at an expression that may be useful in the development of a photodetector, as a change to the carrier concentration therefore modifies the conductivity.

1.5 Optical Absorption

The steady-state concentration of carriers is set by the law of mass action. This is because at equilibrium, only thermal energy can excite electrons into the conduction band. They will reside there for a certain amount of time before recombining with a hole in the valence band. The steady-state of this sort of system gives us the law of mass action. However, we can use light to provide additional energy for the electrons. Here, we treat light as photons that can be absorbed by an electron to provide additional energy. This requires that the photon has enough energy to promote an electron into an unoccupied state. For a semiconductor, this requires that the energy of the photon be greater than the band gap energy - that is:

$$h\nu \geq E_g \tag{1.22}$$

where h is Planck's constant and ν is the frequency of the photon - you may also sometimes see this as f. the frequency is defined as $\nu = \frac{c}{\lambda}$, where λ is the wavelength of light and c is the speed of light.

When a photon is absorbed, it will promote a single electron into the conduction band. If the energy of the photon is larger than the band gap energy, then the additional energy will quickly be thermalized to the conduction band edge by emitting some heat. This is represented in figure 1.8:

At the same time, the intensity of light within the semiconductor must go down, as the total number of photons has been reduced. We can model this using the Beer-Lambert law. Here, we can define the incident intensity in the semiconductor I_0 based on the energy of the photon and the number of photons. Note that Intensity has units of Power/area. The initial intensity can be therefore given as:

$$I_0 = h\nu\Gamma_{ph} = \hbar\omega\Gamma_{ph} \tag{1.23}$$

where Γ_{ph} is the flux of photons inside of the semiconductor and has units of photons/area/s. After absorbing a photon, the flux of photons immediately after it (a distance of δx will be reduced by the *absorption coefficient* α. This is essentially a

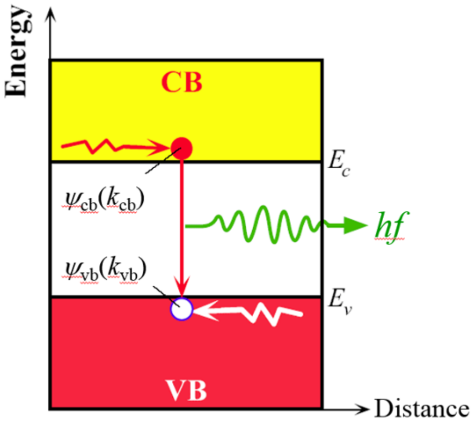


Figure 1.8: Overview schematic of optical absorption in a semiconductor. When a photon energy is greater than the band gap energy, an electron hole pair can be generated through the absorption or loss of a photon. The optical power in the semiconductor goes down while the carrier concentration goes up. Adapted from Kasap.

measure of the ability for the semiconductor to absorb photons within a given unit of length and has units of 1/m. Then, we can write a differential equation such that:

$$\delta I = -\alpha I(x)\delta x \tag{1.24}$$

where the change in intensity δI is related to the absorption coefficient, the current intensity, and the distance traveled into the semiconductor δx. This equation can be solved to yield the Beer-Lambert law, in which the intensity varies within the semiconductor as:

$$I(x) = I_0 \exp(-\alpha x) \tag{1.25}$$

such that the intensity falls off exponentially within the semiconductor. This is represented in figure **??**: Defining the absorption coefficient is captured in the semiconductor depends quite a bit on the band structure of the material. As we have articulated already, the absorption of a photon can only happen if the energy

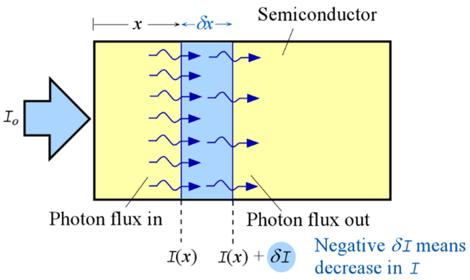


Figure 1.9: Relevant description of the Beer-Lambert equation for absorption, in which absorbed photons within a small differential length δx decreases the photon flux exiting the region. Adapted from Kasap.

of the photon is greater than the band gap. In addition, we require that the there is a state in the conduction band that the electron can be excited into. In all physical processes, we must conserve both energy and momentum. Here, we may note that the momentum of a photo $\hbar k_{photon} = h/\lambda$ is rather small compared to the Brillouin Zone of a crystalline material. Therefore, most diagrams show a vertical line. Regardless, we may additionally note that there is a greater density of states the farther away from the band edges. This suggests that the absorption coefficient will change as a function of photon energy, as there are more states that the electron can be excited into. This is shown in figure 1.10:

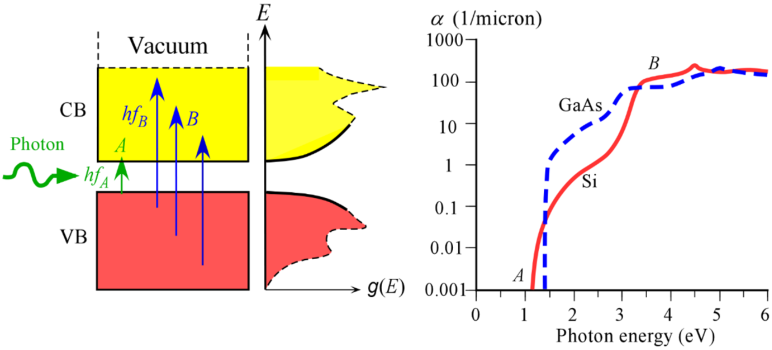


Figure 1.10: Optical excitation of an electron from the valence band to the conduction band requires an open state in the conduction band. The density of states changes within the valence and conduction bands, such that the absorption coefficient is a strong function of energy. Representative measured absorption coefficients for Silicon and GaAs as a function of photon energy are presented. One may notice that the absorption coefficient is essentially 0 below the band gap.

One may notice in figure 1.10 that the slope of the absorption coefficient is very strong in these semiconductors. You may also notice that Silicon has a lower slope thatn GaAs. This is because Silicon is an indirect bandgap semiconductor, which means that the energy maximum of the valence band is not at the same momentum value as the minimum of the valence band edge. As we have articulated, the absorption of a photon can be written down on a band diagram as a nearly vertical line. This is represented in figure 1.11:

This means that an electron excited by a photon at the band gap energy of an indirect band gap semiconductor will not have a state that they can occupy *unless* a phonon can mediate the momentum difference. This is obviously more challenging than direct absorption, so the absorption coefficient does not scale as strongly. once the energy is high enough that there is a state for direct excitation, the absorption coefficient increases dramatically. Conversely, GaAs is a direct bandgap semiconductor, meaning that it can directly excite an electron into the conduction band, which explains why it has a much steeper slope in the absorption coefficient.

Finally, we should note that there must be mechanisms for electrons and holes to relax back to their equilibrium concentration. When an electron and hole are near each other within the crystal, they can recombine with each other, reducing the

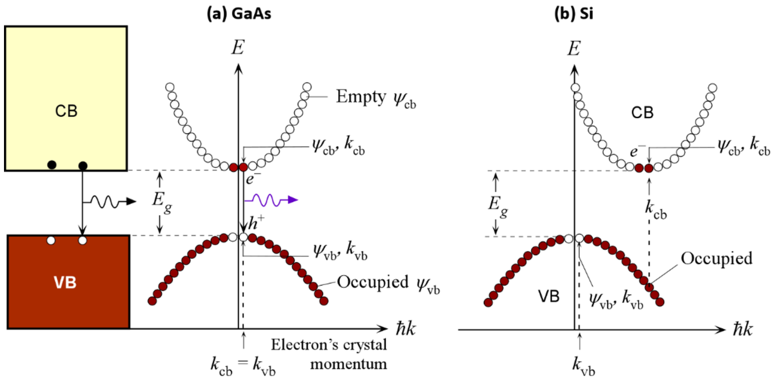


Figure 1.11: Direct bandgap semiconductors can much more easily excite carriers near the band gap energy because of the alignment of the valence band minimum and conduction band maximum. Conversely, an indirect bandgap semiconductor will require additional help thorugh a phonon to facilitate both energy and momentum conservation.

concentration of electrons and holes. This means that, once the light is turned on, there will be some time before a new steady-state carrier concentration is achieved. Once the light is turned off, there will be some time before recombination leads to the thermal equilibrium value of the carrier concentration. This is represented in figure 1.12:

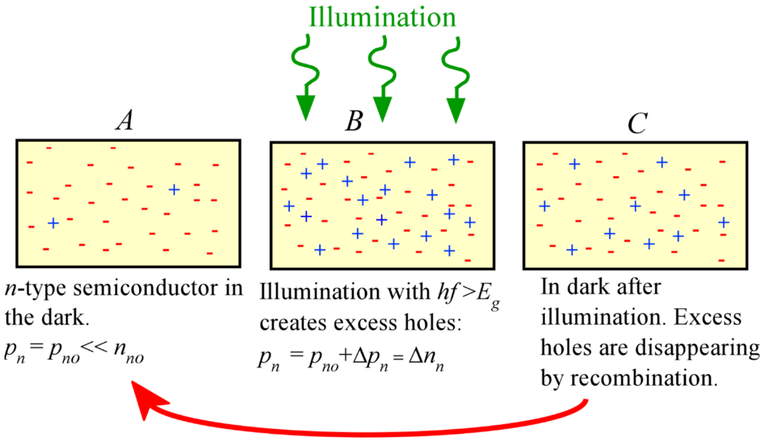


Figure 1.12: Temporal dynamics of optical excitation. In the dark, an n-type semiconductor has an equilibrium concentration of electrons and holes. When light shines on the semiconductor and is absorbed, there are excess carriers generated and a new steady state is formed. Once the light is turned off, recombination occurs to lower the carrier concentration back to its equilbrium value.

The excess concentration of electrons and holes can be represented through the following equations:

$$p_n = p_{no} + \Delta p_n \tag{1.26}$$

and

$$n_n = n_{no} + \Delta n_n \approx n_n \tag{1.27}$$

where

$$\Delta p_n = \Delta n_n \tag{1.28}$$

Here, we use the convention where the subscript refers to the type of semiconductor the charge carrier is within - that is, p_n is the concentration of holes within an n-type semiconductor, n_n is the concentration of electrons within an n-type semiconductor, and p_{no} is the initial concentration of holes within an n-type semiconductor, etc. We see that optical excitation leads to an equal excess concentration of electrons Δn_n and holes Δp_n. Additionally, we have assumed that the light absorption is not so strong such that the concentration of electrons is nominally unchanged (the majority charge carrier), while the concentration of holes is increased substantially.

1.6 Photoconductivity

Let's use this phenomena to our advantage. We have already identified that the conductivity of a semiconductor is strongly dependent on the carrier concentration. We can use the ability to create more carriers through optical absorption to make detectors for light. We will identify a thought problem to understand this, where we connect the semiconductor to a hypothetical circuit to measure the current or conductivity. We are omitting any impact that the metal contact has on the overall circuit. We additionally assume that all light is transmitted into the semiconductor, all light is absorbed, the material properties like τ_h, μ_e, μ_h, etc. do not change with carrier concentration, and that contacts do not block the flow of carriers. This is represented in figure 1.13, with the net result included:

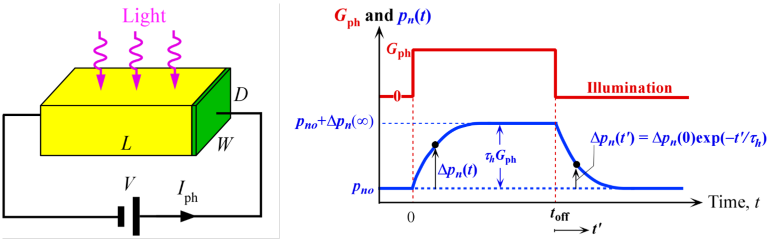


Figure 1.13: Schematic of photoconductivity experiment, in which an n-type semiconductor with depth D, width W, and length L is connected to a circuit. Light shines on the material and is absorbed. The response of the photodetector is included, which as a rise time and a fall time.

We can define the *photoconductivity* σ_{ph} as the change in the conductivity $\Delta\sigma$ of the semiconductor when the light is on (σ_{light}) vs off (σ_{dark}) as:

$$\Delta\sigma = \sigma_{ph} = \sigma_{light} - \sigma_{dark} \tag{1.29}$$

We can expand this to include all of the relevant terms as:

$$\Delta\sigma = en_n\mu_e + ep_n\mu_h - (en_{no}\mu_e + ep_{no}\mu_h) \tag{1.30}$$

Where the first term is the conductivity when the light is on (taking into account the contributions from both the electrons and holes), while the second term is the conductivity in the dark (and therefore uses the initial concentrations within the semiconductor).

We can determine what the conductivity is once we understand the generation and recombination of carriers. Here, we can define the *generation rate* G_{ph} of electron hole pairs as the number of electron hole pairs that are generated due to the flux of photons across the area of the semiconductor, divided by the volume to get a concentration per unit time. We additionally include a quantum efficiency factor η - for a perfectly operating device with every single absorbed photon creating an electron-hole pair, $\eta = 1$. We can express this as:

$$G_{ph} = \frac{\eta A \Gamma_{ph}}{AD} = \frac{\eta \left(\frac{I}{h\nu}\right)}{D} = \frac{\eta I \lambda}{hcD} \tag{1.31}$$

where we have used the geometry of the detector as defined in figure 1.13. In addition, we must write down what the recombination rate R of electrons and holes are. This is defined as:

$$R = \frac{\Delta p_n}{\tau_h} \tag{1.32}$$

where we have included a lifetime τ_h to describe the rate at which these electron hole pairs will recombine. At steady state, the rate of generation must match that of recombination. Thus, we have:

$$\frac{d\Delta p_n}{dt} = G_p h - \frac{\Delta p_n}{\tau_h} = 0 \tag{1.33}$$

such that we can solve for the steady state excess concentration of holes as:

$$\Delta p_n = G_{ph}\tau_h = \frac{\tau_h \eta I \lambda}{hcD} \tag{1.34}$$

And therefore the photoconductivity is:

$$\Delta\sigma = \frac{e\eta I \lambda \tau_h (\mu_e + \mu_h)}{hcD} \tag{1.35}$$

We can additionally solve these equations to determine the temporal profile. First, let us look at a scenario in which the light is turned on at time $t = 0$ and we want to determine the change in carrier concentration as a function of time $\Delta p_n(t)$. To do this, we write down the differential equation again:

$$\frac{d\Delta p_n(t)}{dt} = G_{ph} - \frac{\Delta p_n(t)}{\tau_h} \tag{1.36}$$

We can separate variables and integrate to a given time t such that:

$$\ln\left(G_{ph} - \frac{\Delta p_n(t)}{\tau_h}\right) = -\frac{t}{\tau_h} + C \tag{1.37}$$

We know that at time $t = 0$ there is no excess carrier concentration such that $\Delta p_n(t = 0) = 0$. We can therefore solve for the integration constant to find that $C = \ln(G_{ph})$ and therefore the excess hole concentration has the temporal profile:

$$\Delta p_n(t) = \tau_h G_{ph}\left(1 - \exp\left(-\frac{t}{\tau_h}\right)\right) \tag{1.38}$$

Therefore, there is a characteristic time to achieve the steady state concentration difference that depends on the lifetime τ_h

We can do the same thing at the time $t' = t - t_{off}$ that the light source is turned off to look timescale for the carrier concentration to return to its equilibrium value in the dark. This means that the same differential equation applies, but G_{ph} is defined to be 0. Then, we have:

$$\frac{d\Delta p_n(t')}{dt'} = -\frac{\Delta p_n(t')}{\tau_h} \tag{1.39}$$

This can go through the same process to show that:

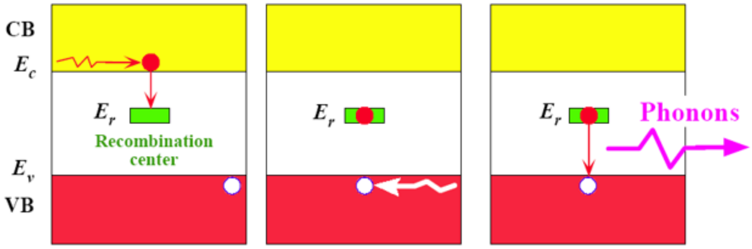
$$\ln(\Delta p_n(t')) = -\frac{t'}{\tau_h} + C \tag{1.40}$$

At time $t' = 0$, we have set that the steady state carrier concentration excess is $\Delta p_n(t' = 0) = \tau_h G_{ph}$ such that $C = \tau_h G_{ph}$. Therefore, the fall time of the detector can be determined from:

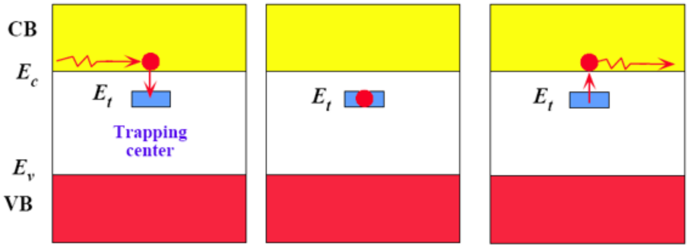
$$\Delta p_n(t') = \tau_h G_{ph} \exp\left(-\frac{t'}{\tau_h}\right) \tag{1.41}$$

which has the same characteristic timescale as the photoresponse, as expected.

Finally, we note that there are a few non-idealities that we need to be aware of. First, there are a variety of ways for electron hole pairs to recombine. In imperfect semiconductors, a variety of defects from impurities, dislocations within the crystal, grain boundaries, and other defects can lead to energy levels within the band gap of the semiconductor. These can lead to unwanted recombination pathways that lead to additional issues when creating a semiconductor device. Free electrons in the conduction band can end up lowering their energy and occupy this energy state (known as a recombination center), before combining with a hole to recombine through the emission of a phonon (lattice vibration, heat). This is an unwanted pathway that depends on crystal purity and processing. In addition, somewhat shallow defect states within the band gap can lead to an effect called trapping, in which a free electron in the conduction band will lower its energy through the emission of a phonon and become localized on this trapping center. It may not move for a significant amount of time before thermal energy can re-excite the electron into the conduction band. This localized charge state will “trap” the carrier and remove its ability to be conducted. This can be a killer of optoelectronic devices - you have gone through all of the effort to find ways to effectively create free carrier electrons in the conduction band only for unwanted defects or impurities to trap them at a defect location and prevent them from moving freely. This can be a limiting factor in the design of LEDs, for example. The various phenomena discussed here are shown in figure 1.14



(a) Recombination



(b) Trapping

Figure 1.14: Recombination and trapping within a semiconductor. Unwanted defect states with energy levels within the band gap can lead to unwanted recombination pathways through the emission of phonons. In addition, sites can trap electrons for a period of time before they can be re-excited into the conduction band, which eliminates their ability to conduct through the lattice.

1.7 Drift and Diffusion currents

So far in this course, we have discussed the ability for the carrier concentration to be modified through doping, temperature, and the ability to create additional carriers through optical absorption (of light above the semiconductor's band gap). However, we need to understand the factors that lead to charge being inserted and extracted from materials to leverage these effects.

To compute the current that can come from a material, we need to articulate the different methods for charge to flow in a material. For a semiconductor, both electrons and holes are charge-carrying particles that can be used for conduction. If we know the number of particles entering/exiting the material (or a plane within the material), we can know what the current is across that boundary.

1.8 Diffusion current

Previous lectures have discussed the analogy between electron/hole movement in a solid material and the movement of gas molecules in a volume (hence some models referring to a "Free electron gas"). Essentially, we have developed up to this point the notion that the particles can be thought of as having an effective mass m_e^* or m_h^*, which is modified from a free electron/hole mass through the band structure of the material. These electrons and holes can freely move about the crystalline lattice with a mobility μ_e or μ_h that is determined by the frequency these particles scatter off of impurities or lattice sites. There is a strong analogy between the flow of electrons in this scenario and aspects of diffusion. If there is a higher concentration of charges in one part of the semiconductor, we would expect this free movement to lead to a diffusion of charge through the crystal.

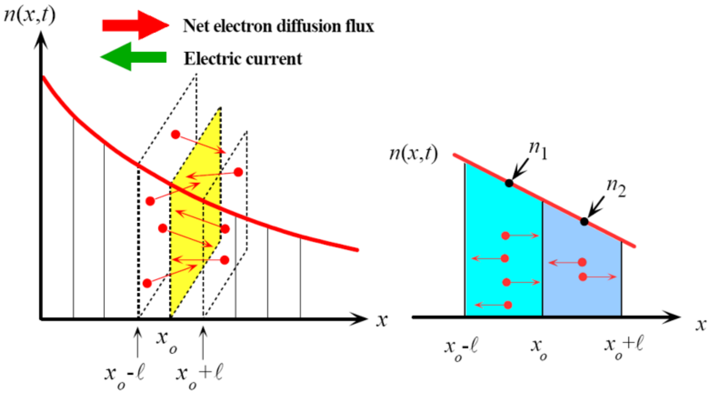
To develop this, let us first quantify the *particle flux* for electrons and holes. If we imagine a plane within the material with an area A, and can count the net number of particles ΔN that pass through the plan per unit of time Δt, then we can define a particle flux as:

$$\Gamma = \frac{\Delta N}{A \Delta t} \tag{1.42}$$

Here, the particle flux will have units of $\#/(m^2 t)$. Noting that each electron or hole has a charge Q, we can then define the current density J as:

$$J = Q\Gamma [=] A/m^2 \tag{1.43}$$

We will first try to understand the current density of electrons when there is a difference in electron concentration, and finish this section by understanding the difference between this *diffusion* current and *drift* current. Here, consider a concentration profile of electron density as indicated by figure 1.15 below:



(a) (b)

Figure 1.15: a) Nominal electron concentration profile in one dimension. y-axis is electron concentration at a given moment in time, while x axis is position within semiconductor. Drawing planes in/out of page can therefore define the flux of electrons at position x. b) The flux of electrons Γ_e is approximated by taking the average concentration within regions to the left and right of the plane of interest at x_0. The net flux going in the $+x$ direction is given by the difference of flux from either direction. Figure adapted from Kasap 4th edition

Here, we see that the concentration $n(x)$ of electrons decreases as a function of distance x. To understand the current density, we need to define the particle flux.

Let us imagine a series of planes coming in an out of the page that hae an area A. We can place these along the x axis with some spacing determined by some length l, which is approximately the distance an electron will move before scattering. The electrons at each position x can freely move, and will move randomly in all directions. In general, they will move in a straight line for some amount of time τ until they scatter off of a lattice site, a dopant, dislocation or other defect, etc.

In this first thought problem, we assume that they have equal probability of moving either to the left or the right. If they are moving either to the left or to the right, then we can compute the particle flux through a plane at position x_0. Here, we approximate the flux as the number of electrons from the plane at position $x_0 - l$ that move to the right minus the number of electrons from the plane at $x_0 + l$ that move to the left. The number then describes the total flux of electrons that are moving in the positive x direction.

Since each electron has an equal probability of moving to the left or right, the net number of electrons from each plane that move in the appropriate direction is half of the total electron concentration at that plane. Mathematically, this means that the flux is defined as:

$$\Gamma_e = \frac{\frac{1}{2}n_1 A l}{\tau} - \frac{\frac{1}{2}n_2 A l}{\tau} \tag{1.44}$$

Where the first term is the flux of electrons moving right from the plane at $x_0 - l$ and the second term is the nubmer of electrons moving to the left from the plane at $x_0 + l$. Simplifying, we find that:

$$\Gamma_e = \frac{\frac{1}{2}n_1 A l = \frac{1}{2}n_2 A l}{\tau} = -\frac{l}{2\tau}(n_2 - n_1) \tag{1.45}$$

In this simple picture, we can approximate the difference in concentrations with a differential such that:

$$n_2 - n_1 \approx \left(\frac{dn}{dx}\Delta x\right) = \frac{dn}{dx} l \tag{1.46}$$

Which immediately suggests that:

$$\Gamma_e = -\frac{l^2}{2\tau}\left(\frac{dn}{dx}\right) = -D_e \frac{dn}{dx} \tag{1.47}$$

Here, we see a powerful statement: From a simple picture regarding the movement of electrons, we can relate the distance an electron can move in a given time frame to a parameter we call the *Diffusion coefficient*. The net flux of charge carriers is therefore directly related to the concentration difference within the material, with a coefficient that describes how straightforward or easy it is for the charge carrier to move through the lattice. This also has a functional form essentially identical to Fick's first law for diffusion, and represents the connection between particle flux and the relevant driving force. This is similar to the conductivity σ of a material, which connects the current density J to the applied electric field E as:

$$J = \sigma E = -\sigma\left(\frac{dV}{dx}\right) \tag{1.48}$$

Because this equation is related to the ability for these charge carriers to freely move or diffuse through a solid, we call the current density arising from this the *diffusion current density.*

We note two things in this section. First, this derivation applies to holes as well, which is to say that we can relate the current density of holes ($J_{D,h}$ to a diffusion coefficient (D_h) and the concentration gradient. Putting everything together, we can define the current density from this effect as below:

$$J_{D,e} = -e\Gamma_e = eD_e\frac{dn}{dx} \tag{1.49}$$

and

$$J_{D,h} = e\Gamma_h = -eD_h\frac{dp}{dx} \tag{1.50}$$

Note that the difference in sign between the above equations is related to the sign of the charge - the hole has a positive charge while the electron has a negative charge.

The second thing we should note about this is one small detail regarding the derivation. Here, we assumed that all electrons (or holed) would move in the x direction for a given time τ to move a distance l within the crystal to “derive” the diffusion coefficient. This is not the case in real materials, as electrons will have an ensemble average scattering time of τ and can move in all directions. While it is true that half of the electrons will move with the x component of their velocity in the positive or negative directions, the velocity of each electron may have components in y or z. A more complete derivation of this can be done using statistical mechanics. The end result is that the scaling relationship l^2/τ still holds, but the factor of 1/2 disappears. That is to say, the “real” diffusion coefficient is:

$$D_e = l^2/\tau \tag{1.51}$$

All of the important physics is still there, but the absolute value is different by a factor of two.

1.9 Drift current

Consider the following situation in figure 1.16, in which a semiconductor has light shining on it to generate additional carriers. In addition to this, the semiconductor is connected to a circuit that enables a voltage (and therefore an electric field) to be applied. In the absence of an electric field, we may expect a concentration gradient to be created, as the intensity within the semiconductor will decrease as more light is absorbed within the semiconductor. Once we apply an electric field, the electrons and holes may move in response to this electric field. The current that is induced by this field is termed the *drift current density*. The current density is related to the electric field through the conductivity as:

$$J_{drift,e} = \sigma E \tag{1.52}$$

Inserting the expression for conductivity of electrons, we have:

$$J_{drift,e} = en\mu_e E_x \tag{1.53}$$

Therefore, the total current density for electrons (J_e) is given by the sum:

$$J_e = en\mu_e E_x + eD_e\frac{dn}{dx} \tag{1.54}$$

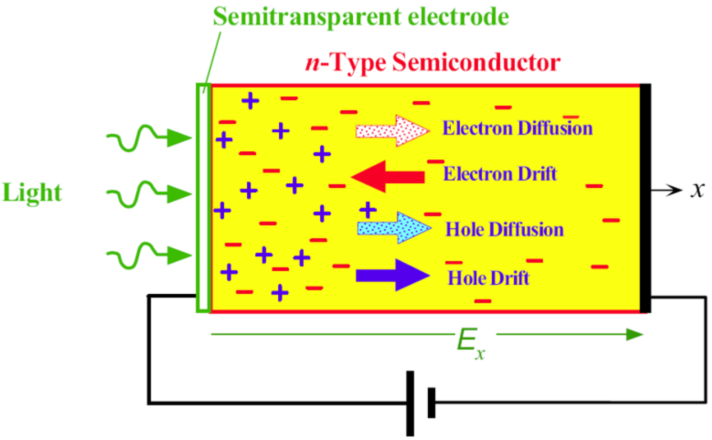


Figure 1.16: A more general case of charge in a uniformly doped semiconductor. Here, light is shining onto the semiconductor to create electron hole pairs which may have a concentration that depends on depth into the semiconductor. There are additionally electrodes placed on the semiconductor such that a voltage can be applied to collect a current. The voltage induces an electric field within the semiconductor that creates a net drift current density. Here we are neglecting any interaction between the metal and the electrode. Figure adapted from Kasap 4th edition.

And the equivalent expression for holes (J_h) would be:

$$J_h = ep\mu_h E_x - eD_h \frac{dp}{dx} \tag{1.55}$$

Where the hole concentration gradient is given by the derivative on the right hand side.

If we assume that $D_e = \frac{l^2}{\tau}$, we can determine an additional relationship for the diffusion coefficient that depends on temperature. It may stand to reason that the ability for an electron to diffuse within a material is related to temperature, as it is the energy source for the electron to move. We can estimate the length l based on the velocity of the electron v_x through:

$$l_x \approx v_x \tau \tag{1.56}$$

which is just a re-statement of the distance a particle can move in a particular direction given its velocity. If this is the case, then the diffusion coefficient is given as:

$$D_e = \frac{v_x^2 \tau^2}{\tau} = v_x^2 \tau \tag{1.57}$$

We can relate this velocity to the particle's kinetic energy. The average kinetic energy for the particle moving in this direction is $1/2k_BT$, such that:

$$\frac{1}{2} m_e^* v_x^2 = \frac{1}{2} k_B T \tag{1.58}$$

which implies that

$$v_x^2 = k_B T / m_e^* \tag{1.59}$$

Inserting into the expression for the diffusion coefficient, we have:

$$D_e = \frac{k_B T \tau}{m_e^*} \tag{1.60}$$

We may recognize that there are terms that can be grouped together into another measurable quantity, the mobility $\mu_e = e\tau/m_e^*$. We can insert the mobility in the expression to arrive at:

$$D_e = \frac{k_B T}{e}\frac{e\tau}{m_e^*} = \frac{k_B T}{e}\mu_e \tag{1.61}$$

Where we have added a factor of e/e and grouped terms to substitute in the mobility. This is a poewrful expression known as the *Einstein Relationship*, which has an equivalent description for holes:

$$\frac{D_e}{\mu_e} = \frac{k_B T}{e} \tag{1.62}$$

and

$$\frac{D_h}{\mu_h} = \frac{k_B T}{e} \tag{1.63}$$

This essentially says that the diffusivity and mobility of charge carriers are directly related to the thermal energy and charge of the particle.

1.10 Inhomogeneous doping concentrations

So far we have considered the simplest case, in which the dopant density in a semiconductor is uniform throughout. This is not the case in general. For example, one of the simplest ways to dope a semiconductor (and one that is very relevant to modern microelectronics fabrication) is the start with a nearly intrinsic wafer of semiconductor (say, Silicon). Then, the wafer may be heated in a chamber that contains a gas with the desired dopant. For example, AsH_3 can be used to dope the wafer with Arsenic to make the wafer n-type. This is a classic diffusion problem such that the doping concentration at a specified depth can be determined based on the temperature, time, and concentration of reactant gas at the surface. However, any solution to the diffusion equation with these boundary conditions will lead to a doping concentration that is not linear with depth. This is schematically represented in figure 1.17, where we take a simple approximation that the As concentration has an exponential dependence within the semiconductor:

If this is the case, then we may expect a few things to happen. First, we would expect (at room temperature) that all charge carriers are activated, meaning that all of the As dopants will have donated an electron to the semiconductor to provide an electron in the conduction band, leaving behind a charged $As+$ core in the lattice. This will lead to a density of conduction electrons $n(x)$ that is generally dependent on the position within the crystal.

As we discussed in the previous section, this concentration distribution has a non-zero gradient $\frac{dn}{dx}$, which means that we should expect to see a diffusion current $J_{D,e}$ of electrons. However, the movement of electrons will leave behind core $As+$ atoms in the lattice, which cannot easily move at room temperature. The distance

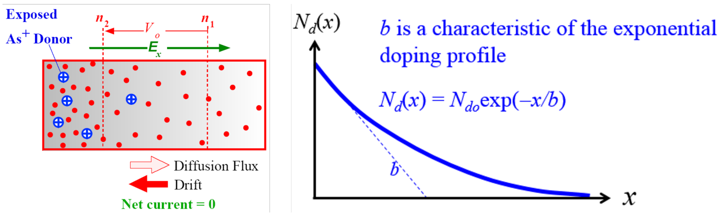


Figure 1.17: An *As* doped silicon wafer may have an inhomogeneous doping concentration. The concentration gradient may be approximated as a decaying exponential away from the surface. The concentration gradient of donors leads to a gradient of electron density, which will undergo a diffusion process. This is resisted by the built-in field generated by the dipole between the stationary donors and the mobile electrons, such that drift and diffusion balance each other.

between the charged $As+$ atoms and negatively charged $e-$ electrons will therefore induce an electric field E, which creates a *Built in Voltage*. This electric field should induce a current density $J_{drift,e}$ that opposes the diffusion current $J_{D,e}$. These will balance each other at equilibrium, leading to a permanent built in field.

At equilibrium, we expect no net current in this structure (essentially, we are at open circuit and no current is entering or leaving). We can write the equation down for the current density as:

$$J_e = en\mu_e E_x + eD_e\frac{dn}{dx} \tag{1.64}$$

Here, E_x is related to the built in voltage by $E_x = -\frac{dV}{dx}$. We can insert this into the expression for the current density and set the different terms equal to each other to arrive at:

$$J_e = -en\mu_e\frac{dV}{dx} + eD_e\frac{dn}{dx} \tag{1.65}$$

such that

$$en\mu_e\frac{dV}{dx} = eD_e\frac{dn}{dx} \tag{1.66}$$

We can simplify this equation by removing the electron charge from both sides and dividing the mobility over. Using the Einstein relationship this then becomes:

$$n\frac{dV}{dx} = \frac{k_BT}{e}\frac{dn}{dx} \tag{1.67}$$

Separating variables, eliminating the dx terms, and integrating between points 1 and 2 (i.e. the region with voltage 1 to voltage 2 that has corresponding electron densities n_1 and n_2, we find the equation:

$$\int_{V_1}^{V_2} dV = \frac{k_BT}{e}\int_{n_1}^{n_2}\frac{dn}{n} \tag{1.68}$$

which has the solution:

$$V_2 - V_1 = \frac{k_BT}{e}\ln\left(\frac{n_2}{n_1}\right) \tag{1.69}$$

We can fully solve for this if we know exactly what the electron density is at each position. A reasonable assumption is to say that the electron concentration has not changed dramatically from the density of donors. As a first approximation we therefore assume that:

$$n(x) \approx N_d(x) \tag{1.70}$$

Here, we will use a simplistic (but not techically correct) solution to the dopant profile to be:

$$N_d(x) = N_0 \exp\left(-\frac{x}{b}\right) \tag{1.71}$$

This is essentially approximating the diffusion of dopants into silicon to be a decaying exponential with a concentration N_0 at the wafer surface and an empirical parameter b that describes how quickly the exponential decays. If we insert this into the expression for the concentration, we arrive at:

$$V_2 - V_1 = \frac{k_BT}{e} \ln\left(\frac{N_0 \exp\left(-\frac{x_2}{b}\right)}{N_0 \exp\left(-\frac{x_1}{b}\right)}\right) \tag{1.72}$$

Which simplifies to:

$$V_2 - V_1 = \frac{k_BT}{e} \ln\left(\frac{-x_2}{b} + \frac{x_1}{b}\right) \tag{1.73}$$

Noting that $\frac{dV}{dx} \approx \frac{V_2 - V_1}{x_2 - x_1}$, we can re-arrange this expression to arrive at the key conclusion, which is that the built in field for an inhomogeneously doped semiconductor is:

$$E_x = -\frac{dV}{dx} \approx \frac{V_2 - V_1}{x_2 - x_1} = \frac{k_BT}{be} \tag{1.74}$$

Why did we go through all of this effort? This shows that the properties of the semiconductor and its composition profile determines not only the number of free carriers, but additionally leads to dynamics that create electric fields natively in the material. As we've seen in 351-1, the application of an electric field also distorts or shifts the band structure of a metal or semiconductor (and is also essentially the reason for conduction of electrons). This has strong implications for the operation of devices with more complicated doping concentrations, as well as devices with metals, whose fermi energy may not be perfectly aligned with the fermi level of the semiconductor. This effect can be leveraged to create the basic operation of semiconductor switches when we discuss pn junctions.

1.11 The continuity equation

So far, we have discussed methods that can lead to the generation of electrons and holes, their recombination, and the ability to conduct through diffusion and drift. There are many possible ways, therefore, that the electron and hole concentration can change, which can lead to rich dynamics. If you are to design an electronic device that uses these semiconductors, it would therefore be advantageous to have an equation that can predict the movement of electrons and holes as a function of time and throughout the semiconductor. This section will provide such an equation and a simple use case for solving it - As you can imagine, the dynamics can be complex and therefore can be quite challenging to solve for analytically for a general set of boundary conditions. Computational software is therefore quite helpful to solve these equations and could be implemented, for example, in MATLAB or COMSOL.

We will first consider a generic n-type semiconductor as indicated in figure 1.18. We will consider a differential volume within the semiconductor that may have a disturbed concentration of electrons and holes because of drift currents, diffusion, generation of charge carriers, and potential recombination of excess charge carriers. We are assuming in this derivation that the semiconductor is nominally uniform in the y and z directions, but there may be variation in the hole concentration density in the x direction. We assume that the semiconductor has some nominal area A and consider two planes very close to each other at positions x and $x + \delta x$. This essentially describes a differential volume within the material. Light may generate carriers within this volume, charge carriers may recombine, and there may be a flux of carrier density due to a flux into or out of the volume.

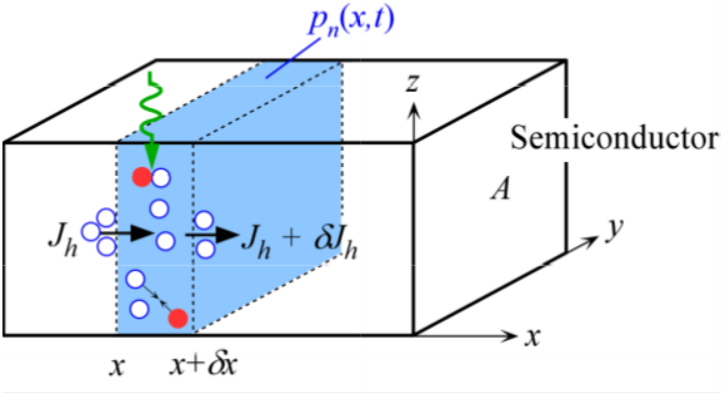


Figure 1.18: A generic n-type semiconductor. We consider the differential volume of this semiconductor between x positions x and $x+\delta x$. The semiconductor is uniform in y and z but may vary in x. The hole concentration of the n-type semiconductor is $p_n(x,t)$. The current density into the plane at x is J_h while the flux out of the plane $x + \delta x$ is $J_h + \delta J_h$. Figure adapted from Kasap.

If there is any generation of charge carriers or otherwise a disturbance of the carrier density away from the equilibrium values for the hole and electron concentrations, we would expect a flow of these species to compensate that will evolve in time. We can quantify this. The shaded area has a volume of $A\delta x$ and a hole concentration $p_n(x,t)$. The current density at x due to flow into the volume is given as J_h, while the flow out of the volume at $x + \delta x$ is defined as $J_h + \delta J_h$. If δJ_h is negative, then the current leaving the volume is lower than the current arriving in it, meaning that the hole concentration should increase. This means that we can define the *rate of increase in hole concentration due to* J_h as:

$$\frac{1}{A\delta x}\left(\frac{-A\delta J_h}{e}\right) \tag{1.75}$$

We additionally expect there to be generation and recombination in this volume. We can write a general rate equation that can describe the spatial and temporal evolution of this effect:

$$\frac{\partial p_n}{\partial t} = -\frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) - \frac{p_n - p_{n0}}{\tau_h} + G_{ph} \tag{1.76}$$

Where the temporal derivative is the time evolution of the hole carrier density at a given position. The first term is the *rate of increase in hole concentration due to* J_h. The second term is the *recombination rate* for the concentration of holes (negative because recombination should reduce the carrier concentration. The last term is the *generation rate* or carrier concentration due to absorption.

We can write a similar equation for electrons, but note that the current density term will have a sign change due to the difference in sign of the charge:

$$\frac{\partial n_n}{\partial t} = \frac{1}{e}\left(\frac{\partial J_e}{\partial x}\right) - \frac{n_n - n_{n0}}{\tau_h} + G_{ph} \tag{1.77}$$

As you can imagine, this can be a tough differential equation to solve in general - and this is only the 1-D case. Depending on boundary conditions, the solution can vary greatly. In general, this governing equation can be used by engineers to design things like doping concentration, illumination conditions, electrode placement and operation, etc. to solve for the dynamics of the relevant charge carriers. We will spend the rest of this section thinking about a simplified version of this equation, which is similar to the equation used for the earlier photoconductivity equation.

1.12 Continuity equation at steady state

We will investigate a simple scenario as depicted by figure 1.19. In this example, light is shining onto an n-type semiconductor and is absorbed within a small layer at the front of the semiconductor. We will assume that all light is absorbed in a distance x_0 into the semiconductor, and use the continuity equation to solve for the steady state characteristics of the semiconductor past this point.

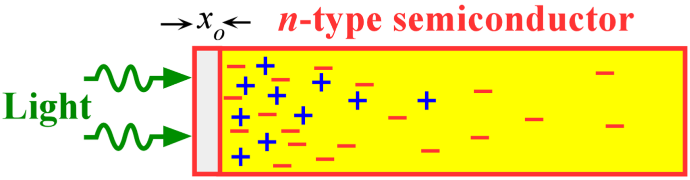


Figure 1.19: A semiconductor with uniform doping concentration has light illuminating it from one side. Assuming that all of the light is absorbed in a small distance, we can determine the steady state currents and carrier concentrations caused by this carrier generation. Figure adapted from Kasap.

At steady state, the temporal derivative in the equation will become 0. By construction, we are looking at the region of the semiconductor after the light absorption region, meaning the generation term will become zero. Under these conditions, the continuity equation for holes becomes:

$$-\frac{p_n - p_{no}}{\tau_h} = \frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) \tag{1.78}$$

If we assume that the electric field within the semiconductor is small, then we can simplify the current density to be $J_h = ep\mu_h E_x - eD_h\frac{dp}{dx} \approx -eD_h\frac{dp}{dx}$. Then, we have:

$$-\frac{p_n - p_{no}}{\tau_h} = \frac{1}{e}\left(\frac{\partial J_h}{\partial x}\right) = \frac{1}{e}eD_h\frac{d^2p_n}{dx^2} \tag{1.79}$$

Recognizing that the top portion of the recombination term $p_n - p_{no} = \Delta p_n$ and that the spatial derivatives of the carrier concentration will have $dp_n/dt = d\Delta p_n/dt$, this equation becomes:

$$D_h\frac{d^2\Delta p_n}{dx} = \frac{\Delta p_n}{\tau_h} \tag{1.80}$$

Using the Einstein relationship we can then obtain:

$$\frac{d^2\Delta p_n}{dx^2} = \frac{\Delta p_n}{L_h^2} \tag{1.81}$$

Where we have define the *diffusion length of holes* to be:

$$L_h = \sqrt{D_h\tau_h} \tag{1.82}$$

In this n-type semiconductor, we assume that $p_{no} << n_{no}$. we will assume that we are in the "weak injection regime", the increase in holes from generation is significantly less than the initial electron concentration - that is: $\Delta p_n << n_{no}$. When carriers are generated at the surface of the semiconductor, we expect a change in carrier concentration at the surface of $\Delta p_n(x = 0)$. This should lead to a diffusion of holes to the right. At the same time, there should be recombination of excess holes with the conduction electrons within the semiconductor, which suggests that the hole concentration should decrease deeper into the semiconductor. Very far away, we should expect that the carrier concentration $p_n = p_{no}$ - that is, the carrier concentration far from the area where light is generating new carriers should be the equilibrium value. The solution to this equation at steady state takes the form:

$$\Delta p_n(x) = \Delta p_n(0)\exp\left(-\frac{x}{L_h}\right) \tag{1.83}$$

This means that there is a *diffusion current* $I_{D,h}$ within the semiconductor due to the charge generation that will have the same spatial dependence. If the semiconductor has an area A, then:

$$I_h \approx I_D, h = -AeD_h\frac{dp_n(x)}{dx} = \frac{AeD_h}{L_h}\Delta p_n(0)\exp\left(-\frac{x}{L_h}\right) \tag{1.84}$$

At steady state, we expect that the holes generated will be conducted away by this hole current such that the carrier concentration at the surface will remain constant. This means that:

$$Ax_0 G_p h = \frac{1}{e} I_{D,h(0)} = \frac{AD_h}{L_h} \Delta p_n(0) \tag{1.85}$$

Where the left term is the total carrier generation rate (volume * generation of carrier density) is equated to the hole current at $x = 0$. Inserting this into the above equations can enable us to solve for $\Delta p_n(x = 0)$ as:

$$\Delta p_n(x = 0) = x_0 G_p h \left(\frac{\tau_h}{D_h} \right)^{1/2} \tag{1.86}$$

We have solved for the carrier concentration and relevant currents for the holes, but we must also do this for electrons. The generated electrons will also diffuse towards the bulk of the material, but in general the diffusion coefficients and diffusion lengths for electrons D_e and L_e will be different than for that of holes. But, the same functional form should still hold:

$$\Delta n_n(x) = \Delta n_n(0) \exp\left(-\frac{x}{L_e} \right) \tag{1.87}$$

Then, we can determine the current of electrons similarly to be:

$$I_{D,e} = AeD_e \frac{dn_n(x)}{dx} = -\frac{AeD_e}{L_e} \Delta n(0) \exp\left(-\frac{x}{L_e} \right) \tag{1.88}$$

The field at the surface must be zero at steady state, such that the generated electrons within the surface region of the semiconductor must additionally be removed by the electron current. Then, we expect a similar equation for electrons as the equivalent for holes:

$$\Delta n_n(0) = x_0 G_{ph} \left(\frac{\tau_h}{D_e} \right)^{1/2} \tag{1.89}$$

which gives the ratio:

$$\frac{\Delta p_n(0)}{\Delta n_n(0)} = \left(\frac{D_e}{D_h} \right)^{1/2} \tag{1.90}$$

If the semiconductor is at open circuit, then the total current within the device should become 0. That is:

$$I = I_{D,h} + I_{D,e} + I_{drift,e} = 0 \tag{1.91}$$

Here, we assume the drift current for holes is very small because of the weak injection limit (the carrier concentration is much lower than the electron concentration). Then, we can solve for the drift current of electrons to be:

$$I_{drift,e} = -I_{D,h} - I_{D,e} \approx Aen_{no}\mu_e E \tag{1.92}$$

such that the electric field E is:

$$E = \frac{I_{drift,e}}{Aen_{no}\mu_e} \tag{1.93}$$

and

$$I_{drift,h} = Ae\mu_h p_n(x)E << I_{drift,e} \tag{1.94}$$

As expected from the weak injection limit.

The relevant results from this section are summarized in the plot below (figure 1.20), which shows the spatial distribution of the components we have solved for above. This exercise should demonstrate both the rich dynamics and physics involved in these devices, as well as the complexities even for this relatively simple set of boundary conditions. If we were not yet at steady state, we could in principle determine the way in which the carrier concentrations evolve in space from the moment light shines onto the semiconductor until the the time it reaches steady state; that is presently outside of the scope of this section of the textbook.

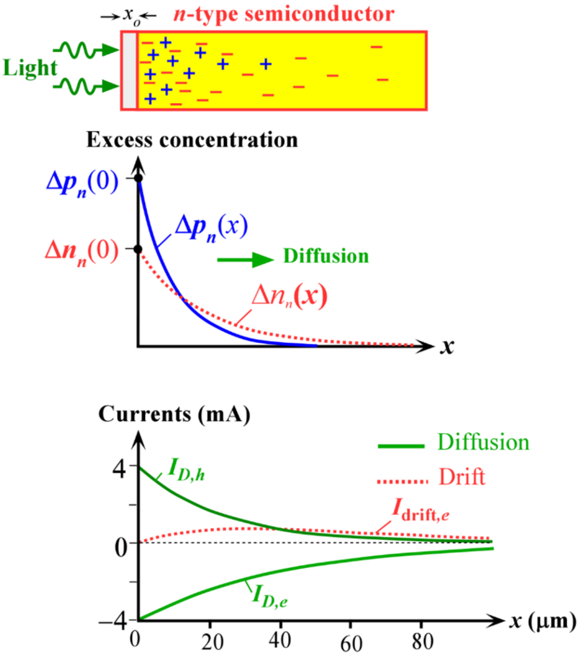


Figure 1.20: The relevant solutions for the changing carrier concentrations $\Delta p_n(x)$, $\Delta n_n(s)$, as well as associated drift and diffusion currents caused by the absorption of light at on end of a very long semiconductor. Figure adapted from Kasap.

1.13 Metal-semiconductor contacts

In this last section regarding semiconductor physics, we will consider what happens when we put a metal in contact (forming a “junction”) with a metal. This is critically important for any device one might want to make, as we need conductors to extract or inject charge or electric fields into our devices. From our discussion of the band theory of solids, we recognize that the energy levels of an electron with respect to the vacuum level are dependent on the energy states within the solid. The Fermi level helps us describe the energy level where we would expect, on average, a 50% chance of a state being occupied by an electron. However, we can also consider the Fermi level as the *chemical potential* for the electrons.

We expect that electrons will attempt to reduce their energy by finding energy levels with lower energy to occupy - a colloquiual phrase is that electrons want to “fall down” to lower energy. Conversely, a hole wants to “bubble up” to higher energy levels in the band. For a semiconductor and metal, we find that the thermal generation term G_0 provides a driving force to excite electrons to higher energy levels (through thermal heat providing this energy), while the recombination rate balances this to arrive at the equilibrium concentration. However, if we have a junction between two dissimilar materials, the relative position of the Fermi level will lead to electrons and holes to re-arrange themselves to minimize their overall energy. This is manifested first in a metal-semiconductor junction, while the next section will discuss this in the context of two regions of semiconductors that have different majority carriers (i.e. a p-type semiconductor in contact with an n-type semiconductor).

In general, the work function Φ of a material is defined to be the energy difference between the vacuum level for a free electron and the Fermi energy of the material at 0 Kelvin, i.e.

$$\Phi = E_{vac} - E_F \tag{1.95}$$

Recall that this fermi level for a semiconductor is dependent on the doping concentration, as you may expect an increase in electrons to increase the Fermi level. Additionally recall that the work function of semiconductor is related to the fermi energy and the electron affinity through:

$$\Phi_n = \chi + (E_c - E_{Fn}) \tag{1.96}$$

where the last term may be related to the electron concentration as:

$$n = N_d = N_C \exp\left(-\frac{E_C - E_{fn}}{k_B T}\right) \tag{1.97}$$

We can broadly categorize different metal-semiconductor junctions into two categories: *Schottky* and *Ohmic*. We will start this section with Schottky contacts and finish with Ohmic contacts.

1.14 Schottky contacts

For an n-type semiconductor, the definition of a Schottky contact is that:

$$\Phi_m > \Phi_n \tag{1.98}$$

where Φ_n is the work function of an n-type semiconductor. In this case, we expect that the Fermi level of the electrons in the *semiconductor* is higher than the Fermi level of the metal, meaning that the metal has unoccupied energy states that have lower energy than the semiconductor. Upon making contact, we would then expect that the system will reduce its overall energy if electrons tunnel from the semiconductor to the metal, and then thermalize their electrons to the lowest available state. This is represented in figure 1.21.

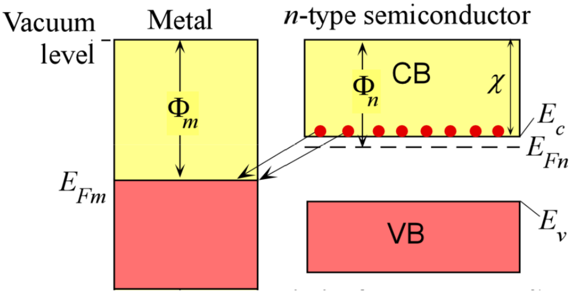


Figure 1.21: A Schottky junction between a metal and semiconductor can form when the work function of the metal is greater than the semiconductor. Electrons in the semiconductor will reduce their energy by finding energy states within the metal at lower energy. Figure adapted from Kasap, fig. 5.39.

This will locally *deplete* the semiconductor of electrons, forming a *depletion region* within the semiconductor. The remaining dopant atom cores leads to a built in field that eventually prevents additional electrons from moving into the metal.

We discussed that the fermi level represents the chemical potential of the system. This means that, at equilibrium, the fermi level of the metal and semiconductor must be the same. If they were not, then work could be done while electrons move from a region of higher potential to lower potential. This requires that $E_F m$ lines up with $E_F n$. However, the depletion region has a lower concentration of conduction electrons electrons tunelling into the metal. We see in the above equation that the carrier concentration n is related to the difference in energy between the conduction band minimum and the Fermi level $E_C - E_{Fn}$. Thus, we expect $E_C - E_{Fn}$ to increase as n decreases, which manifests in the bands of the semiconductor *bending up* within the depletion region. This creates an additional barrier to electron tunneling with a barrier height of:

$$\Phi_B = \Phi_m - \chi = eV_0 + (E_C - E_{Fn}), \tag{1.99}$$

which is called the *Schottky barrier height* Φ_B. All of this is represented in figure 1.22:

Finally, note that the built in potential V_0 develops an electric field E_0 within the depletion region. This region is also called a *Space charge layer.*

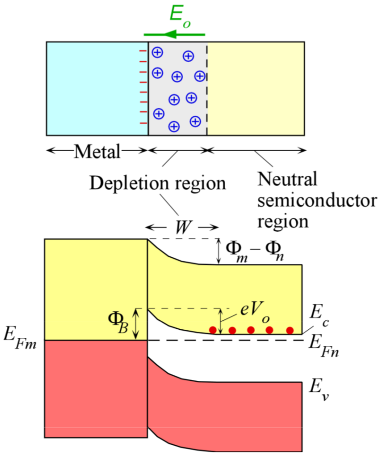


Figure 1.22: At equilibrium, the Fermi levels of the metal and semiconductor are aligned, and the concentration of electrons near the metal lower relative to the rest of the semiconductor, leading to band bending that creates an additional potential barrier for electrons to tunnel into the metal. Figure adapted from Kasap, fig. 5.39.

To understand how electrons can be conducted through this interface, we need to consider hte mechanisms for charge carriers to move around and how they are generated. Without any electric field being applied (i.e. at open circuit), we expect the total current is 0. The two components to current flow across the junction, which are related to electrons moving from the metal to the semiconductor and vice versa. Any electrons in the metal that have enough thermal energy to overcome the Schottky barrier height may flow to the right, which must be balanced by the electrons that have enough thermal energy in the semiconductor to overcome the built-in field. The energy distribution of electrons is related to the Boltzman factor, such that:

$$J_{M->S} = C_1 \exp\left(-\frac{\Phi_B}{k_B T}\right) \tag{1.100}$$

Where $J_{M->S}$ is the current density from the metal to the semiconductor and C_1 is a constant. For the current density from the semiconductor to the metal ($J_{S->M}$, we have

$$J_{S->M} = C_2 \exp\left(-\frac{eV_0}{k_B T}\right) \tag{1.101}$$

This equation is the same, but for the barrier for electrons in semiconductor. At equilibrium these must be equal. To visualize the difference between the two different paths, see figure 1.23.

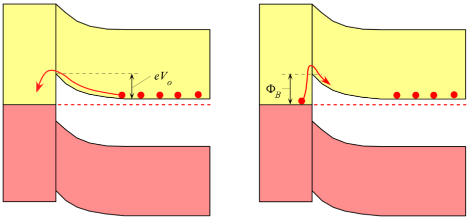


Figure 1.23: The different locations of mobile electrons in the metal and semiconductor lead to different energy barriers to overcome when determining the current density from thermal excitation. For the metal (right), the barrier height set by the metal and semiconductor fermi energies is the barrier to overcome for charge to move into the semiconductor. For the semiconductor (right), the electrons are within the conduction band, meaning that only the induced voltage from the Schottky junction is the barrier to overcome to travel left. At equilibrium, these are equal. Figure adapted from Kasap chapter 5.

When we apply an electric field, things change. We should note that there are now nominally 3 componenets in series in this circuit - the metal, the depletion region, and the semiconductor. The material with the highest resistivity is the depletion region, as the density of charge carriers is the lowest. When an electric field is applied, the majority of the voltage will be dropped across the highest resistance region, which is the depletion region. Because of this, we may expect that the built in voltage V_0 will be modified by the applied voltage V regardless of the direction it is applied in. This will cause different levels of band bending in different directions, which will manifest in a very different Current-Voltage (I-V) curve. The band bending is visualized in figure 1.24 and the ramifications on the electronic transport is explained below.

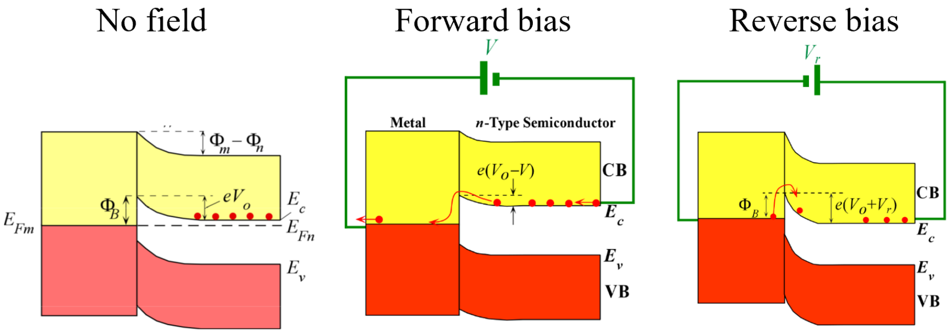


Figure 1.24: Band bending under applied external fields. Without an applied field (left), the Schottky junction has bent bands at the metal-semiconductor interface due to the charge flow from the semiconductor to the metal from the work function difference. Under forward bias (middle), the semiconductor is biased on the negative terminal of a circuit. The majority of the voltage is dropped across the depletion region, which reduces the built in field by the applied voltage. This is manifested in a lower degree of band bending, as the conduction band and valence band are increased in energy due to the potential energy of the applied voltage. This makes the conduction from semiconductor to metal stronger. The conduction in the reverse direction is not affected, as the barrier height has not changed. Under reverse bias (right), the band bending is enhanced as the voltage applied adds to the built in voltage. This leads to rectifying behavior, as the current density from the semiconductor to metal becomes much harder to conduct, and the current is nominally independent of voltage and determined by the current from metal to semiconductor (the barrier height does not change with applied voltage).

We may expect that under *forward bias*, where the semiconductor is connected to the negative terminal of a circuit, that the built-in voltage will therefore be reduced by V such that the new built in voltage is $V_0 - V$. The Schottky barrier does not change. You can think of this as the band structure outside of the depletion region uniformly increasing in energy (relative to everything else) by eV. Now that the built in voltage is reduced, the barrier for electrons from the semiconductor to move into the metal is reduced, but the reverse is not. That means we may expect a non-zero current to flow. The total current density is given as:

$$J = J^{for}_{S->M} - J_{M->S} = C_2 \exp\left(-\frac{e(V_0 - V)}{k_B T}\right) - C_2 \exp\left(-\frac{eV_0}{k_B T}\right) \tag{1.102}$$

where we have used the fact that, before, the electric field is applied, the current density in either direction must be equal. We can simplify this to:

$$J = C_2 \exp\left(-\frac{eV_0}{k_B T}\right)\left(\exp\left(\frac{eV}{k_B T}\right) - 1\right) = J_0\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \approx J_0\left[\exp\left(\frac{eV}{k_B T}\right)\right] \tag{1.103}$$

Where we have grouped like terms and re-written the first term as a constant that depends on the construction of the device. The approximate equation is valid when the exponential term is much larger than 1. Clearly, this is a nonlinear function with an exponential dependence on the applied field.

Conversely, in reverse bias we find that the applied field leads to a voltage that *increases* the built in voltage using the same arguments as before. This essentially further increases the barrier height, meaning that we do not expect similar behavior under a reverse bias. If we perform the same analysis, we find that the reverse current from the semiconductor to the metal $J^{rev}_{S->M}$ is given by:

$$J^{rev}_{S->M} = C_2 \exp\left(-\frac{e(V_0 + V_{rev})}{k_B T}\right) << J_{M->S} \tag{1.104}$$

as the voltage increases. This means that the current is limited to the current flow from the metal into the semiconductor, which is defined entirely by the battier height Φ_B (and therefore nominally independent of voltage once the reverse voltage V_{rev} is large enough that the inequality holds.

Putting this together, this means that under *forward bias*, we expect a Schottky junction to exhibit a current-voltage characteristic (I-V curve) that is exponential with the applied voltage (very different than a standard $V = IR$ dependence. however, if we reverse the voltage, we find that the voltage quickly is limited and becomes a small constant value with increasingly negative voltages applied. This is *substantially* different than the IV characteristic of a metal or insulator, and driven entirely by semiconductor physics! The IV curve for a typical Schottky junction is shown below in figure 1.25. This is sometimes called a *diode*

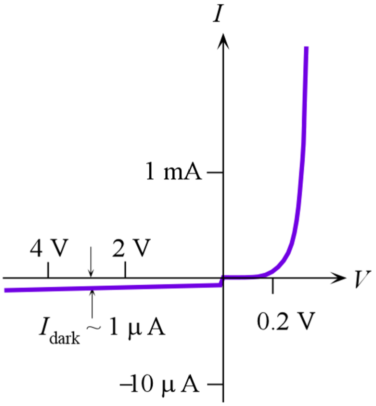


Figure 1.25: Diode response from a Schottky junction. Under forward bias we have an exponential dependence on the current due to the Boltzmann factor for thermal carrier generation and the reduced band bending. Under reverse bias we quickly get to a constant current set by the Schottky barrier height. The completely different behavior under forward and reverse biasing leads to this diode behavior, where current is essentially only allowed to go in one direction. Observe that the scale under reverse bias is scaled differently than forward bias to see the actual current value.

1.15 Application: Solar cell and photodetector

We can use the built in voltage to our advantage in the form of a solar cell or fast photodetector. Here, consider what would happen if light is absorbed in the depletion region. An electron and hole will be generated in a region in which the bands are bending. The applied electric field can act on the additional charges, which are then separated and travel in opposite directions due to their different charges. The electron will travel to the right in the figure, whereas the hole will travel to the left. This leads to an increase in electrons in the semiconductor. Connecting both sides to a load can drive a current to deposit this extra electron into the metal side.

You can also envision this from the earlier definition of how the system will reduce its overall energy. The electron will follow the slope of the conduction band down to a lower energy outside of the depletion region, whereas the hole will travel up the valence band towards the metal. With the appropriate external load, the energy of the photon can be harvested.

We can imagine that the speed with which the charge carriers will separate will depend on the built in voltage or electric field. This means that if we were to reverse bias a schottky junction we can increase this field and therefore speed up the separation process. By measuring this current, we can use this to detect photons - thereby creating a photodetector. These are just a few examples of the power that semiconductor physics can bring to an overall device design. The engineering of Fermi levels, doping concentrations, and metal contacts directly inform the ability to create forces and fields that can control and guide electrons within semiconductor devices that can be manifested in many different ways. This will be explored in more detail when we discuss pn junctions and MOSFETs.

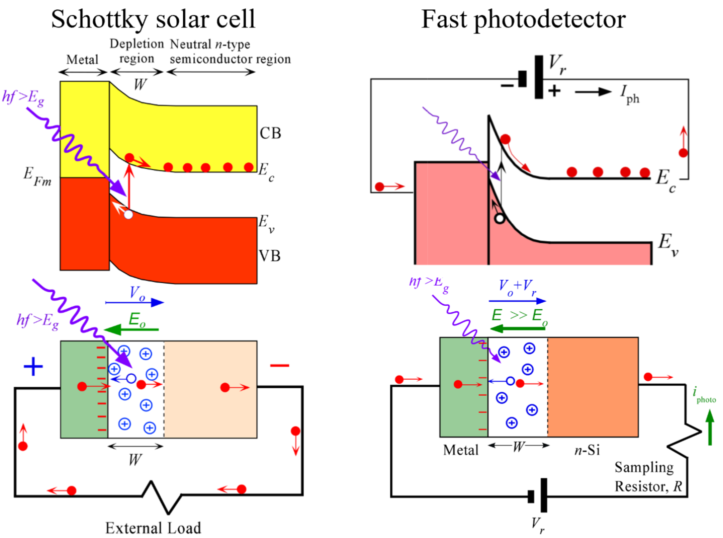
A drawing with the two cases is provided in figure 1.26, which shows the potential configurations for the Schottky junction solar cell (left) and fast photodetector (right).

1.16 Ohmic contacts

As a final note, we discuss the other case, which is an Ohmic contact. As you can imagine, the rules to create an Ohmic contact will be opposite that of a Schottky contact. For an n-type semiconductor, this requires that:

$$\Phi_m < \Phi_n \tag{1.105}$$

Now there are electrons that are at a higher energy than allowed states within the semiconductor, meaning that the electrons will want to tunnel into the semiconductor conduction band to balance the fermi levels. As this occurs, the electrons at the metal-semiconductor interface lead to an unbalanced charge and an accumulation of electrons, leading to an *accumulation region* that bends the bands. This time, the bands bend down. The bending occurs for the same reason - looking at the equation that relates the carrier density to $E_C - E_{Fn}$ we expect that increasing n will decrease $E_C - E_{Fn}$, leading to the conduction band to bend down. Figure 1.27 shows the band bending as a result of contacting the metal with



Schottky solar cell

Fast photodetector

Figure 1.26: Some use cases for a Schottky junction. When illuminated with light, electron-hole pairs generated in the depletion width of the semiconductor are swept in opposite directions due to the built in electric field. With the appropriate external load, this can act as a solar cell to collect energy from the absorbed light. Reverse biasing the junction can create a strong electric field that can speed up the drift velocity of the photogenerated carriers, which can lead to a fast photodetector. The voltages can be quite large, as the depletion width tends to be relatively small within the semiconductor.

a semiconductor in this case.

The conductivity in this region is more straightforward to evaluate. The higher resistance region in this circuit is now the semiconductor, as the accumulation region has a higher concentration of electrons. Therefore, the current flowing through this composite structure is simply defined as $J = \sigma E$, where the conductivity σ is from the bulk semiconductor. This gives a linear I-V curve that is reminiscent of Ohm's law (hence "Ohmic". We should note here that the contacts are not *inducing* a linear conduction relation, but rather that the electrodes are not *limiting* the conduction of the sample they are contacting. This is in contrast to the Schottky contact, in which the interaction of the metal with the semiconductor leads to a difference in conduction.

1.17 Summary an p-type Semiconductors

The conclusion of this section is simply to highlight the complexity of semiconductor materials as one begins to think about making useful devices. The ability to locally dope and control the majority/minority carrier concentrations, generate carriers, and apply fields to make them drift leads to a rich set of possibilities for controlling electrons and holes in devices. Now that we understand the impact that the Fermi level has when connecting two dissimilar materials we can move forward into more complex devices, which in this class will manifest with the metal-oxide-semiconductor field-effect transistor.

As a final note, we should highlight the rules again for a Schottky or Ohmic contact. We discussed the case of an n-type semiconductor, which has the following relationships to create a Schottky contact:

$$\Phi_m > \Phi_n \tag{1.106}$$

and an Ohmic contact:

$$\Phi_m < \Phi_n \tag{1.107}$$

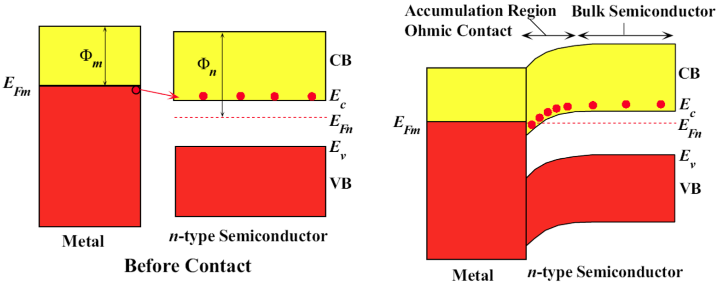
If we have a p-type semiconductor, the majority carriers are holes rather than electrons. Therefore, we need to think about things with respect to the valence band rather than the conduction band, and the inequalities are flipped. For a p-type semiconductor, we find that a Schottky contactor junction will form when:

$$\Phi_m < \Phi_p \tag{1.108}$$

and an Ohmic contact will form when:

$$\Phi_m > \Phi_p \tag{1.109}$$

Where the subscripts for the work function Φ_p refer to the p-type semiconductor. It is instructive to draw out the band diagrams for this case to convince yourself of the same behavior.



Before Contact

After Contact

Figure 1.27: An Ohmic contact, where the work function of the metal is less than the semiconductor. Before contact (left) we see that there are electrons in the metal that are at a higher energy level than the Fermi level of the semiconductor, meaning that there are states that the metal electrons can find in the semiconductor to lower the overall energy. Upon contact (right), the electrons move into the semiconductor to create an accumulation region. The bands bend downward accordingly to reduce the distance between the fermi level and the conduction band. This leads to a region with higher electron concentration and therefore lowere resistance, which facilitates electron conduction.

1.18 Transport in semiconductor devices question

In class we began discussing The ramifications of transport in Schottky and Ohmic contacts. To recap our discussion from class, if we have an *Ohmic* contact, the charge flow through the junction is not limited by the metal-semiconductor interface - it is instead limited by the higher resistance semiconductor. In this case we expect to have a *linear* relationship between current I and voltage V. With a *Schottky* contact, we expect a strongly nonlinear behavior.

For the homework problem, we need to consider a semiconductor with potentially two different metals that could have either ohmic or schottky contacts. To understand the behavior for positive or negative voltages, we need to understand what will limit or determine the current. This is a circuit that is in series, so we can look at the components resistance to determine the limiting factors. It should be apparent that the metals have much higher conductivity than the semiconductor due to the much higher concentration of carriers. Because of this, the problem boils down to whether the semiconductor or the metal-semiconductor junction limits the overall behavior.

The definition of an ohmic contact is that the metal-semiconductor junction does not impede current flow. Then, the limiting component of this circuit is the semiconductor itself, which has a resistance set by the resistivity. Then, it should be clear that you expect a linear relationship between current and voltage if the limiting component is the semiconductor. So, if you have two Ohmic contacts, then you expect the current to have a linear relationship (i.e. $V = IR$) for all voltage values in this case.

The unique aspect of a Schottky junction is that the ability for charge to flow is strongly dependent on the sign of the voltage that is applied. Under *forward bias*, the applied voltage reduces the built in voltage at the metal-semiconductor voltage. This dramatically increases the conductivity at this interface. This means that when a Schottky barrier is under *forward bias*, we expect that the bulk semiconductor will still be the limiting resistance in this case. Conversely, under *reverse bias*, the built in field grows with the voltage. This means that it becomes *harder* for current to flow from the semiconductor to the metal, so the current is limited by the thermal generation of carriers that overcome the barrier energy Φ_B. This value is almost independent of voltage, so you expect the IV curve to very rapidly go to a nearly constant value.

To summarize, this means that if you have a junction that has *any* Schottky junction under reverse bias, you expect that the IV curve to be nearly constant. If you have a Schottky junction under forward bias, the limiting factor in conduction will not be the Schottky junction.

By way of example, we can plot the scenario where we have an Ohmic contact and a Schottky contact. in this scenario in figure 1.28, positive voltages correspond to a reverse bias for the Schottky diode. Then, we expect the voltage to be limited when a positive voltage is applied, so you expect to see a nearly flat current with increasing voltage. However, when we apply a negative voltage, the Schottky junction is under forward bias. In this scenario, the largest resistance to current flowing is the semiconductor. This scenario we would expect a linear relationship between voltage and current.

To summarize: To determine the current-voltage relationship, consider first what the most resistive part of your circuit will be. In these scenarios they appear to be either a reverse biased Schottky junction or the bulk semiconductor. If there is a reverse biased schottky junction, the current will quickly saturate to a constant value. If the limiting factor is an Ohmic contact, then you expect a linear relationship with a slope related to the resistance of the semiconductor.

2 pn Junctions and associated devices

2.1 Introduction

We have spent a significant amount of effort developing the fundamentals of charge transport inside semiconductors. We learned that different dopants can increase the electron or hole concentrations, and that optical excitation and absorption can be used to increase the concentration of electrons and holes beyond thermal equilibrium (at least transiently). Finally, we discussed how electric fields can move electrons and holes, and how the contact of two dissimilar materials can induce a built in electric field, cause drift and diffusion currents of charge carriers, and potentially impede the flow of charge carriers through the creation of a barrier.

In this chapter we will explore some prototypical examples of semiconductor devices that are formed with both p-type and n-type semiconducting materials. Here, we will discover that these materials can form diodes and junctions that are similar to a

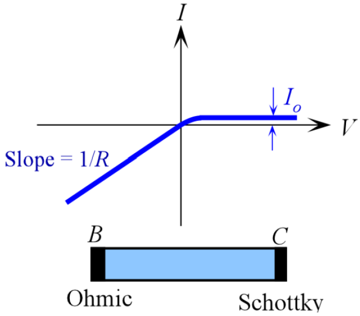


Figure 1.28: IV Characteristic of a semiconductor with an ohmic and schottky contact. Under positive voltage, the diode is reverse biased, leading to a current limited by the Schottky junction. For negative voltages, the Schottky junction is forward biased, so the slope is related to the resistance of the semiconductor. This is a highly asymmetric response.

Schottky junction at the metal-semiconductor interface. We will develop this further and explain its ramifications on the design of solar cells and LEDs. In principle these can be extended to lasers and other complex devices. We will wait to discuss the transistor until we have had a chance to discuss how dielectric materials function, as we will use this as part of the description of a Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET).

2.2 Junction formation and built in fields

Similar to the formation of a Schottky junction, we consider the case of two semiconductors with different doping concentrations. Here, we have one p-type semiconductor and one n-type semiconductor. These can be formed, for example, through differential doping of parts of a silicon wafer. Before these materials make contact, we note that the Fermi levels are different. Because this represents the chemical potential of the electrons in the system, we expect there to be a flow of electrons to reduce their overall energy. They will recombine with holes in the p-type semiconductor to reduce the concentration of electrons and holes in their respective materials. The remaining charged dopant cores will induce a space charge that will balance this effect. After equilibrium is established, the Fermi level will be constant across the junction and we expect to see band bending, like we did for the Schottky junction. This is represented in figure 2.1:

We can compute the built-in field from this sketch by first noting that the charge density will induce an electric field through:

$$\frac{dE}{dx} = -\frac{d^2V}{dx^2} = \frac{\rho_{net}}{\epsilon_0\epsilon_r} \tag{2.1}$$

Such that if we know the charge density of the depletion region, we have essentially solved the problem. First, we can assume that the charge density will be constant in each layer and defined by the doping densities (this is a reasonably good approximation, but note that n and p in these regions is not identically zero, just much smaller than their initial concentration. Then, in the p region we have a charge density of eN_a for the acceptors and eN_d for the donors. We will

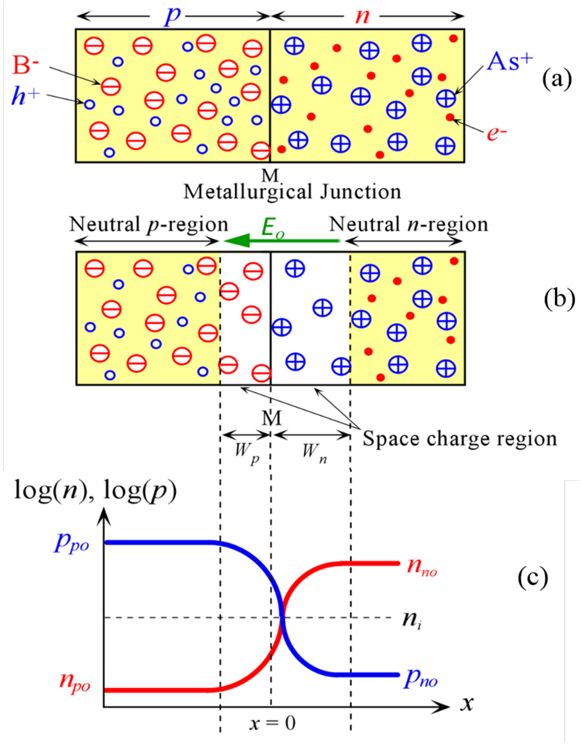


Figure 2.1: Depletion region of a pn junction, which includes a depletion width that goes into both the p-type and n-type semiconductors. Adapted from Kasap

independently define the widths, with the region of the p type depletion region as W_p and the n type region as W_n, and we will set the origin $x = 0$ at the junction.

We can compute the electric field as:

$$E(x) = \frac{1}{\epsilon_0 \epsilon_r} \int_{-W_p}^{x} \rho_{net}(x) dx \tag{2.2}$$

where we begin at the farthest left portion of the charge density and can integrate to any point x within the depletion region. This is shown in figure 2.2:

Note that the highest electric field is right at the junction (here, the field uses the phrase "metallurgical junction" to describe the exact p-n connection or junction point).

Once we know what the electric field looks like, we can compute the voltage as:

$$V(x) = -\int_{W_p}^{x} E(x) dx \tag{2.3}$$

to show that the voltage is a function of space. Finally, we can note that hte potential energy that an electron or hole will feel given this voltage is defined by

$$PE = qV(x) \tag{2.4}$$

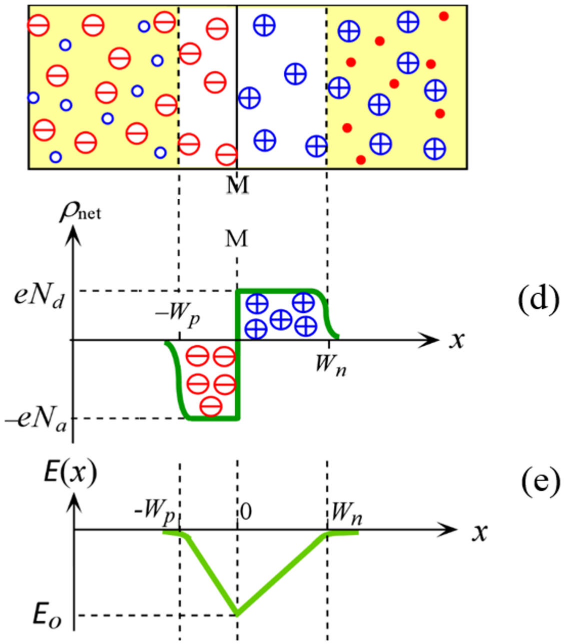


Figure 2.2: Depletion region of a pn junction, and the built in electric field from the charge density. Adapted from Kasap

Such that they feel a different potential energy because of the charge. This is all schematically illustrated in figure 2.3:

Since we know that electrons and holes must recombine in pairs, we can set bounds on what the depletion widths will be based on the doping density. This can be equated as:

$$N_a W_p = N_d W_n \tag{2.5}$$

Where we can clearly see that this is essentially balancing the total acceptors and donors that are exposed in the depletion region. We notice from figure 2.2 that the electric field is approximately linear in each depletion region, and that we have assumed that the carrier density is constant. Then, we can show that the maximum electric field will be given at $x = 0$ and will be:

$$E_0 = -\frac{eN_d W_n}{\epsilon_0 \epsilon_r} \tag{2.6}$$

Given that the voltage is the integral of the field and that the elctric field looks like a triangle, we can simply write down that:

$$V_0 = -\frac{1}{2} E_0 W_0 = \frac{eN_a N_d W_0^2}{2\epsilon_0 \epsilon_r (N_a + N_d)} \tag{2.7}$$

Where we have defined that $W_0 = W_a + W_d$.

To actually define the built-in voltage and therefore compute the depletion width, we need to do some math. We know that the concentration of carriers in a semiconductor is proportional to a Boltzmann factor that takes into account the energy

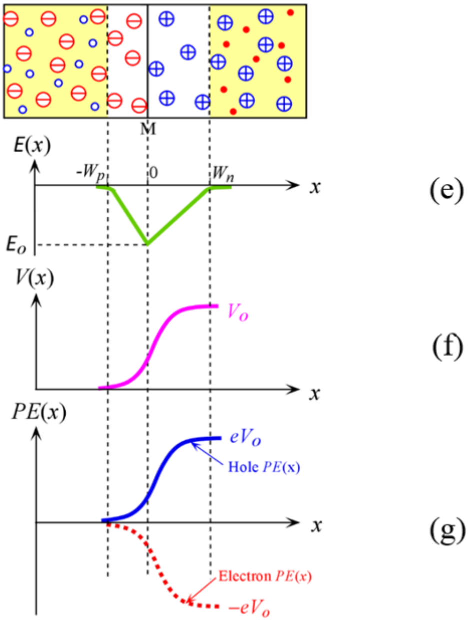


Figure 2.3: Built in field and potential energy within the pn junction. Adapted from Kasap

gap. That is, $n(E) \propto \exp -E/k_BT$. Since there is an energy difference in the band gap between the n side and p side of the junction, we can take the ratio of the two concentrations and arrive at:

$$\frac{n_2}{n_1} = \exp\left[-\frac{E_2 - E_1}{k_BT}\right] \tag{2.8}$$

where $E = qV$, with q being the charge of the carrier. At the p side of the junction, $E = 0$ and $n = n_{p0}$. At the n side, $E = -eV_0$ and $n = n_{no}$. Then, we can take the ratio to be:

$$\frac{n_{po}}{n_{no}} = \exp\left(-\frac{eV_0}{k_BT}\right) \tag{2.9}$$

And similarly,

$$\frac{p_{no}}{p_{po}} = \exp\left(-\frac{eV_0}{k_BT}\right) \tag{2.10}$$

We use the law of mass action to re-define $p_{no} = \frac{n_i^2}{n_{no}} = \frac{n_i^2}{N_d}$, and recognizing that $p_{po} = N_a$, we can use the above equations to define that:

$$V_0 = \frac{k_BT}{e} \ln\frac{N_aN_d}{n_i^2} \tag{2.11}$$

Where we have now define the built in voltage entirely in terms of material parameters that are decided on by the device design.

We can therefore solve for the depletion width W_0 to be:

$$V_0 = \frac{eN_aN_dW_0^2}{2\epsilon_0\epsilon_r(N_a+N_d)} \tag{2.12}$$

which can be solved to find that:

$$W_0 = \left(\frac{2\epsilon_0\epsilon_r V_0}{e}\frac{(N_a+N_d))}{N_aN_d}\right)^{1/2} \tag{2.13}$$

This is a key result. When we apply a voltage, we essentially modify V_0, the built in electric field, and therefore the potential energy that the charge carriers feel. The depletion with is also modified, which has implications on the transport of electrons and holes.

2.3 Forward bias

When we apply a field that opposes the built-in field, it will (1) increase the minority carrier concentration, (2) reduce the depletion width, and (3) increase the diffusion current of minority carriers. We will sketch this out below. Figure 2.4 shows the carrier concentration distribution and the built in electric field when a pn junction is forward-biased.

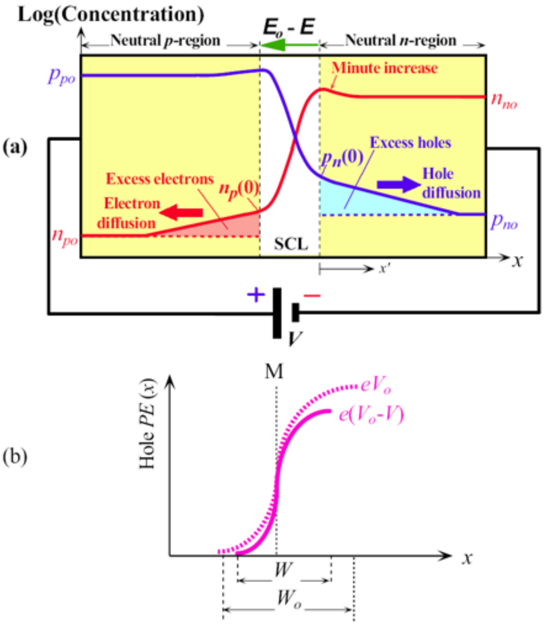


Figure 2.4: Carrier concentrations under forward bias, as well as the built in field and reduction in depletion region width. Compare this distrubtion against the one in figure 2.1. Adapted from Kasap

Additionally note that the fermi level of the p and n type materials have also shifted because of the applied electric field. Like in the Schottky junction case, the bulk of the semiconductor energy level will be modified because the voltage induces a potential energy increase or decrease for the fermi level. This is effectively the reason for which the built in voltage increases or decreases.

We can compute the change in the depletion width, which is simply related to the net built in field. This is given as:

$$W \approx \left(\frac{2\epsilon_0\epsilon_r}{e}\left(\frac{N_a+N_d}{N_aN_d}\right)(V_0-V)\right)^{1/2} \tag{2.14}$$

Additionally, we can compare the relative widths for the depletion region within the n-type or p-type material. To a decent approximation, the depletion width in each region is essentially the weighted average of the doping concentrations of the bulk semiconductor. The depletion width for holes is essentially given as:

$$W_h \approx W\frac{N_a}{N_a+N_d} \tag{2.15}$$

Because the depletion width has decreased and the built in voltage is now lowered, we expect there to be a greater amount of diffusion of carriers, as the force that was establishing equilibrium is now reduced. This difference in concentration of at the edge of the depletion region (here, defined as $x = x_n$) can be expressed as:

$$p_n(x_n) = \frac{n_i^2}{N_d}\exp\left(\frac{eV}{k_BT}\right) \tag{2.16}$$

which essentially is a reflection of the fact that the barrier has been reduced by a value of V and we can compute the Boltzmann factor of additional charge carriers. Note, however, that this is a very strong function, and even modest voltages can dramatically increase the concentration of carriers.
We can also compute the change in electron concentration at the edge of the depletion region in the p type material (here, $x = -x_p$, if 0 is the position of the exact contact position between n and p type semiconductors) as:

$$n_p(-x_p) = n_{no}\exp\left(\frac{-e(V_0-V)}{k_BT}\right) = n_{no}\exp\left(\frac{-e(V_0)}{k_BT}\right)\exp\left(\frac{eV}{k_BT}\right) = n_{po}\exp\left(\frac{eV}{k_BT}\right) \tag{2.17}$$

The increased density of minority carriers at the edge of the depletion region leads to a large diffusion current of minority carriers. We note that the continuity equation will give that the excess concentration of holes in the n type material can be expressed as:

$$\Delta p_n(x) = B\exp\left(-\frac{x}{L_h}\right) = p_n(x) - p_{no} \tag{2.18}$$

Where B is some constant and L_h is the diffusion length for holes (recall that $L_h = \sqrt{D_h\tau_h}$. We can insert the expression above (3.16) into this expression to determine the change in concentration with distance, by noting that $n_i^2/N_d = p_{no}$. We then arrive at:

$$\Delta p_n(x, x > x_n) = p_{no}\exp\left(\frac{x_n - x}{L_h}\right)\left[\exp\left(\frac{eV}{k_BT}\right) - 1\right] \tag{2.19}$$

Using the definition we have for the diffusion flux $J_p = -eD_h \frac{d\Delta p_n(x)}{dx}$, we can evaluate this at the edge of the depletion width in the n type semiconductor to compute the flux of charge carriers. When we do this, we obtain:

$$J_p(x_n) = \left(\frac{eD_h n_i^2}{L_h N_d}\right)\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \tag{2.20}$$

This has been a lot of math. We can do a similar analysis for the electrons, and we will get a similar functional form. Here, we find that there is an exponential decay in the flux as we move away from the depletion width, as the concentration gradient is reduced.

The total current flux needs to take into account both the fluxes of the electrons and holes. The complete solutoin takes into account the contribution of the electrons and holes, and takes the form of:

$$J_{tot} = \left[\left(\frac{eD_h}{L_h N_d}\right) + \left(\frac{eD_e}{L_e N_a}\right)\right] n_i^2 \left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \tag{2.21}$$

where we can define the *Reverse saturation current density* J_{SO} as:

$$J_{SO} = \left[\left(\frac{eD_h}{L_h N_d}\right) + \left(\frac{eD_e}{L_e N_a}\right)\right] n_i^2 \tag{2.22}$$

A drawing of the different contributions from drift and diffusion are shown in figure 2.5:

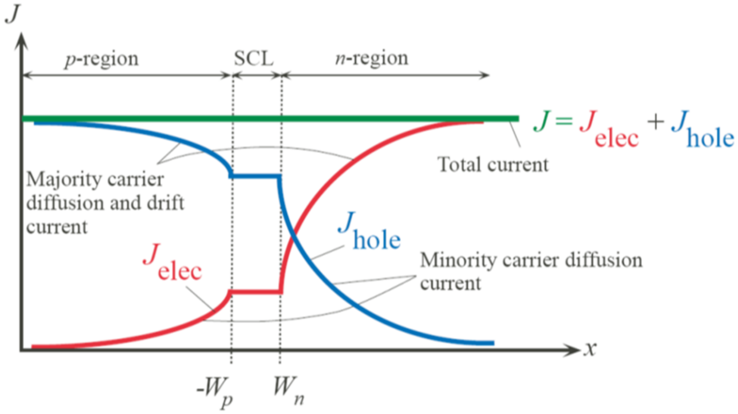


Figure 2.5: Current density under forward bias in a pn junction. The total current density is constant, so the relative components from electrons and holes, drift and diffusion, can change as a function of position. Adapted from Kasap

2.4 Reverse bias and generation current

As you can imagine, applying an electric field that enhances the built in field (i.e. driving hte junction under reverse bias) will decrease the ability for electrons and holes to traverse the junction. Because of this, we expect to be limited by the reverse saturation current density defined above, as the exponential term in the diode equation very quickly disappears and we are left with the only remaining term. However, there is one more detail we have yet to cover. If there are electron-hole pairs that are generated within the depletion region, we would find that the built in electric field will cause the electrons and holes to move in opposite directions. This will sweep out generated charge carriers from the depletion region, which will introduce an additional small (but non-zero) component to the current density. The *generation rate* of these carriers will thus depend on the width of the depletion region (as carriers can be generated anywhere in it, and therefore more or less will be created if the depletion region is increased/decreased). It will also depend on the generation time and the intrinsic concentration of carriers. We will not derive the term but merely reproduce it. The Total reverse current density is:

$$J_{rev} = \left[\left(\frac{eD_h}{L_h N_d}\right) + \left(\frac{eD_e}{L_e N_a}\right)\right] n_i^2 + \frac{en_i}{\tau_g} W(V) \tag{2.23}$$

Where the first term is the reverse saturation current density and the last term is the *generation current* that is induced due to electron-hole pair generation in the depletion region. Here, we explicitly indicate that the depletion width $W(V)$ is a function of the applied voltage, and have introduced a phenomenological timescale τ_g that we call the *generation time*.

As you can imagine, this indicates that one can have charge carriers generated in the depletion region of the pn junction. The effective voltage then creates a driving force to separate the electrons and holes, which then is turned into an additional current density. This can be a foundational approach in developing devices like solar cells, in which we want to convert light into electricity. For posterity, I have included the band structure and the depletion width for the reverse bias case in figure 2.6:

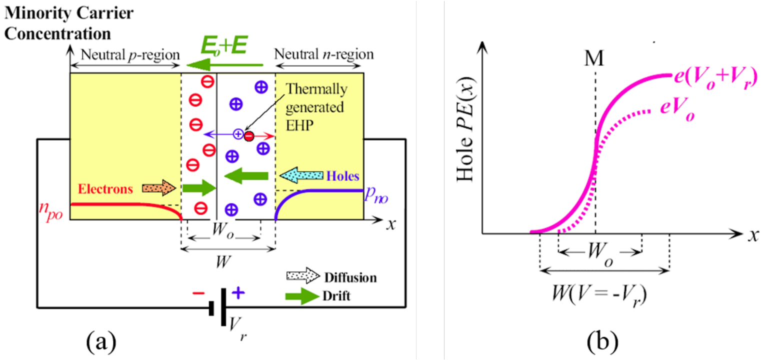


Figure 2.6: Schematic of reverse biasing a pn junction. This widens the depletion region. The electron and hole concentrations will be reduced near the depletion regions due to the additional applied field. Adapted from Kasap

2.5 Recombination current during forward bias

While we're discussing all the non-idealities in pn junctions, let's discuss the reason for the "ideality factor" η in the diode equation. We have not discussed what will happen to the current density when a particle undergoes recombination within the space charge region. This will increase the current under forward bias.

Recall that the depletion width W depends on the applied bias, which will change the size of the region where recombinatoin is relevant for this example. This rate additionally will depend on the concentration of charge carriers, which varies in space (and is a difficult calculation to determine). We will not derive the expression, but present one with effective parameters. A more advanced course may delve into this.

The recombination current density takes the form:

$$J_{recomb} = J_{ro}\left[\exp\left(\frac{eV}{2k_BT}\right) - 1\right] \tag{2.24}$$

where

$$J_{ro} = \frac{en_i}{2}\left[\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h}\right] \tag{2.25}$$

Here, the τ values are the effective lifetimes. We must add this source of current density to the total term for the current density under forward bias. We then arrive at the final equation, which is:

$$J = J_0\left[\exp\left(\frac{eV}{\eta k_BT}\right) - 1\right] \tag{2.26}$$

where we have the ideality factor η to take into account this phenomenological factor of recombination.

What does all of this mean? If we drive current under forward bias through a pn junction and there is recombination, more current is required at a given voltage. If we use a direct bandgap semiconductor, this recombinatoin can efficiently generate light, which can be used to great *light emitting diodes.*

2.6 Band bending summary and general notes

Wow, this is a lot of content! There are many variables to keep track of, and the overall physics can be a bit complex here. Some general advice: (1) Recall that the law of mass action must apply at equilibrium, which will set the concentrations of carriers. (2) Because we are dealing with two different types of semiconductor in the same problem (p-type and n-type), keeping track of subscripts is quite important. (3) We can use the boltzmann factor analysis to convert between carrier types and concentrations at different points along the structure. (4) The band bending is induced because of the built in field from the depletion region, which essentially acts like a capacitor. (5) Applying an electric field can displace the fermi levels in the p and n type materials compared to at thermal equilibrium, which can reduce or enhance the band bending.

This last point is discussed in figure 2.7, where we consider all of the various cases from earlier parts of this section.

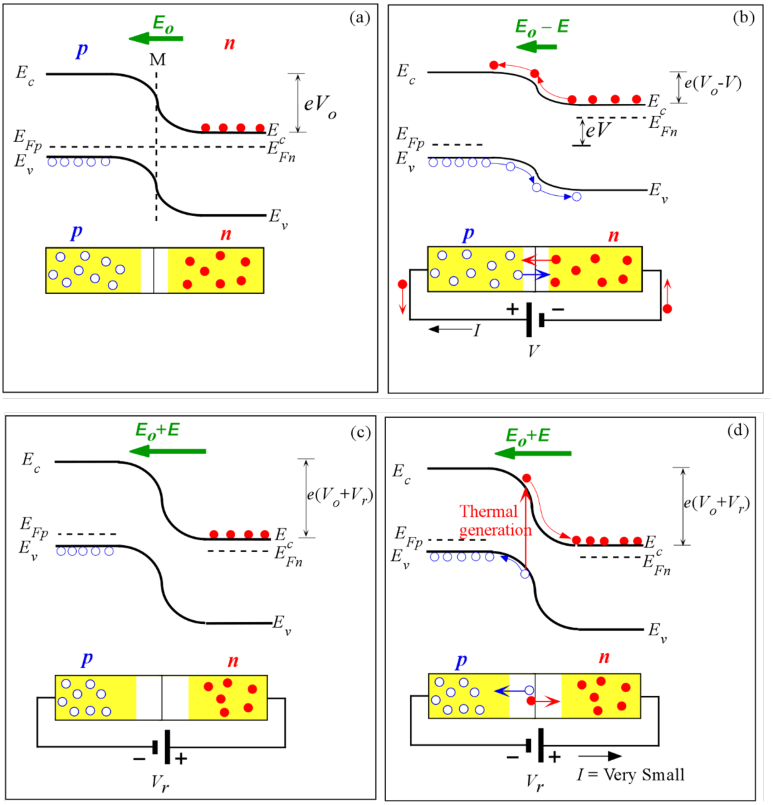


Figure 2.7: Schematic of the band bending in a pn junction. a) is the case where there is no applied field, with the bending defined entirely by the concentration of dopants in the n and p type materials. b) shows the case of forward bias, in which an applied field opposes the built in field and reduces the band bending by increasing the energy level (and fermi level) of the band. c) shows reverse bias, where the opposite behavior occurs. Finally, d) shows what happens when electron-hole pairs are generated in the space charge layer (within the depletion layer) of a junction - charge carriers are spontaneously swept out of the junction. Adapted from Kasap

2.7 Solar Cells

Finally, some content about applications of these diodes! A solar cell in the dark that is composed of a pn junction acts just like how we have described so far. When we turn the light on and shine it onto the semiconductor, we expect that any charge carriers that are generated within the depletion region will see the applied electric field and be swept out of the depletion region. The electrons will go one way and the holes will go the other way. We can then extract meaningful current from this. To maximize the efficiency of a solar cell, we want to minimize the non-radiative recombination mechanisms in the diode. A schematic of a solar cell in operation is given in figure 2.8:

Some important parameters of the IV characteristic of a solar cell are the open circuit voltage (V_{OC}) and the short-circuit current (J_{SC}). It can be shown for a

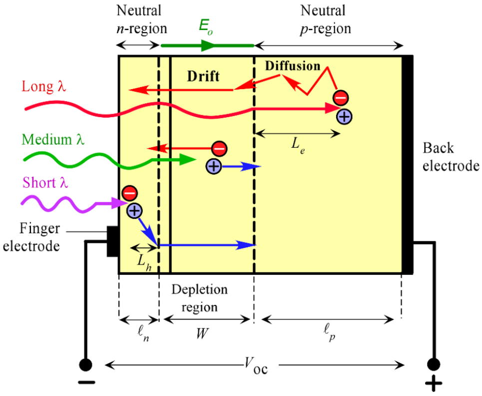


Figure 2.8: Design of a solar cell. Light is shining from the left into the solar cell, which has electrodes on either side to collect current. The depletion region spontaneously separates charges due to the built in voltage, which can be used to harvest energy from the absorbed light.

solar cell that these take the form:

$$V_{OC} = \frac{\eta k_B T}{e} \ln\left(\frac{J_{SC}}{J_{SO}}\right) \tag{2.27}$$

where the reverse bias saturation current J_{SO} is defined to be

$$J_{SO} = en_i^2\left(\frac{D_e}{L_e N_a} + \frac{D_h}{L_h N_D}\right) \tag{2.28}$$

Clearly, we would like for the reverse saturation current to be minimized to maximize the open circuit voltage. In addition, one may notice that a longer lifetime τ increases the open circuit voltage V_{OC}, which in turn provides higher power and efficiency. This makes sense if we need the generated charge carriers to last long enough to be collected by the electrodes.

The power for a solar cell is defined as:

$$P = V_{OC} I_{SC} FF \tag{2.29}$$

where FF is the fill fraction, and is defined as:

$$FF = \frac{I_M V_M}{I_{SC} V_{OC}} \tag{2.30}$$

where I_M and V_M are the operating current and voltage that can maximize this function. You can do this graphically. Figure 2.9 shows the relevant curves you will find for solar cells as a function of the optical power that shines on them. the short circuit current J_{SC} is the current at 0 voltage (i.e. crosses the vertical axis), while the open circuit voltage V_{OC} is the voltage when there is no current (i.e. crosses the horizontal axis). To optimize the power, we must find a value along the curve where the IV is maximized. You may notice, therefore, that an ideal solar cell will have a nearly constant IV characteristic until a turning point, in which it would quickly reduce in current until it reaches the open circuit voltage. This notion of “Fill fraction” is essentially measuring how close the IV characteristic is to a rectangle in shape.

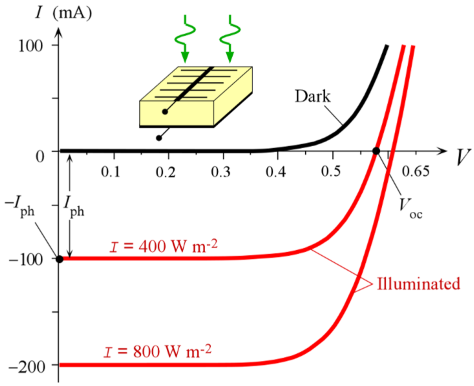


Figure 2.9: IV characteristic of a solar cell with the light on and off. Different powers lead to different short circuit currents. Here, the short circuit current is also labeled as $-I_{ph}$ to represent that the current is caused by the absorption of light generating charge carriers. Adapted from kasap.

2.8 LEDs

We will introduce the basics of an LED. Note, however, that some of the most efficient LEDs are based on heterostructures that can create quantum wells to trap charge carriers. This might be worth looking at - in fact, this won a nobel prize for its work.

In a basic pn junction LED, we consider the forward bias of the junction in a direct bandgap semiconductor. This forces carriers through the depletion region. When the electrons and holes recombine, light can be emitted. This is essentially the opposite process as a solar cell, in which we absorb light to collect its energy. Here, we expend energy by creating photons. Figure 2.10 shows the band structure of this case.

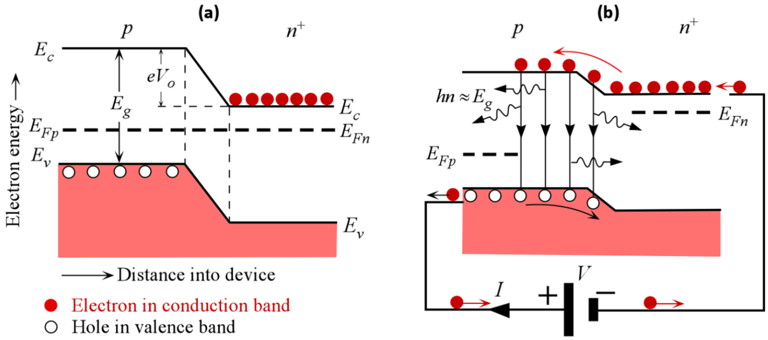


Figure 2.10: Band structure of an LED. Upon forward ias, the fermi energies are no longer in equilibrium. Charges can move through the depletion region more easily, which can enable more efficient extraction of light through the recombination of these carriers within the depletion region.

Some general notes to consider. Because the charge carriers are thermally generated, there is a distribution to their color, as electrons with different energies can recombine with holes of different energies to create a variety of frequencies. Based on the density of states, the most intense energy that is emitted is approximately $E_g + 0.5k_BT$. The distribution of colors can also be approximated to be approximately $\Delta h\nu = 1.9k_BT$, or the change in wavelength is approximately:

$$\Delta\lambda = \lambda_0^2 \frac{mk_BT}{nc} \tag{2.31}$$

There are many factors that can impact the efficiency of LEDs, and many of them are fundamentally materials problems! Reombination and trapping centers within direct bandgap semiconductors are a critical challenge, and the formation of the blue LED was awarded a Nobel Prize!

We can compute quantum efficiency of these devices. The internal quantum efficiency is essentially the rate of radiative recombination normalized by the total rate of recombination, which can include nonradiative routes. Then,

$$\eta_{IQE} = \frac{\Phi_{ph}}{I/e} \tag{2.32}$$

which is essentially photons emitted internally per second/total carriers lost per second.

The extraction efficiency η_{EE} is the number of emitted photons externally from the device divided by photons generated internally by recombination. Then, the total output power is defined to be:

$$P_0 = hf \times \eta_{EE} \times \eta_{IQE}(I/e) \tag{2.33}$$

and the external quantum efficiency η_{EQE} is the total conversion of electrons that flow into the LED to the number of generated photons that are emitted from the device.

$$\eta_{EQE} = \frac{P_0/hf}{I/e} \tag{2.34}$$

From this, you can determine the wall plug efficiency if desired.

2.9 Zener Diode

We will spend a brief amount of time discussing the Zener diode, as it will make an appearance in laboratory exercises. While we have covered the general IV characteristics for an IV curve of a diode, we have not discussed the notion of breakdown. If we apply a large enough reverse bias, we may find that the IV characteristic suddenly has a strong current, and devices will typically start to fail. The deviation from the reverse bias saturation current is the onset of breakdown. This can happen due to an effect called avalance breakdown, in which electrons have high energy collisions with atoms and the device starts to become damaged. This is obviously undesirable. The Zener diode attempts to avoid this possibility by creating a situation where electrons can tunnel through the potential barrier rather than destroy the device. While this is still breakdown from the saturation current, it is done in such a way that the device is not damaged or destroyed. This is done by having a high doping concentration on either side of the junction, which creates a very narrow depletion region. When reverse biased, the depletion region is decreased even more substantially, such that electrons are able to tunnel through the barrier. A schematic bandstructure and an example of reverse bias breakdown IV curve is given in figure 2.11. One outcoe of such a device is that you can make these devices operate under very high reverse bias conditions without fear of breaking them.

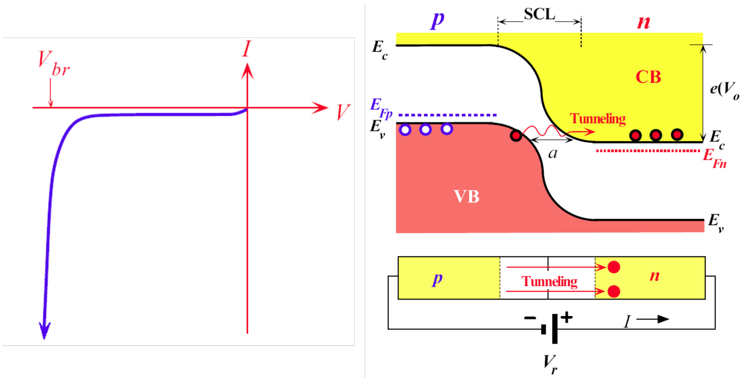


Figure 2.11: left: IV characteristic for strong reverse bias, showing a sudden increase in the current at the breakdown voltage. This is typically bad. right: A Zener diode is composed of a pn junction with high doping concentrations, leading to a very narrow depletion region. Upon high voltage operation, the device will break down but will not destroy the device.

3 Dielectric materials

3.1 Introduction

So far, we have discussed materials that have free charge carriers that can move around in response to an applied electric field, which has created several complex phenomena that gives rise to modern electronics and optoelectronics. We will briefly step away from this to discuss the properties of insulators, which do not have (intentional) carriers. We will need to understand the manner in which they respond to an applied electric field, which will end up polarizing the device. The material mechanisms that give rise to polarization will define the dielectric constant. We will find that the behavior will also depend on the frequency of the applied electric field. This will eventually lead us to understand some aspects of the optical properties of thin films.

3.2 Capacitance

You may recall from introductory physics that the Capacitance of two parallel plates with area A and separation distance d is given by:

$$C = \frac{\epsilon_0 A}{d} \tag{3.1}$$

where C is the capacitance (with units of Farads F), ϵ_0 is the permittivity of free space $\epsilon_0 = 8.85 \times 10^{-}12$ m^{-3} kg^{-1} s^4 A^2 (these units are also equivalent to F/m). A is the area (units of m^2), and d is the distance between the plates (with units of m). The capacitance will relate the charge that can be stored on these plates for a given voltage that is applied. We will see that when a material is placed in between, then this equation will be modified due to the dielectric response of the material. This is the parameter ϵ_r or the *relative permittivity*. The goal of this section is to understand where ϵ_r comes from.

We first start with a hypothetical experiment with a parallel set of metallic plates that are connected to a circuit. All variables and the experiment is shown in figure 3.1.

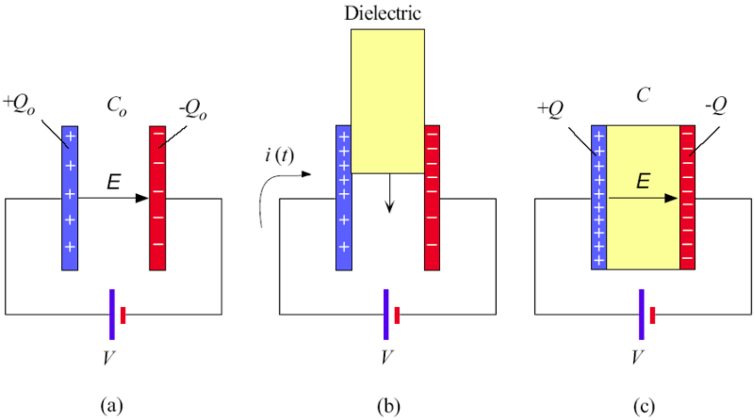


Figure 3.1: Hypothetical experiment with parallel plates. As a voltage is applied, charge is stored on the plates, which induces an electric field. Inserting a dielectric material between the places leads to a current flow as more charge can be stored on the plates. The new capacitance can therefore measure the dielectric constant of the material. Adapted from Kasap

If we apply a voltage, the plates will develop charge of opposite sign on either plate. The metal, with a high density of conduction electrons, can supply the charges easily and they exist at the very edge/surface of the metal (this is in contrast to a semiconductor, in which there is some charge density going into the material). The different in charges at the surface induce an electric field between the region of positive and negative charge.

This geometry is essentially a capacitor. We can relate the Capacitance C to the applied voltage V by understanding how much charge Q can be stored on the plates.

In the case where there is nothing between the plates, we can determine the initial capacitance C_0 based on the initial charge Q_0 that is induced when the voltage is applied. The equation that relates these parameters is given as:

$$C_0 = \frac{Q_0}{V} \tag{3.2}$$

When we start to place a dielectric material inside of the gap, we find that the total amount of charge that can be stored on the capacitor places will go up. Note that we have specified that this material is a dielectric material, which is any non-conductive material. If it were conductive, then the charges on either plate would move in response to the voltage and the circuit would be shorted. In this hypothetical experiment, we find that the net charge Q goes up for a given applied voltage V. This can be measured through a current through the circuit as we insert the dielectric material.

Since we previously defined that the permittivity of free space is defined by the total amount of charge that can be stored on these plates in the absence of a material, we must modify this expression to take into account the response of the material. We define this value as the *relative permittivity* or the *dielectric constant*, ϵ_r that takes this into account. The total permittivity is then given by the ratio of the new charge stored compared to the plates without material Q/Q_0 or the new capacitance compared to the original capacitance C/C_0. Physically, the increased polarization of the dielectric from the applied field is responsible for this behavior. We must understand this to understand the relative permittivity. The consequence of this is the general equation:

$$\epsilon_r = \frac{Q}{Q_0} = \frac{C}{C_0} \tag{3.3}$$

such that the capacitance of this structure is therefore:

$$C = \frac{\epsilon_0 \epsilon_r A}{d} \tag{3.4}$$

3.3 Dipole moments and atomic polarizability

We will begin by developing a microscopic picture of the ability for a dielectric material to polarize. Whenever there are positive and negative charges separated by a distance, we can characterize the *electric dipole moment* that is generated. this is essentially the electrostatic response of a net charge density sparation. We can define the dipole moment as:

$$\mathbf{p} = Q\mathbf{a} \tag{3.5}$$

where $\mathbf{p}$ is the dipole moment, Q is the charge, and $\mathbf{a}$ is the separation between them. Note that this is for a positive charge $+Q$ separated from a negative charge $-Q$. This behavior is in general a vector, as the charges can be separated by a given distance in any spatial direction. We represent the physical picture of a dipole in figure 3.2.

This behavior is useful to think about and can be used in a variety of contexts. If we were to have an atom between the plates, we would expect that the electric

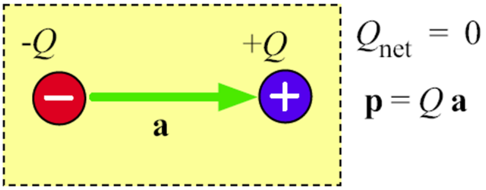


Figure 3.2: Description of a dipole within a volume (yellow area). The net charge in the volume is 0, but the separation of charges Q by a distance a induces a net dipole moment within the solid. Adapted from Kasap

field would cause the electrons to move in one direction, while the atomic core would move in the opposite direction due to having an opposite charge. The electron cloud can much more easily because the relative mass of the electrons is much lower than that of the atom. Thus, we would expect that an electric field applied by the capacitor plates would displace the electrons with respect to the atomic core. We can represent this effect simply by looking at the net charges and distances, and therefore represent this material as a dipole. Because this atom has an electric dipole moment, it is *polarized.* This leads to the ability for a material to polarize inside of a capacitor, leading to the dielectric constant.

The approach we will use to understand the dielectric constant will start with an understanding of how a single atom will respond to an electric field. Once we understand this, we can understand how a collection of atoms can polarize, and from this we can start thinking about the different contributing factors to this collection of atoms.

To capture this physics, the field has decided to use several different terms and variables to represent this that can often become a bit confusing to keep track. We will attempt to make this as clear as possible. First, we want to understand how easy it is for an atom to become polarized in response to an electric field, which we can use to built up intuition about the response of a material. To do so, we will define the *polarizability* α, which is the term that will directly relate the induced dipole moment in a material or atom in response to an applied electric field E:

$$p_{induced} = \alpha E \tag{3.6}$$

Here we see that this relationship is linear, in that it depends linearly on the applied field. There are several materials for which this is not true, and there are a lot of interesting things that can be done with a nonlinear polarization, but we will not discuss these now.

The model that we can use to determine how easily an atom can polarize is essentially every introductory model for a material, the mass on a spring. When we apply an electric field, the electrons will want to move in response to the field. As they move away from the atomic core, the Coulombic interaction creates a restoring force that opposes the applied electric field. If we apply a static electric field, these must balance. This can be represented as a force with a linear relationship with the distance as:

$$F_r = -\beta x \tag{3.7}$$

where the restoring force F is related to the displacement x by the bond strength β. The net force from the applied field is defined by the total number of electrons Z, their charge e, and the applied field E. This can be written down as:

$$ZeE = \beta x \tag{3.8}$$

which means we can define the induced dipole moment as:

$$p_e = (Ze)x = \left(\frac{Z^2e^2}{\beta}\right)E \tag{3.9}$$

where we have noted that the dipole strength is related to the total number of charges as well as the separation distance. We can measure the bond strength by looking at what happens when the electric field is suddenly turned off. If we do this, then the restoring force from the bond will cause the electron cloud to accelerate towards the atomic core, as there is no other force applied to create the dipole. We can then relate the restoring force as the force times acceleartion as:

$$-\beta x = Zm_e \frac{d^2x}{dt^2} \tag{3.10}$$

This is the common mass on a spring differential equation, which has a solution of simple harmonic motion defined as:

$$x(t) = x_0 \cos(\omega t) \tag{3.11}$$

where the resonant frequency ω_0 is

$$\omega_0 = \left(\frac{\beta}{Zm_e}\right)^{1/2} \tag{3.12}$$

With this relationship, we can relate the electronic polarizability α_e to values that can be looked up in tables as:

$$\alpha_e = \frac{Ze^2}{m_e\omega_0^2} \tag{3.13}$$

This resonant frequency is something that can be looked up for many atoms.

3.4 Collections of atoms and polarization vectors

Now that we have a basic relationship for the response of individual atoms, we can consider the cases perhaps more interesting to materials scientists - solid materials.

In a solid material we will have a collection of atoms in particular bonding configurations that will all respond to an applied electric field. Whereas the polarizability

α_e is a descriptor of an individual atom (and is therefore a *local* property descriptor), we need to define the net *Polarization* of a bulk material, which will have an associated set of *macroscopic* terms to consider. If we have n atoms that each have a dipole moment p_i (the subscript i would refer to the dipole of the ith atom), then we can measure the macroscopic polarization P by simply summing up the dipoles and taking the volume average. Note that the dipole is actually a vector property $\mathbf{p}_i$. Then, we can write the net polarization as:

$$\mathbf{P} = \frac{1}{V}\sum_i \mathbf{p}_i \tag{3.14}$$

Note that the net polarization is also a vector quantity - the induced direction wants to point in a particular direction.

We can simplify this to define an aggregate average dipole p_{av} in the solid. Then, this can simply be related to the number density N of dipoles per unit volume and the average dipole moment as:

$$\mathbf{P} = N\mathbf{p_{av}} \tag{3.15}$$

We can visualize this behavior in the thought experiment presented above. If we put a material between the parallel plates and apply an electric field, each atom will polarize in response to this. This is shown in figure 3.3:

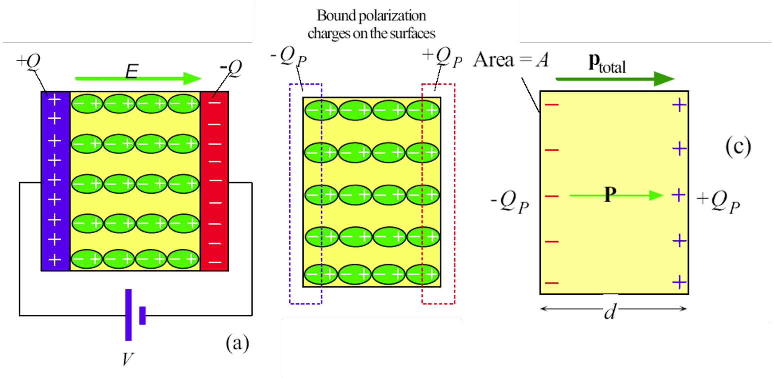


Figure 3.3: When a material is placed in between two plates, the atoms individually polarize in response to the applied field. Within the bulk of the material the net positive and negative charges cancel, but the net charges at the surfaces of the dielectric material remain. To balance this, new charge must be added to each plate, which can be measured to determine the dielectric constant of the material. Adapted from Kasap

When the atoms polarize, there is a net negative charge at the edge of the dielectric material by the positive terminal, while there is a net positvie charge on the edge of the dielectric material at the negative terminal. In order to balance the local charges at the metal plates, more charges must accumulate in the metal plates, which increases the net charge and therefore defines the dielectric constant. Additionally, the positive and negative charges that are in the bulk of the material cancel, leading to a polarized material with the net charges on the surfaces. The Polarization vector $\mathbf{P}$ is related to the net charges on the surface and the distance

d between them.

As a brief aside, we note that we care about the net dipole that is oriented along the applied field. The net dipole may actually be written as $\mathbf{p}_{induced}$, but the dipole moment we care about for these purposes is $\mathbf{p}_{av}$.

As we are thinking about the macroscopic response of the medium, we can express the total dipole moment as the net charge on either side of the dielectric material as:

$$p_{total} = Q_P d \tag{3.16}$$

where we have dropped the vector nature as the dipole is aligned with the electric field. We are using the definition Q_P as defined in figure 3.3. We can take the net polarization $\mathbf{P}$ as the volumetric average of this dipole as:

$$P = \frac{p_{total}}{V} = \frac{Q_P d}{Ad} = \frac{Q_P}{A} = \sigma_P \tag{3.17}$$

where σ_P is the surface charge density on the plates. Therefore, we have a way of directly measuring the macroscopic polarization of the material based on the net charge that is accumulated on a plate with a cross-sectional area A.

We can relate this to macroscopic phenomenal that depends on the applied electric field. As we increase the voltage, the electric field between the plates will increase, and we therefore expect the polarization to increase. For the microscopic picture we had defined the *atomic polarizability*, which described how easy an atom could polarize. Here, we define a new macroscopic value χ_e through:

$$P = \chi_e \epsilon_0 E \tag{3.18}$$

where χ_e is the *electric susceptibility*. This is the constant that relates how easy (or susceptible) the bulk material can polarize in response to an applied electric field.

Because the macroscopic polarization is related to the net induced dipoles that are created in the material, we can find a relationship between the macroscopic descriptors and microscopic phenomena. This can simply be expressed as the product of the number density of dipoles and the induced dipole moment as:

$$P = \chi_e \epsilon_0 E = N p_{induced} = N \alpha_e E \tag{3.19}$$

which implies

$$\chi_e = \frac{N\alpha_e}{\epsilon_0} \tag{3.20}$$

such that:

$$\epsilon_r = 1 + \chi_e = 1 + \frac{N\alpha_e}{\epsilon_0} \tag{3.21}$$

This is the key result of this thought problem - if we can appropriately describe the ability for individual atoms to polarize in response to an electric field, we can define the macroscopic properties that we can use to engineer our materials!

3.4.1 proof of relative permittivity (optional)

This section may be helpful to think through the result above, but is not required. Before any material is placed in between the parallel plates, we can express the elcctric field in the capacitor as:

$$E = \frac{V}{d} = \frac{Q_0}{\epsilon_0 A} = \frac{\sigma_0}{\epsilon_0} \tag{3.22}$$

where we have defined σ_0 as the free surface charge density - that is, the charge density on the plate when a voltage is applied in vacuum.

Note that when the dielectric has been placed inside the capacitor, the electric field has not changed, as the voltage and distance in the material has not changed. However, more charge appears on the plates in response to the effective charge that is induced on the surfaces of the dielectric material. The new charge $Q = Q_0 + Q_P$ defines a new surface charge density σ as:

$$\sigma = \epsilon_0 E + \sigma_P \tag{3.23}$$

Relating the surface charge density induced on the plates to the macroscopic polarization of the dielectric $\sigma_P = P = \chi_e \epsilon_0 E$ we obtain:

$$\sigma = \epsilon_0 (1 + \chi_e) E \tag{3.24}$$

Using our prior definition of the relative permittivity as:

$$\epsilon_r = \frac{Q}{Q_0} = \frac{\sigma}{\sigma_0} \tag{3.25}$$

such that:

$$\epsilon_r = 1 + \chi_e = 1 + \frac{N\alpha_e}{\epsilon_0} \tag{3.26}$$

3.5 Clausius-Mossotti equation

Unfortunately, our life is note quite as simple as the picture painted in the previous expression. If we could simply just sum up the dipoles of individual atoms, we could very quickly identify the permittivity of virtually any material. However, this assumes that the electric field E applied by the electrodes leads to the dipole response. However, we need to consider that the electric field inside of a material may look different locally than the macroscopic picture, which we may expect given the distribution of charge locally within the material. The electric field is not quite uniform at this level, and the crystal arrangement has a large impact on the field that an individual atom will see.

What we care about is the local electric field E_{loc}. If we can accurately describe this local electric field, then we know better how the atoms will polarize. This thought problem is represented in figure 3.4

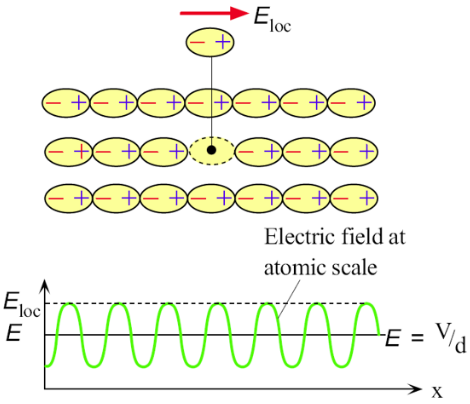


Figure 3.4: The local electric field is complicated. To compute the electronic response of an atom to the applied field, we need to understand the local electric field. To compute the local electric field, we can consider the thought experiment where an atom or molecule is removed. Measuring the electric field in the location where the atom was from the atoms and fields around it can therefore determine the local electric field. Adapted from Kasap

Figure 3.4 describes the thought problem that can be used to determine the local electric field. Here, we need to consider both the voltage applied by the capacitor plates but the additional field induced by nearby atoms. This can be an exceedingly complicated problem to solve, but the key point is that we need to understand the *local* electric field in order to determine the macroscopic response. This can be solved in more advanced texts, and techniques like Density Functional Theory can help predict this. One solution we will present without evidence is the solution for liquids and materials with a cubic crystal structure, in which the local electric field can be represented as:

$$E_{loc} = E + \frac{1}{3\epsilon_0} P \tag{3.27}$$

In these systems, we can therefore represent the induced polarization as:

$$p_{induced} = \alpha_e E_{loc} \tag{3.28}$$

Combined with the expression for the macroscopic susceptibility $P = \chi_e \epsilon_0 E = N p_{induced}$, we can find a relationship for the permittivity of these materials based on the polarizability of the atoms in the crystal. The induced dipole in this equation therefore relates the local electric field to the macroscopic polarization. Using these equations, we can relate the macroscopic permittivity to the atomic polarizability as:

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N\alpha_e}{3\epsilon_0} \tag{3.29}$$

Which is known as the *Clausius-Mossotti* equation. This fundamentally links macroscopic material properties ϵ_r to microscopic details in α_e.

3.6 Visualizing polarization in different kinds of materials

This has been fairly abstract. This section aims to look at how different classes of materials will polarize in response to an applied electric field. First, let's look at figure 3.5, which shows the electron clouds displacing in response to the applied electric field. Here, we see that the macroscopic polarization points in the direction of the applied electric field, and the electrons are displaced with respect to the atomic cores.

3.6.1 Covalent solid materials

In covalently bonded solids, the electron clouds for the valence electrons are delocalized with respect to an individual atom. This makes it much easier for these electrons to move with respect to an applied electric field, which leads to a larger induced dipole. We can rationalize this by looking at the binding energy of these electrons to the atomic cores - the valence electrons take much less energy (for the Semiconductor Si, of order 1-3 eV) to break the bond between the electron and atomic core. The core electrons (say, the 1s orbitals) are much harder to remove from the atom. Because of this, the valence electrons respond much more easily than the rest of the electrons. In fact, this is part of the reason for the high relative permittivity of these materials compared to many others. For example, the relative permittivity of Silicon is 11.9, much higher than that of free space (i.e. vacuum).

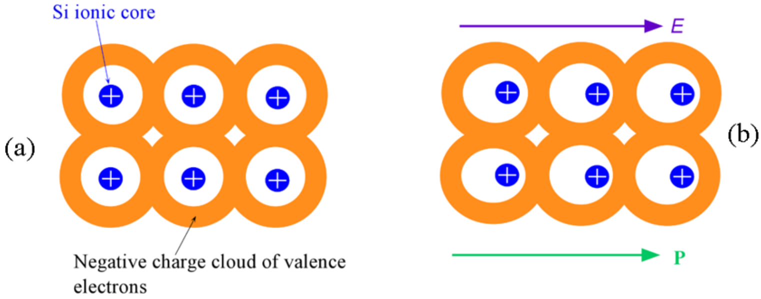


Figure 3.5: The valence electrons inside of a covalently bonded material move in response to the applied electric field. This polarizes the solid, leading to a net polarization. Adapted from Kasap

3.6.2 ionic solid materials

Several materials that are composed of atoms with strongly differing electronegativity are highly ionic crystals. This can include common salts like NaCl. Here, we have a strong localization of electric fields, as each Na atom tends to take on a positive charge characteristic while Cl ions are more strongly negative - that is, we can think of these crystals as having one atom with a predominately negative charge, while the other has a predominately positive charge. Without any applied field, these materials have no net dipole despite this strong localization of charge, due to the crystal symmetry. We can quantify this as a net dipole p_{net} being the sum of dipoles from the positive and negative components as:

$$p_{net} = p_+ - p_- = 0 \tag{3.30}$$

Once a field is applied, the electric field will push the ions themselves in response to the field. This will break the overall symmetry of the crystal and the individual dipoles that are in the solid will no longer cancel. Then, the average dipole in the crystal p_{av} is going to be given by:

$$p_{av} = \alpha_i E_{loc} \tag{3.31}$$

Where we are looking at the local electric field E_{loc} as in previous examples. Here, the ionic polarizability α_i defines how large the average dipole moment will be for a given applied electric field. These can be quite a bit larger than the electronic polarizability α_e in, for example, noble gases. As such, this mechanism can give rise to a reasonably large dielectric constant, with some ionic materials have dielectric constants $\epsilon_r \approx 10$.

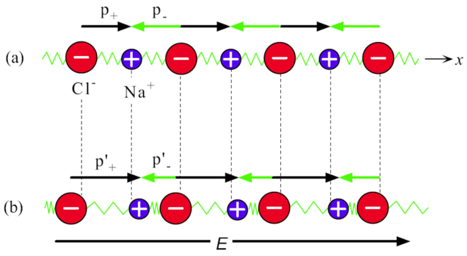


Figure 3.6: Crystals that have predominately ionic bonding will lead to a net zero dipole moment in a centro-symmetric crystal. The application of an electric field will physically move ions around, which induces a net dipole moment that depends on the ionic polarizability of the individual materials.

3.6.3 Orientational polarization

The dielectric constants of liquids can be quite high in general. This is because several molecules, like water, HCl, etc., have a built-in permanent net dipole moment defined by the geometry of the individual molecule. At thermal equilibrium, a liquid like this will have no net dipole moment, as each molecule will be randomly oriented and the net dipole will equal zero. With an applied electric field, this is no longer the case - there is now a force from the electric field that is being applied that will make the molecules re-orient themselves along the direction of the electric field. This is achieved through the electric field applying a torque on molecules that have dipoles that are not perfectly aligned with the applied field. However, this must be balanced by the thermal energy leading to disorder in the liquid. On average, there will be a net dipole, but the molecules will not all be entirely aligned in the direction of the electric field. This is represented in figure 3.7

Computing this in a fully rigorous way is beyond the scope of this course. To bound the potential results, we can consider the limiting cases. If we have a number density

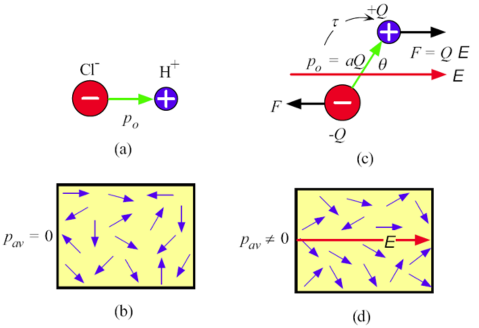


Figure 3.7: A liquid composed of molecules with permanent dipole moments has an average polarization of 0 without an applied field due to thermal fluctuations. An applied electric field will partially align the dipoles of molecules along the direction of the applied field due to a torque applied to the dipole, but thermal energy will lead to some disorder. The net dipole moment is no longer 0. Adapted from Kasap

N of dipoles that each have a dipole moment p_0, then the net polarization P is given as:

$$P = Np_0 \tag{3.32}$$

The net polarization induced will be lower than this due to thermal energy leading to disorder in the orientation of the dipoles. The net effect is called the *orientational polarization.*

The average thermal energy that each molecule has is approximately $5/2k_BT$. This sets the order of magnitude for the applied field to have a meaningful impact on the net polarization of the liquid - if the applied field is much lower than this, then thermal energy will dominate and the collisions of different molecules will keep the liquid mostly randomly oriented. Once the average energy per dipole that the field imparts is comparable or larger than the thermal energy, then we expect a larger proportion of molecules to be aligned with the field, leading to a net dipole moment.

To compute this in general, one would need to compute the average energy that the electric field applies to a molecule that is not in alignment with the field through the torque that would be applied to the dipole. Using the appropriate Boltzmann statistics for this distribution of orientations and the average thermal energy in the system, we would arrive at the following equation for the average dipole moment when the $p_0E < k_BT$:

$$p_{av} = \frac{1}{3}\frac{p_0^2E_{loc}}{k_BT} \tag{3.33}$$

which means that we can define the *dipolar orientational polarizability* α_d per molecule as:

$$\alpha_d = \frac{1}{3}\frac{p_0^2}{k_BT} \tag{3.34}$$

Note that this is a much more temperature-dependent function than any of the other mechanisms we have talked about. This is somewhat intuitive - if the liquid is quite cold, then there is less thermal energy to create molecular collisions that re-orient the dipoles. We may expect, then, that the dipoles will be better aligned with the field in this case, leading to a higher average polarizability.

3.6.4 Interfacial polarization

As a final example we will consider one final case, known as *interfacial polarization*. This will occur whenever there are interfaces between two different materials where charges can accumulate in the dielectric material. This can also happen at defects, grain boundaries, and other crystal defects (or dopants) where charge can accumulate. If there are mobile charges or ions in a material (for example, H+ or Li+ commonly found in oxides or glasses), then the applied electric field can lead to mobility of these ions, which may end up at an interface and accumulate. In response to this, an electrode will counter-act the additional charge by increasing the net charge on that electrode, which appears to have an increase in the dielectric constant of the material. Another case may be when electrons or holes will accumulate at a crystal surface or interface between grains within a material. In general, these phenomena are hard to capture analytically but play important roles when you have significant heterogeneity in the materials. A schematic illustration of this is included in figure 3.8

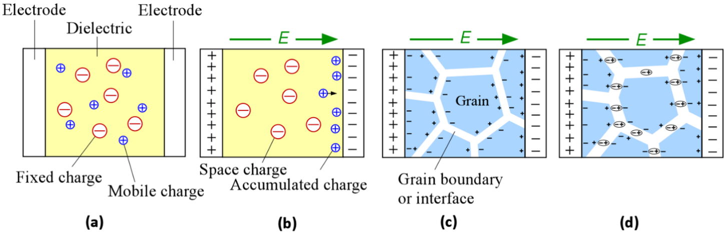


Figure 3.8: Interfaces between dissimilar materials, defects, or grain boundaries, can lead to charge trapping or localization of charges at interfaces. The additional dipole induced by this creates an additional contribution to the polarization vector that leads to an additional contributor to the dielectric constant of a material. Adapted from Kasap

3.7 Putting it all together

We have discussed several mechanisms for the net polarization of a material. In general these are all independent effects, meaning that we can simply sum them to get the net dielectric constant of a material through:

$$p_{av} = \alpha_e E_{loc} + \alpha_i E_{loc} + \alpha_d E_{loc} + ... \tag{3.35}$$

However, note that the inclusion of any dipolar molecules is in general hard to compute, so coming up with a general mapping between that effect and the net permittivity is challenging. If we only include contributions from electronic and ionic polarizabilities, then we can give an approximate value for the permittivity through a modified Clausius-Mossotti relationship:

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{1}{3\epsilon_0}(N_e\alpha_e + N_i\alpha_i) \tag{3.36}$$

3.8 Frequency dependence and losses

As a final section here we will discuss two important details: First, the dielectric constant of a material depends on the frequency with which the applied electric field oscillates - that is, there is a frequency dependence. We will discuss how this is manifested. As a consequence of this frequency dependence we will also find that the dielectric constant is in general a *complex* number. The addition of an imaginary part of the dielectric constant is related to a phenomena known as *dielectric loss*, which can be a limiting factor in the performance of capacitors.

We will consider the general case of two masses with opposite charge separated by some distance, whichis not dissimilar from ions in a crystal. When we apply an oscillating electric field, this will induce a time-dependent dipole moment that may not respond in phase with the oscillating field. We will find that the amplitude of oscillation will depend on the resonant frequency, and that loss will arise when the dipole oscillates 90 degrees out of phase with the driving field.

We will define this system to have a *reduced mass* $M_r = \frac{m_1+m_2}{m_1 m_2}$ and a bond strength β that has a Hooke's-law type restoring force $F_{restore} = -\beta x$.

We previously discussed in this chapter that the bonding in an atom of material can lead to a resonant frequency, in the manner of a mass on a string. We simplified that problem slight, as we neglected physical damping mechanisms, either through coupling to other degrees of freedom (like phonon scattering), or from the fact that an accelerating charge must radiate energy. If we assume that we have a time harmonic electric field of the form:

$$F = QE(t) = QE_0 \exp{(j\omega t)} \tag{3.37}$$

we can observe the amplitude and phase that the dipoles will respond with. Here, Q is the charge of the atoms or electrons, E_0 is the amplitude of the applied electric field, j is the imaginary number (electrical engineers prefer using j rather than i - the author of this set of course notes strongly prefers the opposite, but we will be consistent with the guiding textbook), and ω is the frequency of oscillation. If we further assume that thereis some form of damping through phonon scattering, re-radiation, etc., which takes the form of:

$$-\gamma\left(\frac{dx}{dt}\right) \tag{3.38}$$

where γ is the loss constant and x is the position variable. We previously saw that an atom or dipole that has a bond strength β has a resonant frequency ω_0. The net differential equation for the position of an atom is given as:

$$M_r\frac{d^2x}{dt^2} = QE_0\exp(j\omega t) - \beta x - \gamma\frac{dx}{dt} \tag{3.39}$$

Where we have identified that the net force is equal to mass times acceleration (the left side of the equation), which is related to the field inducing a force, the restoring force based on the bonding, and further reduced by losses. We will see this equation in more detail when we discuss the optical properties of materials, but suffice it to say that we can re-write this in the form of:

$$\frac{d^2x}{dt^2} + \gamma_I \frac{dx}{dt} + \omega_I^2 x(t) = \frac{Q}{M_r} E_0 \exp\left(j\omega t\right) \tag{3.40}$$

where we have defined $\gamma_I = \gamma/M_r$ and $\omega_I = \sqrt{\beta/M_r}$. This model represents the movement of ions in response to a driving electric field, which induces microscopic dipoles in the material

the solution to this equation has a much more complex formulation when we include losses and introduces a *frequency-dependent* amplitude. The solution for the position $x(t)$ is defined as:

$$x(t) = \frac{1}{(\omega_I^2 - \omega^2) - j\gamma_I\omega} \frac{QE_0}{M_r} \exp(j\omega t) \tag{3.41}$$

Here, we see that there is a complex number for the amplitude and a frequency dependence in the response of the dipole to an applied electric field. This has significant impact on the dielectric constant and loss.

The net polarization is dependent on the number density of these dipoles and the relative displacement, such that:

$$P(t) = NQx(t) = \frac{NQ^2}{M_r} \frac{1}{(\omega_I^2 - \omega^2) - j\gamma_I\omega} E_0 \exp(j\omega t) = \epsilon_0 \chi E(t) \tag{3.42}$$

Therefore, we expect that the permittivity to depend on frequency. Since there is a frequency-dependent amplitude, we therefore expect that the polarizability of the dipole will also depend on frequency, as the amplitude is directly related to the polarizability. We can determine the dielectric constant/permittivity by omitting the oscillating field from the polarization and express it as follows:

$$\epsilon(\omega) = \epsilon_0\epsilon_r = \epsilon_0(1+\chi) = \epsilon_0 \left[1 + \frac{NQ^2}{M_r} \frac{1}{(\omega_I^2 - \omega^2) - j\gamma_I\omega}\right] \tag{3.43}$$

We can more explicitly see the real and imaginary parts of this equation if we multiply the the numerator and denomenator by the complex conjugate $(\omega_I^2 - \omega^2) + i\gamma\omega$ such that:

$$\epsilon_r = 1 + \frac{NQ^2}{M_r\epsilon_0} \frac{\omega_I^2 - \omega^2}{(\omega_I^2 - \omega^2)^2 + \gamma_I^2\omega^2} + \frac{NQ^2}{M_r\epsilon_0} \frac{i\gamma_I\omega}{(\omega_I^2 - \omega^2)^2 + \gamma_I^2\omega^2} \tag{3.44}$$

Where we find that the first two terms are entirely real while the last term is entirely imaginary.

Note that this is a very general phenomena and is applicable to many, if not all, of the mechanisms we have talked about so far. For example, the ionic contribution to the dielectric constant will have an associated resonant frequency that will be different that the covalent bonding contribution, etc.

The real part of this equation describes the amplitude of the dipole in response to the driving field that is in phase with the oscillating field, and describes the same permittivity we have discussed previously. The imaginary part of this expression describes the out of phase response, which is proportional to the amount of absorption of energy (at optical frequencies, this would be the absorption of light, for example). Note that the imaginary part is mostly localized around the resonant frequency. Like any other driven oscillator, the amplitude of response is enhanced near the resonance. All of this is manifested in figure 3.9:

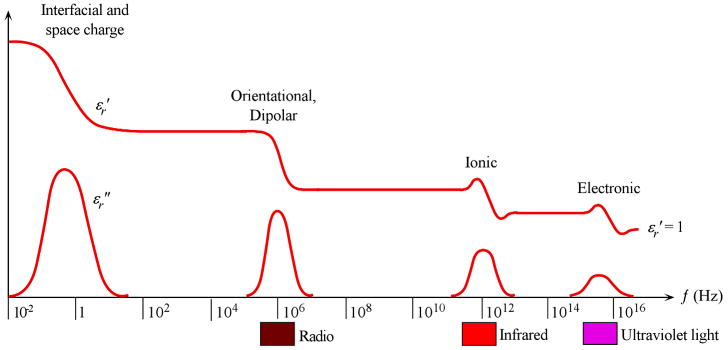


Figure 3.9: Frequency dependence of the dielectric constant. Each contributor to the dielectric constant has a different resonant frequency that depends on the mechanism (ionic movement, electronic movement, etc.), and therefore the dielectric constant is a strong function of frequency. Adapted from Kasap

As a final note, one may find from the plot that the dielectric constant tends to be the highest at low frequencies. This is because each contributing factor is occurring “in phase” with the driving field. As the frequency increases, several of these mechanisms will “freeze” out, and their contribution will decrease. For example, the movement of ions is in general much slower than the movement of electrons, as their masses are much higher. We would therefore expect that the resonant frequency for ionic movement to be lower than electronic movement. If the oscillating field is at a frequency much higher than the resonant frequency, then the atomic motion simply cannot keep up with the driving field, so the ability for ionic movement to induce a dipole or to polarize is reduced. Therefore, the contribution to the net dielectric constant is reduced.

Complementarily, we find that the absorption is localized primarily around the resonant frequency of the oscillation.

3.9 Advanced dielectrics: Piezoelectric, pyroelectric, and ferroelectric materials

For many materials (and all of the materials we have discussed so far), the thermodynamic equilibrium structure has no net dipole in the crystal structure, and all of the dielectric behavior has been due to an electric field modifying the local polarization. This electric field induces a polarization in the material, which we measure as the dielectric constant (or relative permittivity). However, this is not quite the entire picture. Materials that lack inversion symmetry (i.e. are non-centrosymmetric) lack a symmetry that makes the prior description true. That is to say, one can show that a material with inversion symmetry *must* not have a permanent dipole without an applied field.

We can show this schematically by looking at a material with inversion symmetry, like NaCl. NaCl crystallizes into the rocksalt crystal structure. While there are positive and negative charges (as the crystal has strongly ionic bonding) that are spatially arrayed, the net dipole in each unit cell is 0 because of inversion symmetry - that is, the average charge in the unit cell is net 0 because the average position of the positive charges overlaps the average position of the negative charge at the center of the unit cell. If we apply a stress to this material, the crystal will deform. If we do the same analysis, we find that *even with* the deformed structure, the inversion symmetry still requires that the net dipole moment of the unit cell is zero. This is all visualized in figure 3.10:

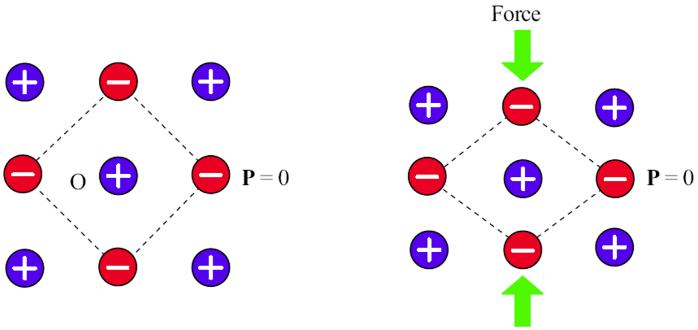


Figure 3.10: The rockstalt crystal NaCl under applied stress. Regardless of how much stress is applied, the crystal structure still has inversion symmetry and therefore the net dipole within the crystal is 0. Adapted from Kasap

Materials that lack inversion symmetry do not necessarily follow this rule. More generally, the lack of inversion symmetry means that the local dipole can be created through other means. Some important classes of materials rely on this extensively. For exampe, a *piezoelectric* material can generate an electric field due to the application of stress. Conversely, an electric field applied to these materials can induce a strain. A few classic examples of these materials include Barium Titanate and Quartz. There are several classes of devices that rely on this effect, including quartz crystal monitors, strain sensors, etc.

The way in which these technologies work can be understood through the crystal structure. In the case of quartz, we see that there is a hexagonal lattice of silicon and oxygen atoms in figure 3.11. Without any applied stress, the relative charges all line up at the center of the hexagon, so the net dipole moment and polarization is still 0. However, if the crystal is deformed due to a force, the hexagon is distorted. The result of this is that the relative position of the positive charges no longer overlaps the net negative charge at the origin. The distance between these two average charges acts as a dipol source, where a net polarization is induced due to the stress. Note, however, that this is a complex effect that can depend on the orientation of the applied stress. In fact, the orientation of the induced polarization does not have to be along the direction of the applied stress.

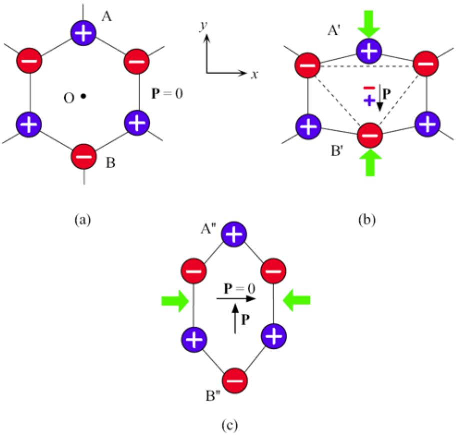


Figure 3.11: Quartz as a piezoelectric material. The crystal structure is hexagonal but lacks inversion symmetry due to the differing charges on the lattice sites. Without an applied stress or force, the net dipole of the structure is still 0. Once a force is applied that deforms the crystal structure, a net dipole forms due to the lack of centrosymmetry. The orientation of the polarization is not necessarily in the direction of the applied force. Adapted from Kasap

This piezoelectric effect is defined through tensors, which can make the mathematics more complicated than we will discuss in this course. In general, we can define the stress T (an unfortunate naming convention to use the same variable as temperature) in the j direction as T_j. For a piezoelectric crystals, we can describe the induced polarization P along the i direction P_i through the following relationship:

$$P_i = d_{ij} T_j \tag{3.45}$$

where d_{ij} is the *piezoelectric coefficient*. Note that the subscript on the piezoelectric coefficient maps the stress in the j direction to a polarization in the i direction. In general, this means that **d** is the *Piezoelectric tensor*, while **P** and **T** are vectors related to the crystallographic directions. Depending on the symmetry of the crystal, some elements in the piezoelectric tensor will be 0. One can use the analysis indicated above to determine which elements will be zero - if you deform the crystal along a direction corresponding to the induced strain, you can determine whether the charges will lead to an induced dipole element. If they do not (like in the rock-salt example), then the element will be zero. These results are tabulated in various textbooks and can be used when designing a material. For example, a particular device you are trying to create may require that you specify a crystal orientation in order to make the device function the way that you would like it.

Figure 3.12 shows some potential use cases for the material using the piezoelectric effect (above). This effect is reciprocal in the sense that you can interchange whether you are measuring or applying an electric field to induce a change to the

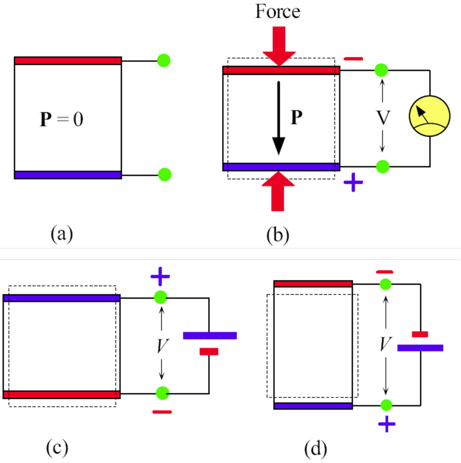


Figure 3.12: The piezoelectric effect can generate a voltage due to the buildup of charges on the surface of a piezoelectric material from applied stress. This is due to the induced dipole polarizing the material. Conversely, the converse piezoelectric effect can use an applied electric field to strain the material. Adapted from Kasap

material. The *Converse piezoelectric effect* can be used to convert electrical energy to mechanical energy by using a voltage to induce a strain in a material (and therefore strain it). We can map this through the same piezoelectric tensor as follows: If we apply an electric field E in the i direction E_i, the piezoelectric coefficient d_{ij} will induce a strain S in the j direction S_j such that:

$$S_j = d_{ij} E_i \tag{3.46}$$

This means that we can use electric fields to deform materials, which can be used in actuators, etc. People working in this field may define a “coupling factor” k that measures this. Here, k^2 is the ratio of the total electrical energy converted to mechanical energy divided by the total energy put into the system.

These materials can also use heat to change polarization, and vice versa. These materials are said to be *pyroelectric*. Here, we can relate the change in polarization a material may experience as it changes temperature through the *pyroelectric coefficient* p (again, unfortunate naming convention) through:

$$p = \frac{dP}{dT} \tag{3.47}$$

This can be used in a variety of sensors - for example, in the far infrared or in another place where the material will absorb ambient heat. These can be extraordinarily sensitive, and can therefore be used when precision is important. For example, a cryostat could use something like this. This is shown schematically in figure 3.13:

Finally, some materials naturally contain an oriented dipole at thermodynamic equilibrium. These are called *ferroelectric* materials. Below a critical temperature, the

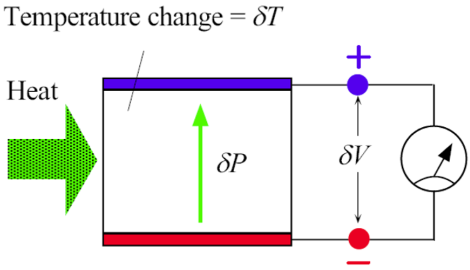


Figure 3.13: The pyroelectric effect maps changing temperatures to changing polarizations through the pyroelectric coefficient. Adapted from Kasap

configuration of atoms is preferred that contain a permanent dipole. This can be due to an atom existing within an interstitial site, but slightly out of the center of the site. Materials like $LiNbO_3$, $BaTiO_3$, and others are ferroelectric. The temperature above which these materials are no longer ferroelectric is called the *Curie Temperature* T_C. A description of this effect is shown in figure 3.14:

This polarization depends on the position of the atoms within the lattice and can be locally switched through a process known as “poling”. A strong enough voltage can give enough energy to the material to move the Titanium atom down to flip the orientation of this polarization.

We should note that the permanent polarization means that we have to express the permittivity of this material slightly differently, as the built in field means we cannot use $P = \epsilon_0(\epsilon_r - 1)E$ to describe the behavior of this material (because P is not 0 when $E = 0$. We instead have to note that there is a built in charge that any electrode would see even with 0 voltage due to the dipole moment. Then, we can define the dielectric constant as:

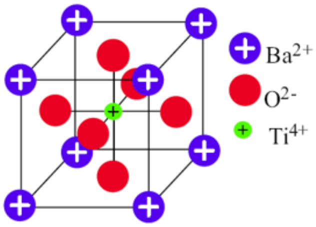
$$\Delta P = \epsilon_0(\epsilon_r - 1)\Delta E \tag{3.48}$$

where the *change* in the polarization due to a change in the electric field is manifested through the dielectric constant.

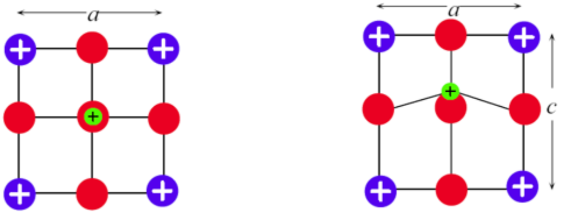
We should additionallynote that all ferroelectric materials are piezoelectric, but not all piezoelectric materials are ferroelectric (for example, Quartz is not ferroelectric). This broken symmetry has important consequences in the utility of these materials beyond the scope of this course.

3.10 The Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET)

The final section of this chapter will combine aspects of the previous chapter to show how combining all classes of electronic materials (metal, semiconductor, and insulator) into one composite device can manifest unintuitive operation that nevertheless underpins all of modern electronics.



(a) $BaTiO_3$ cubic crystal structure above 130 °C



(b) $BaTiO_3$ cubic structure above 130 °C (c) $BaTiO_3$ tetragonal structure below 130 °C

Figure 3.14: The ferroelectric $BaTiO_3$ unit cell. Here, a Titanium 4+ atom exists in the center of the perovskite unit cell. Above the curie temperature (here, 130 C), the crystal is cubic and the net dipole of the unit cell is 0 beause the material is centrosymmetric. Below this temperature, the crystal becomes tetragonal and the titanium atom is displaced slightly compared to the center of the unit cell. This displacement leads to a permanent dipole orientation. Adapted from Kasap

The key point here can help clarify the movement of charge inside of different materials, if possible. This then dictates how the electric field will penetrate into a material, as the electric field is directly tied to the charge density inside of a region of material. First, we can go back to the capacitor, in which two metal plates are separated by some distance. When connected in a circuit and a voltage difference is applied, the metal will create a density of charge at the surface of each plate. We can plot this charge density as a function of distance across the device, and we see that the positive and negative charge densities appear in equal proportion at the edges of the plates. This is shown in figure 3.15:

We can see that the charge is localized to the surface. This is because the concentration of charges that are required to accumulate to screen the field is much lower than the concentration of free carriers in the material. For example, if we apply a 2V potential difference between plates that are separated by 100 nm. The plates are 1 cm x 1 cm. The capacitance of these plates is 8.85 nF, which corresponds to a charge of $Q = 1.77 \times 10^{-8}$C, which corresponds to approximately 1.11×10^{11} electrons on the plate. We can estimate the concentration of electrons in the metal. For a good conductor, approximately one free electron from each atom is available for conduction. The surface density is then approximately $\approx 10^{15}$, as $n_{surface} \approx n_{bulk}^{2/3}$. This is clearly much greater than the required concentration, so only some of the free electrons on the surface need to move in response to the voltage difference. Because of this, there is no disruption or

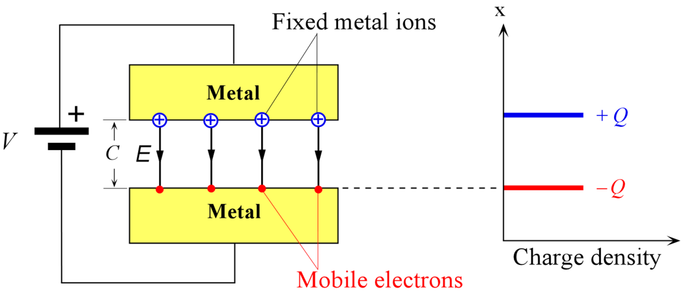


Figure 3.15: A metal-dielectric-metal capacitor. Here, the dielectric is air. When a voltage difference is applied between the plates, charges accumulate on the surface that are localized within approximately one unit cell of the material due to the high charge density in the metal. Adapted from Kasap

change in charge density within the bulk of the metals, meaning that there is no electric field within the metal - it is entirely within the gap between the metal plates.

Now let us consider a p-type semiconductor with a relatively low concentration of acceptors $\approx 10^{15}\text{cm}^{-3}$. If we make the bottom electrode this semiconductor instead of the metal, we can do the same computation. We would find that the surface density of free carriers at the surface is only $\approx 1 \times 10^{10}$ electrons, which is much lower than the number required for the applied voltage. Because of this, a volume of material in the semiconductor must be used to create the necessary charge density. This means that an electric field will penetrate into the semiconductor. This is represented in figure 3.16.

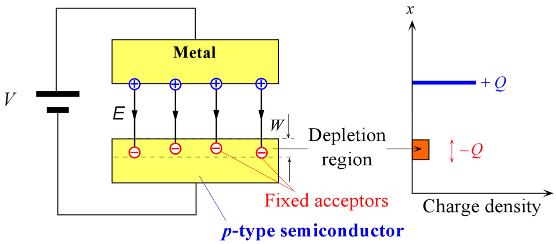


Figure 3.16: A metal-dielectric-semiconductor capacitor. When a voltage difference is applied between the plates, the metal can screen the field within the first unit cell. The density of freely moving carriers in the semiconductor is too low, so the net charge density takes place within a certain distance into the metal. Accordingly, the electric field will penetrate into the depletion region. Adapted from Kasap

In this case, the free holes are pushed away from the surface of the semiconductor, leaving behind the stationary negatively charged acceptor dopants. This creates a *depletion region* within the semiconductor due to the induced electric field. The total charge within the semiconductor in this region must equal the charge on the metal plate, which means that we can determine the depletion width W as:

$$eAWN_a = Q \tag{3.49}$$

where Q is the charge on the metal plate, A is the area of the plate, N_a is the acceptor dopant concentration. In this case, the depletion width that is formed is approximately 1μm. As the applied voltage difference increases, we expect that the depletion width will additionally increase, as the charge buildup on the plate will increase. Note that the semiconductor is still p-type in this depletion region, as the concentration of holes will still be greater than the concentration of electrons.

As the voltage increases, it will eventually become more favorable to acquire electrons at the surface rather than just extending the depletion region deeper into the semiconductor. These electrons may be attracted from the bulk where they were the minority charge carrier. Thermal generation within the depletion region can additionally create the charge carriers can create the required carriers. When this occurs, the additional electron concentration leads to a narrow region where the type of semiconductor has inverted - it is now an n-type semiconductor rather than a p-type semiconductor. As the voltage increases further, the concentration of electrons in the inversion layer increases, rather than increasing the depletion region in the p-type semiconductor. This is shown schematically in figure 3.17:

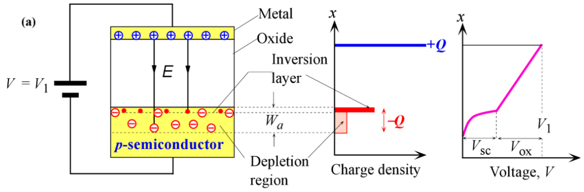


Figure 3.17: A metal-dielectric-semiconductor capacitor leading to inversion. When a voltage difference is applied between the plates, the metal can screen the field within the first unit cell. The density of freely moving carriers in the semiconductor is too low, so the net charge density takes place within a certain distance into the metal. Accordingly, the electric field will penetrate into the depletion region. As the field increases, a thin layer at the surface of the semiconductor will invert to become an n-type semiconductor, with a high concentration of carriers in a thin layer. Adapted from Kasap

This field effect can be very important when creating devices, as it allows a composite device to switch semiconductor type with an externally applied electric field. This will be used in this section to create a transistor.

We will consider the enhancement MOSFET. This structure is shown schematically in figure 3.18. Here, we have a p-type semiconductor in the bulk of the material. By locally n-type doping the semiconductor, some regions will become n-type. At the interface between the n-type and p-type semiconductor, a depletion region will form. We can then deposit metal electrodes to contact the n-type regions, from which we may want charge to flow from one end to the other. In the middle, we place an insulator on top of the p-type material, and put a gate electrode on top of this.

When the gate voltage is not turned on, we have a pnp transistor. No matter what direction we apply an electric field, we find that one of the pn or np sections will be reverse biased, which means that charge does not easily flow through this circuit. We can utilize the field effect from the gate to modulate this. Then, for a set voltage across the source and drain electrodes, we can modify the current that flows through by using this field effect. When we apply a voltage on the gate, we form a depletion region in the semiconductor. This does not help to enable conduction across the region. However, we find that when the gate voltage is above a threshold voltage V_{th} (here, 4V), inversion of the semiconductor under the gate will occur, creating a thin channel of n-type material. This will connect the source and drain terminals of the device, meaning that the material can easily conduct. We see then that the current flow can be turned on in this device through the application of a gate voltage. In this case, we expect that the current-voltage response will be linear, limited only by the resistance in the channel. We can express this as:

$$I_D = \frac{V_{DS}}{R_{n-ch}} \tag{3.50}$$

This linear conduction is not possible forever, because the voltage that is being applied across the source and drain will additionally modify the overall voltage in the inversion region. The voltage V_{DS} applied between the source and drain will lead to a current I_{DS}. The voltage at the drain from this voltage difference is V_{DS}, while the voltage is 0 at the source. The applied voltage in the channel is related to both the source-drain voltage and the gate voltage. In this device, there is a voltage difference between the gate and source V_{GS} and the source and drain V_{DS}. We expect the voltage in the channel next to the source side is then given as V_{GS}. However, the voltage in the channel on the drain side is given as:

$$V_{GD} = V_{GS} - V_{DS} \tag{3.51}$$

In order for there to be a continuous channel from the source to the drain, we require that this voltage be greater than the threshold voltage - that is,

$$V_{GD} \geq V_{th} \tag{3.52}$$

As we can see, this means that the condition for inversion will now depend on the source-drain voltage. If the source-drain voltage increases such that $V_{GD} < V_{th}$, then we do not have the condition for inversion at the end of the channel. This phenomena is known as “pinch-off”, as the channel will no longer connect the source and drain at this point. We then expect that the width of the inversion layer will decrease and end before the drain. We can look at how this impacts the current in figure 3.19. The n-type region will exist only in areas of the channel that have the voltage requirements for inversion. The channel will be pinched off at a point P within the channel that fulfills the requirement:

$$V_{GD} = V_{GS} - V_{DS(sat)} = V_{th} \tag{3.53}$$

The associated current-voltage relationship is no longer linear in this regime. When the drift electrons move through the channel, they will reach this pinch-off point.

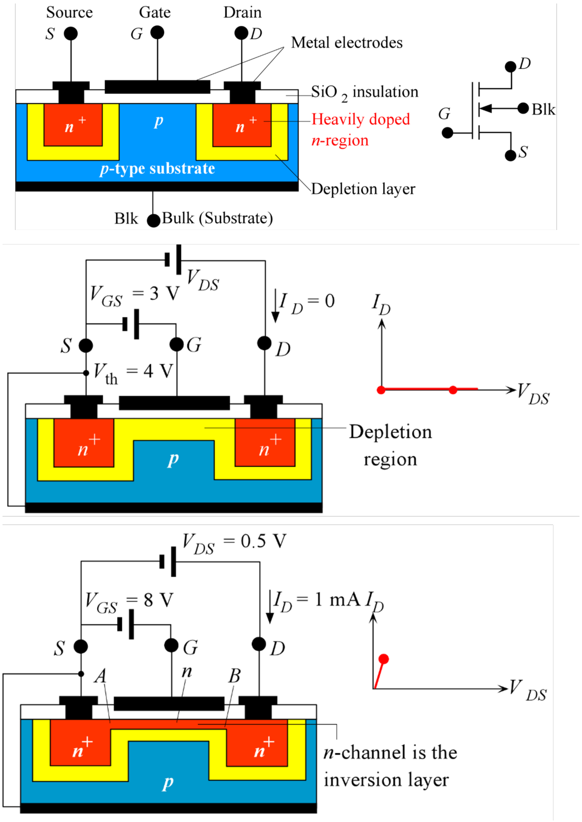


Figure 3.18: An n-MOSFET. Source and drain electrodes are connected to heavily doped n-type materials, which are separated by p-type material. A gate electrode on top of a dielectric uses the field effect to switch the channel semiconductor from p-type to n-type, which can switch the conduction mechanism from rectifying to ohmic. On the right is a standard way that a circuit diagram may represent a transistor, with S, G, and D representing the source, drain, and ground. The back of the substrate is grounded (Bulk, Blk). Once the gate voltage is greater than the threshold voltage V_{th}, semiconductor inversion can enable conduction through the channel. Adapted from Kasap

The strong electric fields in this region will sweep them into the drain, so we still expect some conduction. The current is limited by the ability for electrons in the channel to reach the pinch-off point, which means that the current will still be limited by the channel resistance. As the source-drain voltage increases, we expect that the pinch-off point will move closer to the source. The length of the channel is long enough that typically it does not move significantly, and the resistance in the channel does not change significantly. Therefore, we can assume that the drain current is approximately saturated near this pinch-off point and is given as:

$$I_D \approx I_{DS} \approx \frac{V_{DS(sat)}}{R_{AP'n-ch}} \tag{3.54}$$

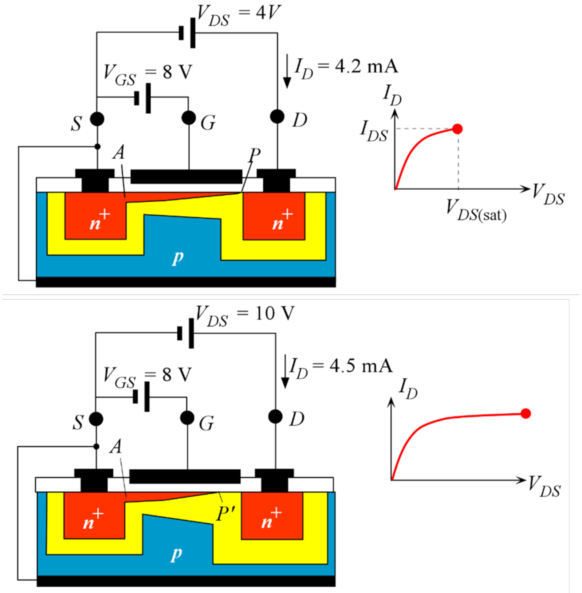


Figure 3.19: Pinch-off and IV curves. Once the voltage at the drain electrode drops below the threshold voltage, the condition for inversion is not fulfilled. We then expect that the channel will revert back to p-type at this point, such that the channel shape now looks different and the channel will "pinch off" from the drain. We observe a turnover on the IV curve, as the relationship is no longer linear. As the source-drain voltage increases, the pinch-off point moves further way from the drain, and the current-voltage relationship is nearly constant. Adapted from Kasap

when $V_{DS} > V_{DS(sat)}$.

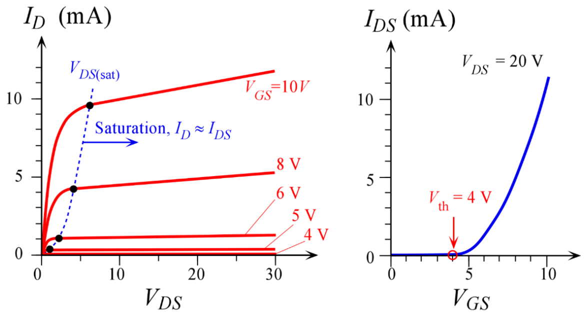
We can represent the current I_{DS} and the gate voltage V_{GS} (notably, when $V_{DS} > V_{DS(sat)}$) through an experimentally derived equation that shows a quadratic relationship as:

$$I_{DS} = K(V_{GS} - V_{th})^2(1 + \lambda V_{DS}) \tag{3.55}$$

where λ is called the "Early voltage" that takes into account the linear relationship of the I-V characteristic at higher voltages. K is the enhancement constant for an ideal MOSFET and is expressed as:

$$K = \frac{Z\mu_e\epsilon_0\epsilon_r}{2Lt_{ox}} \tag{3.56}$$

where Z is the channel width, L is the channel length, μ_e is the electron drift mobility in the channel, ϵ_r is the permittivity of the dielectric material in the gate region, and t_{ox} is the thickness of the dielectric. The typical current-voltage relationships for these devices is plotted in figure 3.20:



(a) (b)

Figure 3.20: IV curves. Experimentally derived trends in the current-voltage relationships are represented in this figure and match the equations above. The source-drain saturation voltage $V_{DS,sat}$ follows a quadratic relationship with the current as defined above. Here, the threshold voltage is 4 V, and the gate voltage V_{GS} is varied. The dependence of the source-drain urrent I_{DS} as a function of the gate voltage V_{GS} for a specific source-drain voltage $V_{DS} > V_{DS,sat}$ Adapted from Kasap

4 Optical properties of materials

4.1 Introduction

The dielectric properties of materials have allowed us to understand the properties of insulators. This showed us that the key material property is the relative permittivity ϵ_r, which is a measure of the ability of a material to polarize in response to an applied electric field. We saw that there were different timescales associated with different polarization mechanisms, which has led to a very complicated frequency dependence of the dielectric properties. This section will introduce the optical properties of materials, which essentially takes this behavior to high frequencies. We will show that the optical properties are essentially the high frequency response of the relative permittivity, and we will show the relationship between this fundamental material property and measurables, including absorption. There will also be a discussion of the role that conduction electrons will have.

NOTE: The definitions for a wave are varied among disciplines. notably, you will find that Kasap will deviate from this derivation because they are electrical engineers and will use $-j$ instead of i as the imaginary number. This has no consequence on the underlying physics, but there will be minus signs that will be *different* than the ones in the Kasap book. Please be careful when looking at different references.

4.2 Maxwell's equations and the wave equation without charges

This class is not long enough to cover Maxwell's equations in detail. We will simply write down the equations and state that a wave equation can arise from it. Maxwell's equations are:

$$\nabla \cdot \mathbf{D} = \rho \tag{4.1}$$

which is Gauss's law relating the displacement field $\mathbf{D} = \epsilon_0 \epsilon_r \mathbf{E}$ to the charge density ρ.

$$\nabla \cdot \mathbf{B} = 0 \tag{4.2}$$

which is Gauss's law for magnetic fields, relating the magnetic field $\mathbf{B}/\mu_0 = \mathbf{H}$ to the magnetic charge density (if you can find a magnetic monopole you can collect a Nobel Prize),

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{4.3}$$

which is Faraday's law that relates a time-varying magnetic field to the generation of an electric field. Finally, we have Ampere's law that shows the opposite - a time varying electric field induces a magnetic field:

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \epsilon_r \frac{\partial \mathbf{E}}{\partial t} \right) \tag{4.4}$$

For the majority of this section we will assume that the current density $\mathbf{J} = 0$. This is not true necessarily in a conductive material, and this can in fact act as a source term that can be important in certain frequency ranges. We will leave this to another course.

We can show that, by combining the two curl equations, a wave equation naturally arises. We will not derive this but simply state the final results, which can be represented for both an electric field and for a magnetic field:

$$\nabla^2 \mathbf{E} = \mu_0 \mu_r \epsilon_0 \epsilon_r \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{4.5}$$

and

$$\nabla^2 \mathbf{B} = \mu_0 \mu_r \epsilon_0 \epsilon_r \frac{\partial^2 \mathbf{B}}{\partial t^2} \tag{4.6}$$

As you can already see, the entirety of the material properties are contained within the relative permittivity ϵ_r and relative permeability μ_r. We will discuss magnetism later, but take for granted now that, at optical frequencies, $\mu_r \approx 1$ for most materials (certainly all the ones we will be discussing). Additionally, we will define $\mu = \mu_0 \mu_r$ and $\epsilon = \epsilon_r \epsilon_0$ for convenience.

These equations have the same general form of a wave equation, which is defined very generally in one dimension as:

$$\frac{\partial^2 f}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 f}{\partial t^2} \tag{4.7}$$

Where the function $f(x,t)$ has a general solution of:

$$f(x,t) = g(x - vt) \tag{4.8}$$

where the function will move with a velocity v that is given as the prefactors in the temporal derivative. From this we can immediately deduce some useful aspects about the wave equation for light, which is that:

$$v = \frac{1}{\mu\epsilon} \tag{4.9}$$

In vacuum, there is no medium to polarize so the permeability and permittivity are simply given as that of free space such that:

$$v = \frac{1}{\epsilon_0\mu_0} = c \tag{4.10}$$

where the speed of light is uniquely defined by these constants (wow!) In a medium, we can also see that the velocity v will decrease, as the values for the permittivity will go up. We can define the *index of refraction* n as the ratio between the speed of light in a vacuum and win the material of interest and determine the material properties that govern this as:

$$n = \frac{c}{v} = \sqrt{\frac{\mu\epsilon}{\mu_0\epsilon_0}} = \sqrt{\epsilon_r\mu_r} \approx \sqrt{\epsilon_r(\omega)} \tag{4.11}$$

where we have assumed that there is not a magnetic response at optical frequencies and we have explicitly noted that the dielectric constant or permittivity is a function of frequency.

A more advanced treatment of maxwell's equations will provide the necessary derivations, but we will present facts below that are related to the wave light will propagate. First, we note that light is an *electromagnetic* wave, as seen above - there is both an electric field and a magnetic field.

A solution to the wave equation is that light waves can propagate as *plane waves* that have the following functional form:

$$\mathbf{E} = E_0 \exp\left[(i(\mathbf{k}\cdot\mathbf{r} - \omega t)\right]\hat{\mathbf{n}} \tag{4.12}$$

This is a complex exponential equation that has a wavevector $\mathbf{k}$, the position $\mathbf{r}$, the frequency of the light ω and the time variable. $\hat{\mathbf{n}}$ is a unit vector that points in the direction of the electric field. There is additionally an amplitude E_0. Note that the bold text indicates that these are in fact vector quantities, such that the unit vector $\hat{\mathbf{n}}$ is the unit vector pointing in the direction of the electric field. The magnetic field can be described similarly and is defined as:

$$\mathbf{B} = B_0 \exp\left[(i(\mathbf{k}\cdot\mathbf{r} - \omega t)\right](\hat{\mathbf{k}} \times \hat{\mathbf{n}}) = \frac{1}{v}\hat{\mathbf{k}} \times \mathbf{E} \tag{4.13}$$

where we see that the magnetic field must be perpendicular to both the electric field and the propagation direction defined by the wavevector. In fact, for all cubic materials (and amorphous materials), one can show that the wavevector $\mathbf{k}$, the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$ must all be mutually orthogonal (and they follow the right hand rule).

The wavevector $\mathbf{k}$ contains all of the material properties and dictates how light will propagate, be absorbed, be refracted, or scattered from a material in general. We can write down the argument of the propagating wave as $kx - \omega t$ (if we assume that the wave is propagating in one direction). From this, we can show that this is equal to $k(x - vt)$, where

$$v(\omega) = \frac{\omega}{k} \tag{4.14}$$

is the *phase velocity*, which describes how quickly this argument term oscillates (essentially how quickly the wave acquires phase). This value is in general frequency dependent. We know that the phase velocity is related to the optical properties of the materials, and we can write down that:

$$k = \frac{\omega}{v(\omega)} = \omega\sqrt{\epsilon\mu} = k^2 \rightarrow k^2 = \omega^2\epsilon\mu \tag{4.15}$$

Note that this applies without any conduction (i.e. not for metals or including conduction electrons). If the material has an isotropic refractive index (like liquids, amorphous material, or cubic materials), we can relate this to the refractive index and the speed of light as:

$$k = \frac{\omega}{v} = \frac{\omega}{\frac{c}{n}} = \frac{\omega n}{c} \tag{4.16}$$

Note that the wavevector itself is defined as:

$$k = \frac{2\pi}{\lambda} \tag{4.17}$$

If the material has a higher refractive index than that of free space (usually the case at optical frequencies), this means that the effective wavelength inside of a material is modified inside of the medium (that is, the light wave acquires phase faster inside of the material in a given distance, such that the wavelength looks shortened). If λ_0 is the free-space wavelength and k_0 is the wavevector in free space, this means that:

$$k = k_0 n \rightarrow k = \frac{2\pi n}{\lambda_0} = \frac{2\pi}{\frac{\lambda_0}{n}} = \frac{2\pi}{\lambda} \tag{4.18}$$

where

$$\lambda = \lambda_0/n \tag{4.19}$$

which shows that the effective wavelength inside of a material is reduced.

With all of this, we have defined the relevant aspects of the wavevector of a material. It is in general a vector quantity that can point in the direction of propagation. In addition, we note that we have 3 coordinates in the vectors (as we live in a world with three spatial directions). If we specify the propagation direction (say, the x direction), we can set the electric field to be in the y or z direction (or any linear combination of the two), and we can always find the orthogonal direction for the magnetic field. The fact that there are still 2 degrees of freedom for the electric field to be placed means that we can have a couple of different *polarization* directions for the electric field. Here, Polarization will be defined as the direction the electric field is oscillating in (for example, the $\pm y$ direction. The field of polarization optics is quite important but generally out of the scope of this class. Noting that there are two different polarizations and that they reflect or refract off of materials with slightly different amplitudes and phases (in general) is essentially the scope of this course. One may note, however, that there is a rich set of materials and systems where polarization is required - for example, polarizing sunglasses.

We can represent these details in figure 4.1:

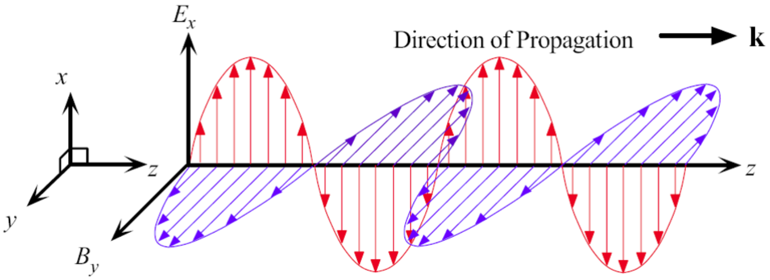


Figure 4.1: Coordinate system of a propagating wave. The wavevector k is along the x direction, and one choice for the polarization of the electric field is defined here. The magnetic field is thus perpendicular to both. Adapted from Kasap

4.3 propagation and absorption

Since all of the ways in which a material will impact the propagation of light are contained within the permittivity, we must try to understand the role that this complex number will have on these properties. First, we then need to re-define the wavevector as $k \rightarrow k' + ik^"$, where k' is the real part and $k^"$ is the imaginary part. Here, $k' = |k|$ describes the propagation of light, while the imaginary part $k^"$ describes the attentuation of light through absorption. If we define the wavevector in this way, we can immediately see that this is the case. For a wave propagating in $\hat{x}$ and for a y-polarized light wave, we then have:

$$E(x,t) = E_0 \exp\left[i(k' + ik^")x - i\omega t\right] \hat{y} \tag{4.20}$$

Grouping terms and noting that $i^2 = -1$ we get:

$$E(x,t) = E_0 \exp(-k^" x) \exp\left[i(k'x - \omega t)\right] \tag{4.21}$$

For a non-zero $k^"$, this means that there will be an exponential decay on the wave. The intensity of a light wave is proportion to the modulus squared (that is, $I \propto |E|^2$) such that:

$$I = I_0 \exp(-\alpha x) \propto exp(-2k^" x) \tag{4.22}$$

This way we can immediately determine a relatoinship between the absorption coefficient α and the imaginary part of the refractive index!

Let's dig into this. First, noting that the refractive index is now a complex value:

$$n \rightarrow n + i\kappa = \sqrt{\epsilon_r} = \sqrt{\epsilon_r' + i\epsilon_r^"} \tag{4.23}$$

Here, we can also define the relationship between the loss κ and the wavevector as:

$$n = \frac{k'}{k_0}, \kappa = \frac{k^"}{k_0} \tag{4.24}$$

and from this we can additionally relate the refractive indices and permittivity differently:

$$n^2 - \kappa^2 + i2n\kappa = \epsilon_r' + i\epsilon_r^{''} \tag{4.25}$$

Therefore, we can see that the propagation of light is impacted by the imaginary part as well (as the real part of the permittivity is related to the imaginary part of the index). From all of this we can determine that:

$$\alpha = 2k^{''} = 2k_0\kappa = \frac{2k_0\epsilon_r^{''}}{2n} = \frac{2\pi}{\lambda}\frac{\epsilon_r^{''}}{n} = \left(\frac{\omega}{c}\right)\frac{\epsilon_r^{''}}{n} \tag{4.26}$$

4.4 Electromagnetic waves in the presence of free charges or currents

If we have conduction electrons or currents in our material, we must modify the wave equations to now be:

$$\nabla^2 E = \mu\epsilon\frac{\partial^2\mathbf{E}}{\partial t^2} + \mu\sigma\frac{\partial\mathbf{E}}{\partial t} \tag{4.27}$$

and

$$\nabla^2 B = \mu\epsilon\frac{\partial^2\mathbf{B}}{\partial t^2} + \mu\sigma\frac{\partial\mathbf{B}}{\partial t} \tag{4.28}$$

where we have a term that is related to the first derivative of the electric or magnetic field and the conductivity σ. We still have plane wave solutions, but the wavevector is now modified to include another term:

$$k^2 = \mu\epsilon\omega^2 + i\mu\sigma\omega \tag{4.29}$$

Where an additional imaginary component related to the conductivity of the material. First we will ignore the frequency-dependence of the conductivity (and therefore σ will become σ_0 to get:

$$\frac{k^2}{\omega^2} = \mu\epsilon + \frac{i\mu\sigma}{\omega} \tag{4.30}$$

such that

$$\frac{c^2}{\omega^2/k^2} = \frac{\mu\epsilon}{\mu_0\epsilon_0} + \frac{i\sigma}{\omega\epsilon_0}\frac{\mu}{\mu_0} \tag{4.31}$$

From all of this, we can see that the permittivity is now modified as:

$$\epsilon = \epsilon_r' + i\epsilon_r^{''} = \mu_r\epsilon_r + i\frac{\sigma_0}{\epsilon_0}\frac{1}{\omega} \tag{4.32}$$

Which shows that free charges induce an additional component of loss. From this we can determine the absorption of a material and its relationship to the DC conductivity (i.e. free carriers) of a material:

$$\alpha = \frac{\omega}{c}\frac{\epsilon_r^{''}}{n} = \frac{\omega}{c}\frac{\sigma}{\epsilon_0}\frac{1}{\omega}\frac{1}{n} = \frac{\sigma}{c\epsilon_0 n} \tag{4.33}$$

This is still an approximation, as free carriers will in principle depend on frequency. Electrons in a material will have a scattering time τ_s within the material, which leads to a frequency-dependent conductivity. Whereas the DC conductivity of a material is related to fundamental constants through:

$$\sigma = \frac{ne^2\tau}{m_e^*} = \sigma_0 \tag{4.34}$$

The AC conductivity instead has a time dependence as:

$$\sigma_{AC} = \frac{\sigma_0}{1 - i\omega\tau_s} = \sigma' + i\sigma^{"} \tag{4.35}$$

This imaginary part shows that the response of the conductivity is out of phase with the driving field at high frequencies, which then is related to loss and absorption. We can define the components of the conductivity as:

$$\sigma' = \frac{\sigma_0}{1 + \omega^2\tau^2} \tag{4.36}$$

and

$$\sigma^{"} = \frac{\sigma_0\omega\tau}{1 + \omega^2\tau^2} \tag{4.37}$$

Such that:

$$\epsilon_r'' = \frac{\sigma_0}{\epsilon_0\omega(1 + \omega^2\tau^2)} \tag{4.38}$$

For a semiconductor, we can then back out an estimated absorption coefficient from this as:

$$\alpha \approx \left(\frac{\sigma_0}{cn\epsilon_0\tau^2}\right)\frac{1}{\omega^2} \tag{4.39}$$

Note, this is the generic response of free carriers. The influence of the band structure of semiconductors is not obvious here, which we will discuss in greater detail as needed. Indirect bandgap semiconductors, for example, will have a different frequency response above their band gap than direct bandgap semiconductors.

Finally, we should note that the imaginary part of the mass on a spring model is very close to zero away from resonance. If you are close to the bond resonant frequency, this will additionally lead to absorption that needs to be represented in the above analysis.

5 Magnetic properties of materials

5.1 Introduction

This section of the course will have a short introduction to magnetism and magnetic materials. This is a rich area of physics that a section of this size in this course does not do great justice to. The goal of this section is to justify the existence of magnetism in materials; the characterize some of the larger classes of magnetic materials; to identify some of the underlying physics of ferromagnetic materials; and to understand magnetic domains as it may relate to useful technologies. This will not give us enough time to discuss superconductors or other associated advanced materials where magnetic effects are interesting.

5.2 Definitions and origins of magnetic fields

First we will write down one of Maxwell's equations that can describe the generation of magnetic fields from currents. Ignoring any time dependence, we can write down that:

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \tag{5.1}$$

where **J** is an electric current density. The cross product guarantees that the magnetic field is perpendicular to the applied field. You may recall from introductory physics that a wire loop with a current running through it will induce a magnetic field perpendicular to the loop, with the field lines inside of the loop uniform and straight lines. This can also manifest as a circular magnetic field that arises from a current in a wire, as indicates in figure 5.1: Conversely, a charged particle will feel the Lorentz force as:

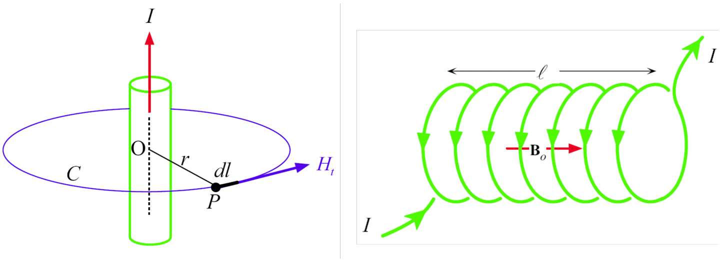


Figure 5.1: Ampere's law, where a current loop can induce a uniform magnetic field. Conversely, current in a straight wire can form a magnetic field that circulates the electric field.

$$F = q\mathbf{v} \times \mathbf{B} \tag{5.2}$$

where a charged particle can be deflected with a magnetic field. The equivalent form for a magnetic field formed from a moving charge is that:

$$\mathbf{B} = \frac{\mu_0}{4\pi} \frac{q\mathbf{v} \times \mathbf{r}}{r^2} \tag{5.3}$$

These fields follow the right hand rule, as indicated in figure 5.2:

We will define the various terms used throughout this chapter here, and show the analogy to the dielectric properties of a material. We will define the *Flux Density* **B**

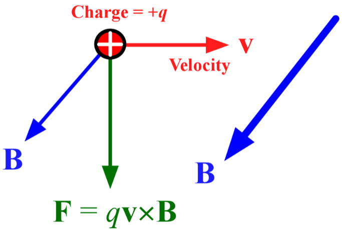


Figure 5.2: The Lorentz force: A moving charged particle experiences a force from a magnetic field perpendicular to the field.

as a function of the *Magnetizing field* **H** and relate it to the *magnetic permeability* μ and the *magnetic susceptibility* χ_m:

$$\mathbf{B} = \mu_0\mathbf{H} + \mu_0\mathbf{M} = \mu_0\mathbf{H} + \mu_0\chi_m\mathbf{H} = \mu_0(1+\chi_m)\mathbf{H} = \mu_0\mu_r\mathbf{H} \tag{5.4}$$

Here, we can see that the form of the equation is similar to that of the displacement field, where we had the $\mathbf{D} = \epsilon\mathbf{E}$. This formally arises from the magnetic susceptibility which, like the electric susceptibility, describes the material response to an external field. Here, the magnetic susceptibilty $\chi_m = \mathbf{M}/\mathbf{H}$ can be positive or negative, in contrast to what we have talked about in the dielectric response of materials (note, however, that this is not always strictly true). In analogy with the dielectric properites of materials, we showed that the macroscopic polarization **P** arose from a summation of the microscopic electric dipoles **p** in a material from an applied electric field. Here, the magnetization **M** arises from the microscopic magnetic dipoles μ_m in a material. One distinct difference is that there is no equivalent to a magnetic charge (a magnetic monopole) like electric fields may have. The magnetic dipole moment can be defined as:

$$\mu_m = IA\hat{\mathbf{u}}_n \tag{5.5}$$

where we note that the dipole moment in general is a vector quantity.

5.2.1 Orbital angular moment and magnetic dipole moments

To motivate the existence of a magnetic field for matter, we can take this understanding an apply it quasi-classically to the atomic level. in the same way that a current in a loop will generate a magnetic field, we can consider the classical approach of an electron orbiting a nucleus that would then induce a dipole that depends on the frequency the electron will orbit the nucleus. The microscopic current can be represented as $I = -e/\tau$, where τ is the period of the orbit. We can then relate this to the orbiting frequency ω as $I = -e\omega/(2\pi)$. From this, we can define the *orbital* magnetic dipole moment μ_{orb} as:

$$\mu_{orb} = -\frac{e\omega r^2}{2} = \left(-\frac{e}{2m_e}\right)L_{orb} \tag{5.6}$$

noting that the orbital angular momentum $L_{orb} = mvr = m\omega r^2$. Note that the component in the parentheses in the above equation is known as the *gyromagnetic ratio*. This orbital dipole moment is visualized in figure 5.3, which additionally shows the definition of a magnetic dipole:

Of course, we know that this model of an electron is incorrect, and a more full quantum mechanical analysis is required to compute the actual L_{orb}. In general, these values are non-zero but tends to be small.

5.3 electron spin magnetic moments

The spin of an electron is also a source of magnetic dipole moments and is based on the intrinsic angular momentum of these particles that have a magnitude of 1/2.

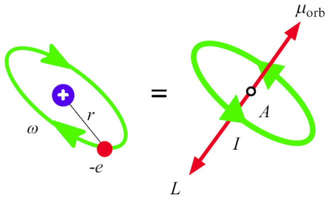


Figure 5.3: An illustration of the magnetic field induced from an orbiting electron around an atom, and the associated representation of the magnetic moment that comes from this.

This intrinsic angular momentum will can create a spin magnetic moment μ_{spin} as:

$$\mu_{spin} = -\frac{e}{m_e}S \tag{5.7}$$

such that when there is a magnetic field applied in the z direction the dipole moment can be represented as:

$$\mu_z = -\frac{e}{m_e}S_z = \frac{-e\hbar}{2m_e} \tag{5.8}$$

which is known as the *Bohr Magneton* and is equal to 9.27×10^{-24} Am^2 (units can also be represented as J/T, with T being tesla). We should note that the spin in the z direction S_z is related to the spinas $S_z = m_s\hbar$, where m_s is the spin and takes the value of 1/2 for an electron.
This electron spin will repond in a similar way to the orbital angular momentum, and the definitions of a magnetic dipole moment that is generated from an applied magnetic field on the dipole is indicated in figure 5.4:

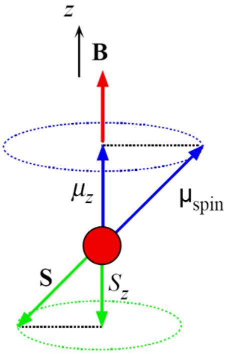
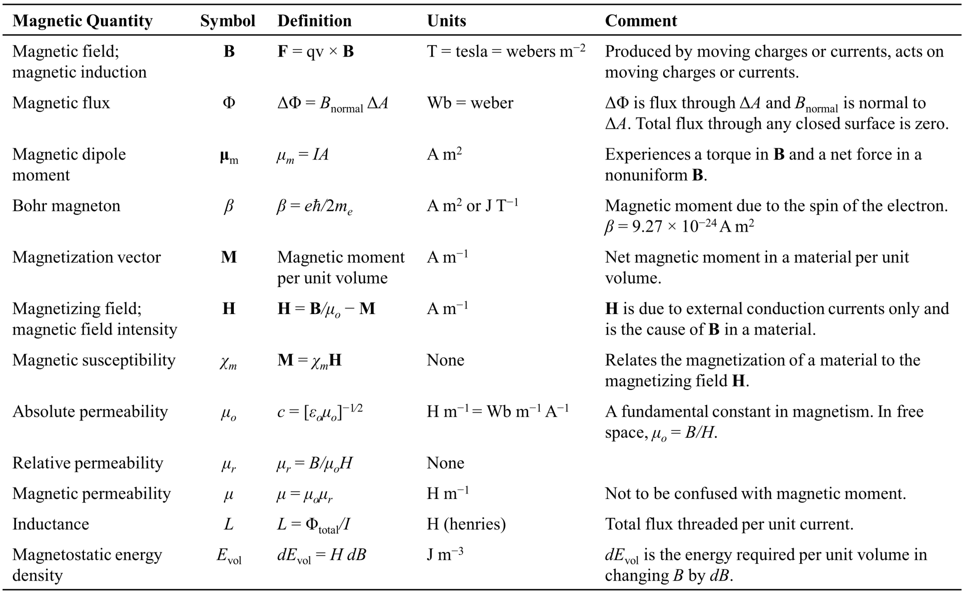


Figure 5.4: An illustration of the induced dipole moment in the z direction from a magnetic field acting on an electron with a spin magnetic dipole. The associated spin, etc. are indicated.

Finally, we should note that the units of these values are all quite complicated. The units of magnetization **M** are in units of A/m, which can be thought of as Am^2/volume, which is essentially a measure of the dipole moment per unit volume. The magnetic flux density **B** has units of J/(Am^2) which is the unit of Tesla T and has units of energy/dipole moment. The relevant units and properites are included in the following table 5.5:



Magnetic Quantity	Symbol	Definition	Units	Comment
Magnetic field; magnetic induction	$\mathbf{B}$	$\mathbf{F} = q\mathrm{v} \times \mathbf{B}$	T = tesla = webers m^{-2}	Produced by moving charges or currents, acts on moving charges or currents.
Magnetic flux	Φ	$\Delta\Phi = B_{\mathrm{normal}}\,\Delta A$	Wb = weber	$\Delta\Phi$ is flux through ΔA and B_{normal} is normal to ΔA. Total flux through any closed surface is zero.
Magnetic dipole moment	$\boldsymbol{\mu}_{\mathrm{m}}$	$\mu_m = IA$	A m^2	Experiences a torque in **B** and a net force in a nonuniform **B**.
Bohr magneton	β	$\beta = e\hbar/2m_e$	A m^2 or J T^{-1}	Magnetic moment due to the spin of the electron. $\beta = 9.27 \times 10^{-24}$ A m^2
Magnetization vector	$\mathbf{M}$	Magnetic moment per unit volume	A m^{-1}	Net magnetic moment in a material per unit volume.
Magnetizing field; magnetic field intensity	$\mathbf{H}$	$\mathbf{H} = \mathbf{B}/\mu_o - \mathbf{M}$	A m^{-1}	**H** is due to external conduction currents only and is the cause of **B** in a material.
Magnetic susceptibility	χ_m	$\mathbf{M} = \chi_m\mathbf{H}$	None	Relates the magnetization of a material to the magnetizing field **H**.
Absolute permeability	μ_o	$c = [\varepsilon_o\mu_o]^{-1/2}$	H m^{-1} = Wb m^{-1} A^{-1}	A fundamental constant in magnetism. In free space, $\mu_o = B/H$.
Relative permeability	μ_r	$\mu_r = B/\mu_o H$	None	
Magnetic permeability	μ	$\mu = \mu_o\mu_r$	H m^{-1}	Not to be confused with magnetic moment.
Inductance	L	$L = \Phi_{\mathrm{total}}/I$	H (henries)	Total flux threaded per unit current.
Magnetostatic energy density	E_{vol}	$dE_{\mathrm{vol}} = H\,dB$	J m^{-3}	dE_{vol} is the energy required per unit volume in changing B by dB.

Figure 5.5: Magnetic quantities and units

5.4 Types of magnetism

In this section, we will highlight the relevant physics and classifications of different responses to a magnetic field. We will start with paramagnetism, then explain diamagnetism, ferromagnetism, and antiferromagnetism. The physics of ferromagnets will be dealt with in depth in a later section.

5.5 Paramagnetism

In dielectric materials, an applied electric field will induce a polarization. In liquids, the applied electric will apply a torque on a dipole that is not aligned with the electric field, which will cause the dipole to align with the field. This is represented in figure 5.6:

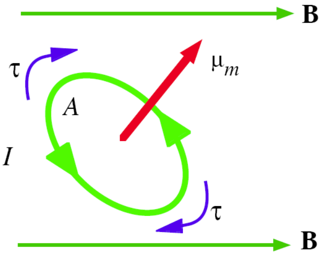


Figure 5.6: An illustration of the torque applied to a magnetic dipole with a magnetic field, which causes a driving force to align it with the field.

A similar thing will occur in paramagnets. With a random set of orientation of spins, a magnetic field will provide a driving force to cause them to align. This is represented schematically in figure 5.7:

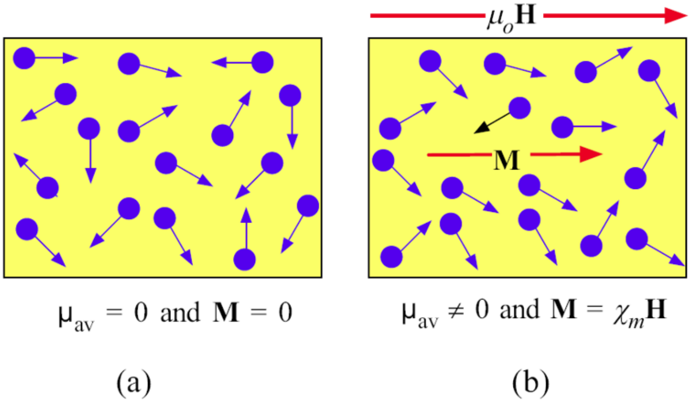


Figure 5.7: A paramagnet will have randomly oriented spins within the material align in response to a magnetic field.

Just like the dipolar response of a dielectric material with orientational permittivity contributions, thermal energy will combat the driving force from the magnetic field. To truly solve for this response, statistical mechanics is required. The result will be introduced here, which shows that the total magnetization $\mathbf{M}$ induced by the magnetic field is defined as:

$$\mathbf{M} = N\mu_m \tanh\left(\frac{\mu B}{k_B T}\right) \tag{5.9}$$

where N is the number density of dipoles, μ_m is the dipole strength, μ is the permeability. As we can see, the argument of the function is the driving force (the applied magnetic flux that favors spin alignment) divided by the thermal energy that favors random orientations. When the temperature is high relative to the applied magnetic field, the net magnetization is low. When the field is high compared to the thermal energy, the magnetization is higher. There is no built-in magnetic field when there is no applied field. This equation works in general, but for low magnetic fields (often the case), we can linearize this to:

$$\overline{\mu_z} = \frac{\mu_m}{3}\left(\frac{\mu B}{k_B T}\right) \tag{5.10}$$

As we can see from this equation, the magnetization is linear with the applied field (for low field) but generally positive when the magnetic field is positive. This implies that the susceptibility χ_m is:

$$\chi_m = \frac{M}{H} > 0 \tag{5.11}$$

Metals can exhibit paramagnetism. We can visualize this through the electronic bands of the materials. Conduction electrons will sit in the band structure with either positive or negative spin ($m_s = \pm 1/2$) and there is no preference for spin. When a magnetic field is applied (here, for example, in the z direction), this is not necessarily the case anymore. With the additional magnetic field, spins that are oriented along the direction of the applied field is lower than spins that take an opposite spin. In the applied field, all states that have a spin aligned with the field will be lowered in energy by βB_0, where B_0 is the magnetic field and β is the Bohr magneton. Conversely, energy states with a spin that is aligned against the field will have their energy increased by the same amount. Because the system wants to minimize its energy, the states with lower energy (i.e. those aligned with the field) will be filled first. This will lead to a net spin that is aligned with the field, providing a magnetic response. This tends to be weak but non-zero. We can determine the approximate susceptibility from this effect, termed *Pauli spin paramagnetism* as:

$$\chi_{para} \approx \mu_0 \beta^2 g(E_F) \tag{5.12}$$

where the density of states at the fermi energy $g(E_F)$ essentially determines the susceptibility. This effect is visualized in figure 5.8:

5.6 Diamagnetism

In contrast to paramagnetic materials, a diamagnetic material has a *negative* susceptibility that is small. Silicon, for example, has $\chi_m = -5.2 \times 10^{-6}$, which leads to a net relative permability that is slightly lower than 1. This means that an applied electric field $\mu_0 H$ will induce a magnetization M that *opposes* the applied field, which means that the resulting B field in the material is less than the applied

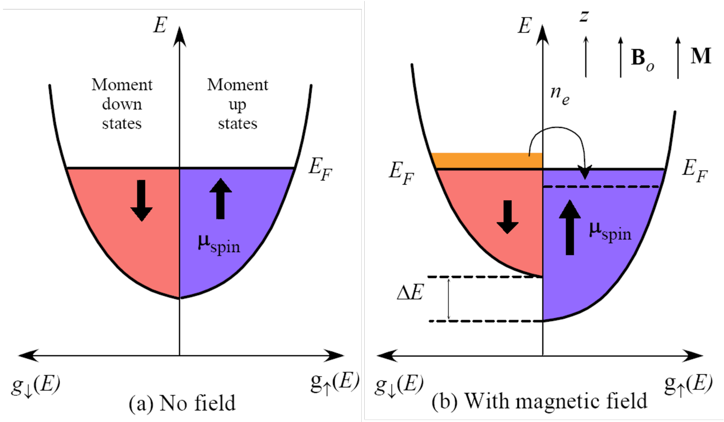


Figure 5.8: Pauli spin paramagnetism in metals. Conduction electrons will lower their energy in an applied magnetic field by finding states with lower energy. Those states are for states with a spin that is aligned with the applied magnetic field. This leads to a net spin in the direction of the applied field, which gives rise to a positive susceptibility

field. Essentially the induced magnetic field from the dipoles in the material is in the opposite direction, and the spins in the material act to try and repel or expel the magnetic field from insite of the material. The net magnetization without an applied field is zero. This will induce a net force that will repel the diamagnet away from the source of the applied magnetic field. This is represented in figure 5.9 We should note that the diamagnetic response is very strong in superconductors, and $\chi_m = -1$ in this case. This is manifested in the Meissner effect, in which a superconductor will expel all magnetic fields within them below the critical temperature and a superconductor will float above a magnet.

5.7 Ferromagnetism, Antiferromagnetism, and Ferrimagnetism

So far, we have discussed materials that do not have a built in magnetic field, like most dielectric materials we talked about. Ferromagnets, Antiferromagnets, and Ferrimagnets are all classes of materials that have a net magnetization without an applied field. The physics of this will be discussed later.

Ferromagnetic materials can have a large permanent magnetization, which implies that χ_m is positive and very large. The relationship between the magnetization $\mathbf{M}$ and applied field $\mu_0\mathbf{H}$ is complicated, nonlinear, and can have hysteresis. At high magnetic fields, the magnetization can saturate. This is an implicitly quantum mechanical. In later sections we will look at the magnetic domains that can occur in these materials that gives rise to this hysteresis, as well as the reasons for which they form. This sort of ordering inside of a crystal has interesting and useful dynamics. One should note that ferromagnetism has a critical temperature T_C above which the material will act as a ferromagnet.

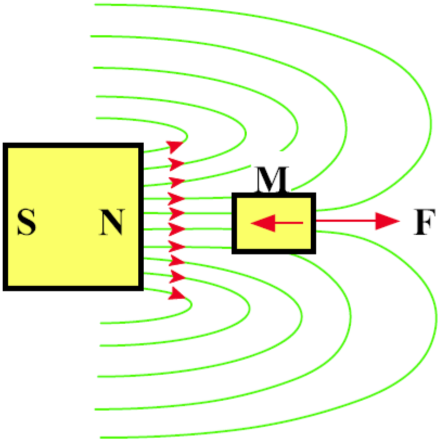


Figure 5.9: A diamagnet will oppose the induced magnetic field and has a negative susceptibility. This will lead to a force that attempts to move the material away from the magnet.

A defining feature of a ferromagnetic material is that (within a domain) all of the spins are aligned in a specific direction that defines the directoin of the magnetization. Conversely, Antiferromagnetic materials (including materials like Cr) have a small but positive susceptibility and do not have any magnetization in the absence of an applied field. This is distinct from a paraelectric for the following reasons. There is strong ordering of spins within the material, but are aligned precisely to cancel out the net magnetization. This essentially means that the spins alternate in orientation such that the net magnetization is zero. These additionally have a critical temperature (called the Néel temperature T_N) above which the material acts like a paramagnet.

Finally, a Ferrimagnetic material (called so because materials like Ferrite, Fe_3O_4) is a behavior that is somewhat in between. There is still ordering within the crystal, and the material acts as a ferromagnet below the curie temperature T_C, but not all of the spins are aligned. Here, a sublattice within the material may be arranged in one orientation while the other sublattice is arranged in a different orientation, but the net spin orientation is not zero. Thus, we still have a net magnetization direction. These materials are often insulators, meaning that they have some interesting applications in electronic materials and devices.

All of this behavior is higlighted in figure 5.10, which gives examples of the ordering required to give specific types of ferromagnetism:

All of this is somewhat complicated, so a table is provided below in order to highlight the differences. Notably, the degree of ordering, a critical temperature, and the sign of the susceptibility can all be used to classify the type of magnetism in these materials.

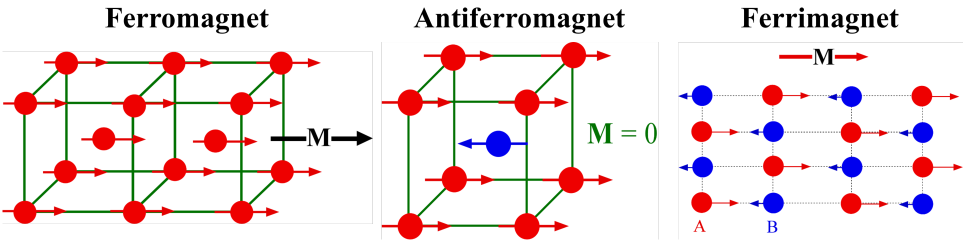


Figure 5.10: Ordering of spins in different materials. When the spins are all aligned in one direction in the absence of an applied field, the material is a ferromagnet (left). Conversely, when the spins are all anti-aligned to cancel out the net spin direction, the material is an antiferromagnet (middle). When the spins have different alignemtns but a net magnetization exists, the material is a ferrimagnet (right). All materials display some level of ordering without an external field, and have a positive susceptibility. They all also have a critical temperature above which they act as a paramagnet.

5.8 Physics of ferromagnetic materials

In this section we will highlight some of the behavior of ferromagnetic materials that can give rise to this specific level or ordering and the associated phase transition to paramagnetism above the critical temperature.

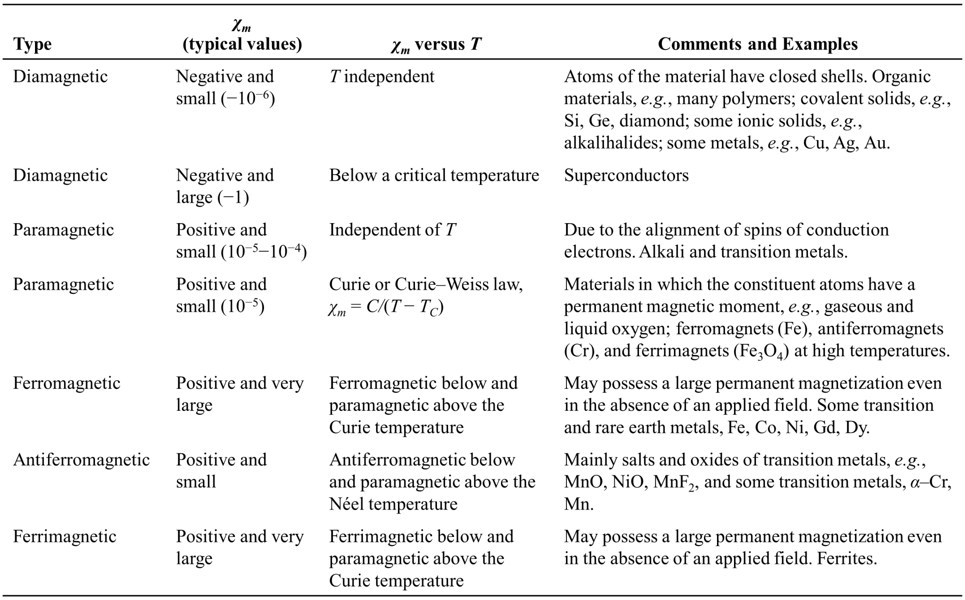
The vast majority of materials do not have an ordering of their spins at room temperature. This level of ordering in a material suggests that there is a driving force to lower the energy of the system to make this the most stable configuration. Iron, Cobalt, and Nickel are all ferromagnetic at room temperature, while some rare earth materials like gadolinium and dysprosium are magnetic below room temperature. Some of the strongest magnets are made with Neodymium.

Consider Iron, which has the electron structure $[\mathrm{Ar}]3\mathrm{d}^6 4\mathrm{s}^2$. Because of Hund's rules, the electrons will be arranged such that each orbital has at least one electron in it to lower its overall energy. The system will try to minimize its overall energy, which means that it must select a configuration of spins and orbitals that both obey Pauli exclusion and minimize the electrostatic interaction energy. Together, these become the *exchange interaction* potential that can create materials in which there are high numbers of unpaired electrons with the same spins. This is shown schematically in figure 5.12:

We should note that, in general, the number of spins that will be aligned on average depends on the strength of this exchange interaction. We can compute the exchange energy that arises from this as:

$$E_{ex} = -2J_e \mathbf{S}_1 \cdot \mathbf{S}_2 \tag{5.13}$$

where J_e is the exchange integral, which integrates the wavefunctions f the potential energy interaction terms for the system. This is a complex calculation. $\mathbf{S}_i$ is the spin angular momentum of electron i. When the sign of J_e is negative, it means that the spins should be antiparallel (which is the case for most elements). For



Type	χ_m (typical values)	χ_m versus T	Comments and Examples
Diamagnetic	Negative and small (-10^{-6})	T independent	Atoms of the material have closed shells. Organic materials, *e.g.*, many polymers; covalent solids, *e.g.*, Si, Ge, diamond; some ionic solids, *e.g.*, alkalihalides; some metals, *e.g.*, Cu, Ag, Au.
Diamagnetic	Negative and large (-1)	Below a critical temperature	Superconductors
Paramagnetic	Positive and small (10^{-5}–10^{-4})	Independent of T	Due to the alignment of spins of conduction electrons. Alkali and transition metals.
Paramagnetic	Positive and small (10^{-5})	Curie or Curie–Weiss law, $\chi_m = C/(T - T_C)$	Materials in which the constituent atoms have a permanent magnetic moment, *e.g.*, gaseous and liquid oxygen; ferromagnets (Fe), antiferromagnets (Cr), and ferrimagnets (Fe_3O_4) at high temperatures.
Ferromagnetic	Positive and very large	Ferromagnetic below and paramagnetic above the Curie temperature	May possess a large permanent magnetization even in the absence of an applied field. Some transition and rare earth metals, Fe, Co, Ni, Gd, Dy.
Antiferromagnetic	Positive and small	Antiferromagnetic below and paramagnetic above the Néel temperature	Mainly salts and oxides of transition metals, *e.g.*, MnO, NiO, MnF_2, and some transition metals, α–Cr, Mn.
Ferrimagnetic	Positive and very large	Ferrimagnetic below and paramagnetic above the Curie temperature	May possess a large permanent magnetization even in the absence of an applied field. Ferrites.

Figure 5.11: A table with all of the relevant behavior of different types of magnetic materials.

magnetic materials with ordering, this corresponds to the antiferromagnetic state (such as Cr). When this exchange integral is positive (like for Fe, Co, Ni, and Gd), the exchange energy is negative if the two electrons are parallel (which would reduce the overall energy of the system). This means that there is an energy term that favors ordering to create a permanent magnetic dipole orientation in some materials. This is shown in figure 5.13.

Since this behavior requires that there is an energy penalty to not align the spins, this would suggest that a phase transition can occur when there is enough thermal energy to overcome this barrier. At low temperature, the maximum magnetization can be achieved when all of the spins are aligned. This value is called M_{sat}. In Iron, this is 2.2 Bohr magnetons, which is also a measure of the average number of electrons that have spins aligned in the same directoin. This means that the magnetic field inside of the material is $\mu_0 M_{sat} = 2.2$ T. However, this is not always the case. As the temperature increases, we can expect that thermal vibrations and thermal energy can make the system more energetic, and some spins will begin to not align in the same directoin. Once the temperature gets to a high enough temperature (the Curie temperature T_C), all of the spins are randomly oriented. This is related to the exchange interaction energy and can be used as a rough guide for the Curie temperature. This would require that $k_B T \approx E_{ex}$. For iron, $E_{ex} = 0.09$ eV, which corresponds to a temperature of about 1000K. The magnetization as a function of temperature is displayed in figure 5.14:

Above the Curie temperature, the susceptibility follows the Curie-Weiss law, which models the susceptibility as:

$$\chi = \frac{C}{T - T_C} \tag{5.14}$$

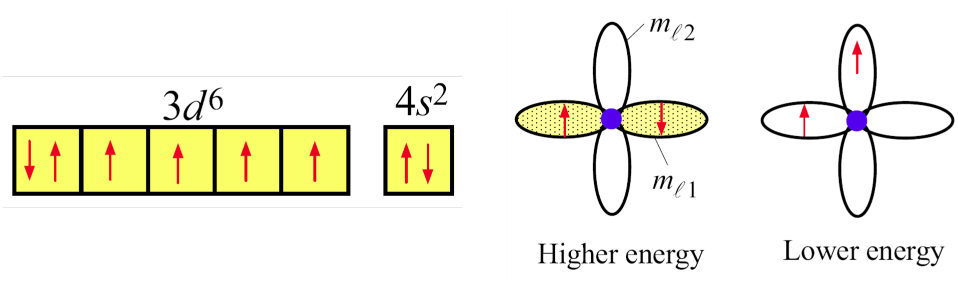


Figure 5.12: Hund's rule in Iron. Left: Electron spins and placement in the outermost orbitals. There is a net of 4 electrons that are spin up, leading to a spin magnetic moment of 4β in the atom. right: Schematic of the orbital energy levels, where a combination of Pauli exclusion and electrostatic interactions of the different orbital states can create situations in which the lowest energy state is to have the spins aligned in the same orientation on different orbitals rather than paired together.

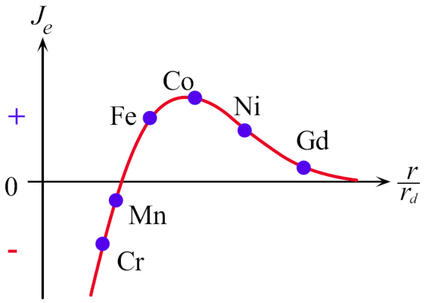


Figure 5.13: The exchange integral J_e calculated for a variety of materials, plotted with the x axis as the radius of the material (r_d is the radius of the d orbitals). A positive value indicates that it is favorable to order the spins. A negative interaction implies the opposite.

where C is a curie constant that depends on the material, and T_C is the Curie temperature.

5.9 Magnetic domains, domain wall movement, and hysteresis

Having discussed the details of different kinds of magnetism, we will focus the remaining time on ferromagnets to understand their behavior. This underpins a variety of devices that are important in modern technology, including speaker systems, engines, magnetic data storage, and materials for electronics.

A ferromagnet can be heated above its Curie temperature to become a paramagnet. Upon cooling, the material will become ferromagnetic, but the average magnetization within the entire volume of material will be zero! This is because, while the spins will align locally, there is no driving force to make all of the spins in the material form into a single direction. There will be a series of magnetic domains, not unlike a grain microstructure in a material. There will additionally be magnetic domain walls at their interfaces. An external magnetic field that aligns these domains

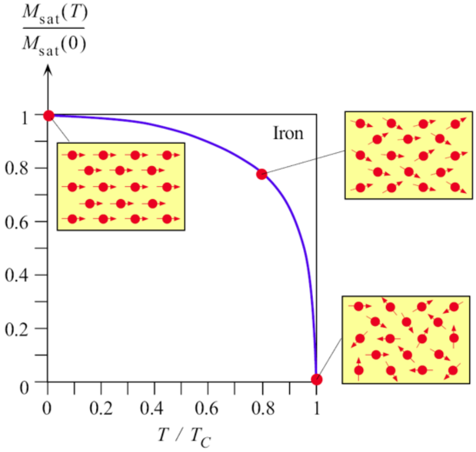


Figure 5.14: Magnetization of a ferromagnet as a function of temperature. As the temperature rises, the average number of spins aligned in a given direction decreases. At the Curie temperature, all of the spins are randomly oriented.

into a single direction will lead to the hysteresis curves observed in experiments. We can rationalize this by considering the energy contained in magnetic fields. If the magnet were to have a single domain upon cooling, then magnetic field lines will exit the material. These can contain potential energy, called the magnetostatic energy. This potential energy can be reduced if the fields outside the material can be reduced. If a domain wall were to form in the material, then there can be regions of opposite magnetization within the material. This will reduce the external magnetic fields, as the magnetic field lines that connect the north and south sides of the magnet do not need to traverse as far. By including domains and domain walls at different angles, the fields outside of the material can be nearly eliminated, reducing the overall energy. While there may be energy costs associated with creating a domain wall (as the domain wall will not have any aligned spins), the overall energy of the system is reduced by including this randomized order. A schematic of this process is indicated in figure 5.15

One should note that the domain walls are not necessarily a single unit cell of the material. The orientation of the spins will flip from the orientation of one domain to the other over a characteristic distance that minimizes the overall energy of the domain wall, but this could be several unit cells in size. This is visualized in figure 5.16:

This has a lot of overlap in the microstructure of many materials, and indeed acts very similarly However, this is a description of the orientation of spins in the material, not necessarily the underlying order. The Curie temperature for a ferroelectric -> paraelectric transition does not need to be at the same temperature as a structural phase transition, but is rather just a transition from an ordered state of spins to a disordered state.

The magnetization of the domains in these materials will be along specific directions, as it may be easier to align the spins in a specific direction of the lattice. This

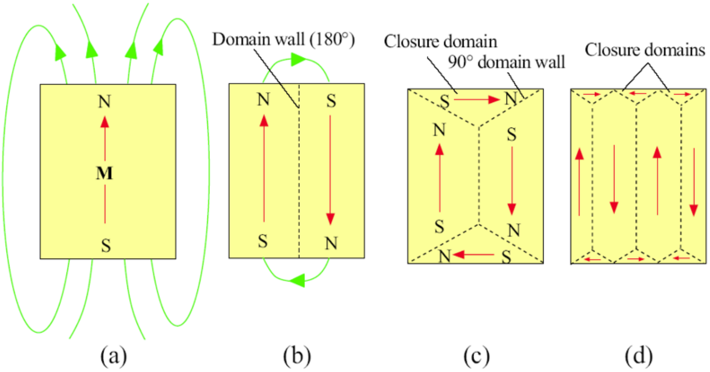


Figure 5.15: Domain wall formation in a ferromagnet. The magnetostatic energy is reduced by forming domains. A material cooled below its curie temperature in the absence of an applied field will have domains.

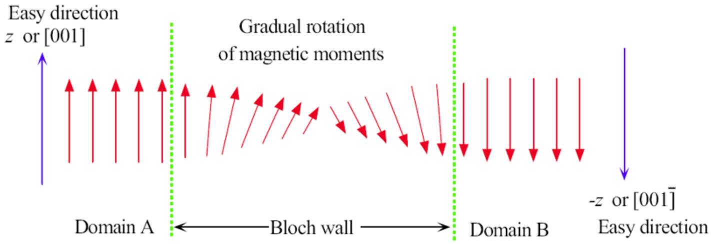


Figure 5.16: A domain wall will gradually rotate the orientation of the magnetic moments as it moves from one domain to another. This is sometimes called a Bloch wall.

can be seen by solving the exchange interaction integral, and can be correlated in experiments. In iron, it is easiest to form domains along the (100) directions, and are therefore called the **easy directions**.

Applying a magnetic field can cause these domains to change in shape and to flip the orientation of ferromagnetic domains. This is because the spins that are aligned with the field have lower overall energy than the spins that are not aligned with the field. As the applied field increases in strength, we may then assume that the domains that are aligned with the field will grow, as the overall energy of the system will be reduced. The spins that are aligned against the orientation of applied field will shrink. Thus, we can imagine that an applied magnetic field can give a net magnetization due to the modification of the domain structure and distribution. This is shown in figure 5.17:

Because the ferromagnetic domain formation arises from the energy reduction for spins to order, it is reasonable to assume that there may be preferred orientations to the spins with respect to the underlying lattice. This is born out in experiments. For example, magnetizing a ferromagnetic sample along the [100] directions will require less energy (a lower applied magnetic field) because it is along the easy direction. If you were to apply the same applied magnetic field along the [111]

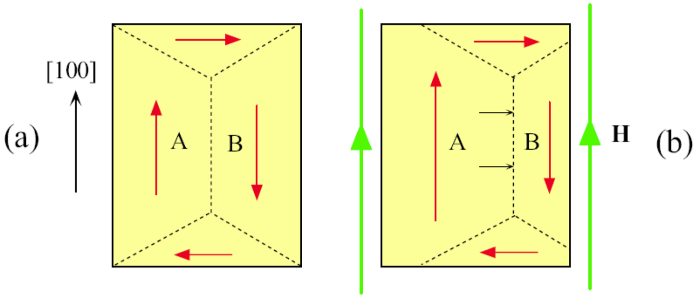


Figure 5.17: A domain wall will move under an applied magnetic field in order to reduce the overall energy of the system. The domains aligned with the field will grow, while the domains aligned opposite of the field will shrink.

direction, you would find that the net magnetization along this direction is lower. It will take more energy to magnetize the material to the same level. This behavior is known as **Magnetocrystalline anisotropy** and is a reflection of the fact that these spins exist in a crystalline material, and their interaction with the underlying lattice can influence their overall behavior. This is not a small value either - in BCC Iron, it required 4 times that amount of field to magnetize the material along the [111] direction compared to the [100]. Some experimental values are shown in figure 5.18:

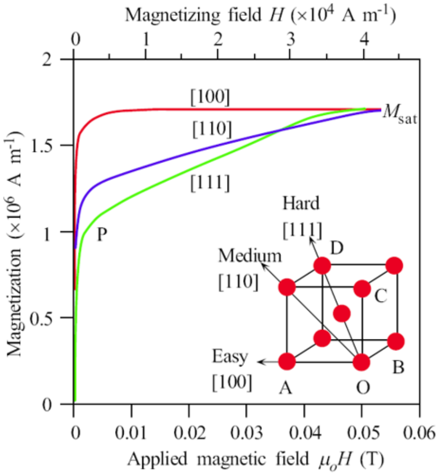


Figure 5.18: Magnetization of Iron along different crystallographic directions, showing a significant different in saturation fields along the easy and hard directions in the crystal.

When a domain wall moves in the crystal, it can be affected by the underlying crystal. For example, any strain in the material will change the interatomic spacing, which then changes the overall electronic environment, as the bond lengths will have changed. This is similar to the piezoelectric effect in some dielectric materials, as a strain can induce an electric field. Conversely, an applied magnetic field can additionally induce a strain on a material (similar to the converse piezoelectric effect). This behavior is known as magnetostriction. This can potentially be used for sensors and other devices, but has ramifications for the behavior of ferromagnetic domains. Since there is now strain energy in the magnetized material, this can impact the behavior and movement of the ferromagnetic domains.

Other aspects of the material can also impact the movement of domains in a material. The local strain environment around dislocations and point defects, for example, can interact with the magnetic field and magnetization. This means that domain walls can be “pinned” to defects, grain boundaries, and other impurities in the crystal material itself.

The result of this is that the microstructure and polycrystalline nature of many magnetic materials can have a large impact on the overall response of these domain walls when an external magnetic field is applied. Even without an applied magnetic field, the domains of the material can have multiple ferromagnetic domains within them. Defect pinning can lead to individual grains that have different domain orientations, whereas sufficiently small grains may have a single magnetic domain. The summary of all of these topics is included in figure 5.19:

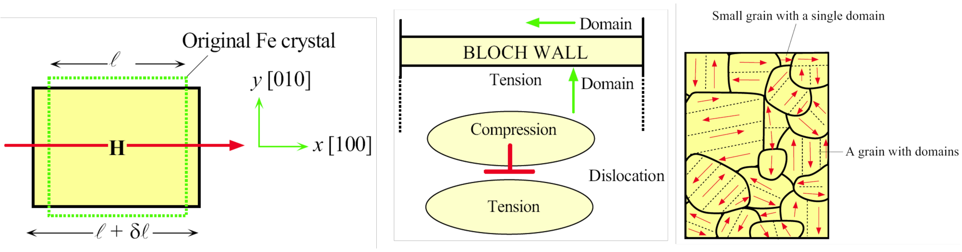


Figure 5.19: Magnetostriction can mean that magnetic fields induce strains in materials, while strains in the material can impact the movement of domain walls. Defects can lead to domain wall pinning, as the overall energy of the system can be lowered at these locations. Domains within grains can also vary, and domain wall pinning at grain boundaries is common.

With all of this together we can begin to understand the evolution of the ferromagnetic domain structure in a material when an external magnetic field is applied. A material with no average magnetization without an applied magnetic field will have a variety of ferromagnetic domain orientations. Measuring the magnetization inside the material as a function of applied magnetic field will sweep out the behavior of this material as it goes from no net magnetization to full magnetized. Applying a magnetic field will first lead to reversible domain wall movement. As the applied field increases, the domain walls will move irreversibly, often in sudden changes that are due to domain walls moving across pinning defects, grain boundaries, dislocations, etc. Domains walls that are pinned to defects or other parts of the material require additional magnetic field to unpin them and move them within the crystal. As the field increases, the material will eventually reduce or eliminate domains that are not aligned with the field. First, the domains will be oriented along the easy directions of the crystal. When the material is fully saturated, the domains will align with the field, whether or not it is along the easy direction in the crystal.

If we were to remove the magnetic field, the magnetization would relax back into one of the easy directions, and some domains may flip orientation. Therefore, we would expect the magnetization at 0 applied magnetic field to be lower than fully saturated. However, we find that the magnetization is still nonzero. The remaining amount of magnetization is called the *remnant magnetization*. Flipping the sign of the applied magnetic field will lead to a decrease in the net magnetization. Eventually the average magnetization will become 0 in the sample for some negative value of the applied magnetic field. The magnetic field required to get to this point is called the *coercive field* or the *coercivity*. This curve shows hysteresis, as the value of the magnetization depends on the previous magnetic field history and does not return to zero when the field is removed. The entire path of magnetic domain structure and applied fields is shown in figure 5.20:

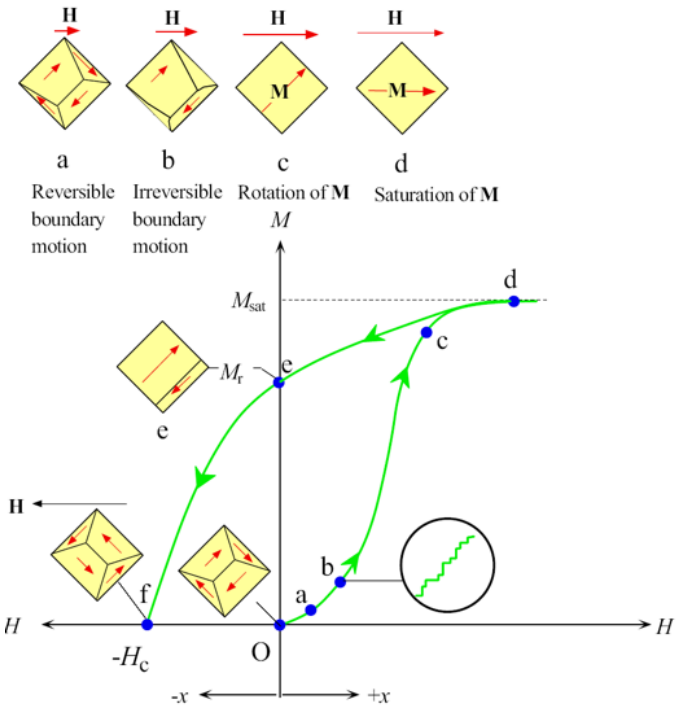


Figure 5.20: M vs. H behavior of a previously unmagnetized polycrystalline iron specimen. An example grain in the unmagnetized specimen is that at O. (a) Under very small fields the domain boundary motion is reversible. (b) The boundary motions are irreversible and occur in sudden jerks. (c) Nearly all the grains are single domains with saturation magnetizations in the easy directions. (d) Magnetizations in individual grains have to be rotated to align with the field, H. (e) When the field is removed the specimen returns along d to e. (f) To demagnetize the specimen we have to apply a magnetizing field of Hc in the reverse direction. Adapted from Kasap.

This hysteresis loop experiment can show a lot of the physical behavior of ferromagnetic materials, and also can contain information about the amount of energy required to magnetize materials. As domain walls move, there is energy required to flip the sign of the spins in the material. In addition, any domain wall movement can create lattice vibrations that will dissipate as heat. The energy loss from a magnetization cycle (such as a path from 0 -> +$\mu_0 H$ -> -H -> 0) can be computed by integrating the area within the hysteresis loop if plotted as B a function of H (rather than M in the previous example). Combined with the coercivity, saturation magnetization, and remnant magnetization, we can begin to categorize materials and use this behavior for a variety of different device applications. For example, we can classify a magnet as a "hard" magnet if there is a large coercitivty, which requires a strong magnetic field to fully magnetize the material. This means that it requires more energy to magnetize (and can be relatively hard to do so), but it also takes a significant amount of energy to demagnetize. These can be used as permanent magnets. These may be used in motors, loudspeakers, receivers, MRIs, and other devices. Conversely, a "soft" magnet will have a relatively low coercivity and will therefore exhibit a stronger slope in a B-H curve. This means that a high magnetization can be achieved with relatively low mapplied magnetic fields. Some materials may include ferrites. These can be used in a variety of power electronic devices, transformers, and other machinery. Some materials, like Silicon iron (2 percent Si), or some Nickel alloys, have low coercivity and low power loss, which make them useful in RF electronics such as in audio devices. Standard M-H and B-H curves, as well as a description of soft and hard magnets, are found in figure 5.21:

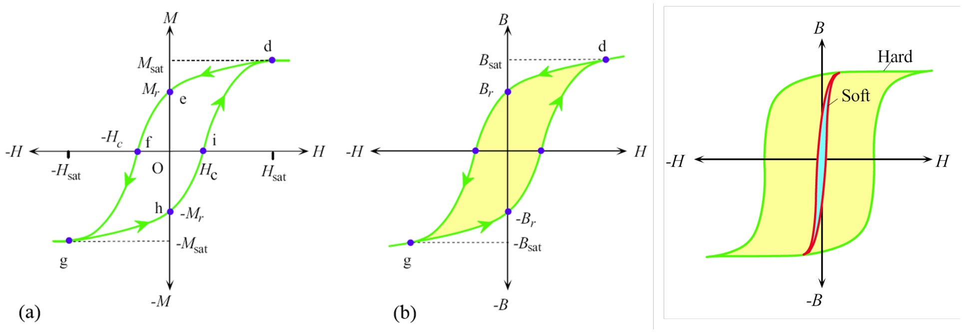


Figure 5.21: M vs. H of a magnet. Adding in the magnetizing field, the y axis can be computed as $B = M + \mu_0 H$ (middle). The area within the hysteresis loop describes the energy loss per cycle. Finally, the relative area, coercive field, etc. can help to classify magnetic materials as "soft" or "hard", and may find different uses in modern technologies. Adapted from Kasap.

6 Thermal properties of materials

6.1 Introduction

In this final section of the course, we will discuss some of the aspects of thermal properties of materials. This, as with other sections, cannot discuss the entirety of thermal transport in solid materials given the time frame of the course. However, some of the key outcomes of this section are to identify that there are analogues to the electronic band structure of materials. We will briefly discuss some of the ways in which thermal energy is transported, but will spend the majority of the time discussing phonons, or quantized lattice vibrations that can carry thermal energy. As bosons (rather than fermions), they have different occupation statistics than electrons. The thermal heat capacity can be estimated from simple models that take into account this quantization, and a brief description of thermal properties will be discussed. Bond strength, crystal structure, and atomic weight all play a role in the ability for materials to carry heat.

6.2 Types of thermal transport

There are broadly 3 different types of thermal transport. They are: Conduction (in solids, most of this class), convection (in gases and liquids, ignored almost entirely in this course), and radiation. We have nominally covered some aspects of radiation in previous portions of this course, where photons can carry energy that can be absorbed as heat. Radiation of photons in the infrared electromagnetic spectrum is often called thermal radiation. In fact, radiation of light and heat from a warm body is one of the critical advances that quantum mechanics helped to answer. This, along with the photoelectric effect, were some of the first demonstrations that quantization is important, and the emergence of particle/wave duality is well represented in these phenomena. Here, we will briefly highlight some of the salient properties of blackbody radiation as it relates to the quantum theory of light (also a boson).

6.2.1 Blackbody radiation

We can say with generality that all objects will absorb and emit electromagnetic radiation. The amount (both in terms of emitted power and the spectral distribution) depend on temperature. in equilibrium, the total amount of energy emitted by the body matches the amount of energy absorbed by its environment. That is, if an object is at the same temperature as its surroundings, then the net power flux into the object will be 0. A perfect blackbody will radiate light with a spectral irradiance I_λ as follows:

$$I_\lambda = \frac{2\pi h c^2}{\lambda^5 \left(\exp\left(\frac{hc}{\lambda k_B T}\right) - 1\right)} \tag{6.1}$$

in which we note that the frequency of a photon (in free space) is defined as $\omega = 2\pi c/\lambda$ such that the exponential takes on the familiar form $\hbar\omega/k_B T$. We may note that the exponential term in parentheses also looks familiar. Recalling that the fermi-dirac distribution is given by $\frac{1}{\exp\left(\frac{E}{k_B T}+1\right)}$, the difference is in the sign in the denominator. We will find that the Bose-Einstein distribution takes on the form in the spectral irradiance, which will become important when discussing the thermal properties of phonons. While beyond the scope of this course, the occupation statistics of bosons are substantially different than electrons (for example, no Pauli exclusion), which is manifested in these particles obeying a different distribution function.

All of this is to say that the derivation of blackbody radiation required the existence of a quantized object called a photon that followed these statistics, and was the only way in which experimental data could be explained. The classical theory based on a modern (at the time) understanding of electromagnetism led to the so-called “ultraviolet catastrophe”, in which the emission of energy of a body at a finite temperature scaled as λ^{-4}, which implied that any object at any temperature would emit more and more light in the ultraviolet. This not only did notmatch experimental measurements, but does not conserve energy (and would additionally imply that everyone and everything would be emitting ultraviolet light, x-rays, and even shorter wavelengths with greater and greater intensity, which is a bit silly).

We can also remark on a few other details of this distribution. First, we can determine the total power that is emitted by a blackbody at a certain temperature. The total power is given as an integral of the spectral irradiance. That is,

$$P = \int_0^\infty I_\lambda d\lambda = \sigma_S T^4 \tag{6.2}$$

where σ_S is the stefan boltzmann constant and is defined as:

$$\sigma_S = \frac{2\pi^5 k_B^4}{15c^2h^3} \tag{6.3}$$

This immediately implies that the power radiated from a warm object increases monotonically (and as a reasonably strong function of) temperature. In addition, one can show that the radiated intensity at each wavelength will additionally increase with temperature - that is, there is never a scenario in which warming up an object will lead to a wavelength being emitted with less intensity. In fact, one can show that the maximum wavelength emitted by a warm body will change as a function of temperature. This is known as Wien’s displacement law and is given as:

$$\lambda_{max} T \approx 2.98 \times 10^{-3} m \cdot K \tag{6.4}$$

This matches our general understanding - for example, a fire or a hot metal will first appear a dull red. As it warms up, it appears yellow, then eventually “white-hot”. From this, one can deduce an estimated temperature of the object by measuring the peak emitted intensity.

There are two more points to make regarding the radiation of thermal energy. First, a material that can perfectly emit thermal radiation may be referred to as a blackbody. This is manifested in a parameter called the “emissivity”. A perfect blackbody will have an emissivity of $\epsilon = 1$. Objects that do no emit as well will have an emissivity less than 1 and are sometimes referred to as a graybody. Essentially, the ratio of the emission of energy normalized by that from a perfect blackbody will define the emissivity. While beyond the scope of this course, this emissivity can be a function of wavelength and can be rationally engineered, which may have impacts in technologies like thermophotovoltaics or other thermal transport phenomena.

Finally, the total power that is emitted into an environment from a warm body can be computed by knowing this emissivity and knowing the temperature difference between the warm body and its environment. This is known as the Stefan Boltzmann equation and is defined as:

$$P = S\epsilon\sigma_S(T^4 - T_0^4) \tag{6.5}$$

where S is the surface area of the object, ϵ is the emissivity of the object, T is the temperature of the object, and T_0 is the temperature of the environment. As can be seen from this equation, the net power emitted by an object will become 0 if it is at the same temperature as its surrounding environment.

6.3 Introduction to phonons

As you may have discovered in this course, many aspects of the world can be represented by the "mass on a spring" model and can be thought of as both a particle and a wave. This is no different for thermal conduction, which is the focus of this part of the course. When we discuss the thermal conduction, we can think about the energy that causes atomic motion as quantized sets of lattice modes, in which the collective oscillation of atoms with a specific frequency and momentum can create the basis set to understand how energy is transferred into thermal energy within the crystal. To do this, we need to develop a physical model for the crystal. Here, we will use the hamronic oscillator (again) to represent the bonds between different atoms within the crystal. Note that this is approximate, as every interatomic potential will have higher order terms than this. However, if we simply model an atom as connected to two unmoving points through springs that have some spring constant/bond strength β, then we can easily solve for the set of allowed modes in this system.

For a mass on a spring system, the potential energy from stretching or compressing the spring is given as:

$$V = \frac{1}{2}\beta x^2 \tag{6.6}$$

where β is the spring constant and x is the displacement from the equilibrium position. We can insert this into the Schrodinger equation as:

$$\frac{d^2\psi}{dx^2}\frac{2M}{\hbar^2}(E - \frac{1}{2}\beta x^2)\psi = 0 \tag{6.7}$$

to find the the solutions. We will not go through the entire derivation (as it is quite tedious and not directly related to the course). However, we can find that the eigenenergies for this model take the following form:

$$E_n = (n + \frac{1}{2})\hbar\omega \tag{6.8}$$

where ω is the oscillation frequency and n is an integer that is greater than or equal to 0. Essentially, this means that any solution to the energy that can be given to an atom to oscillate is quantized in integer units of the oscillation frequency. By giving the system a specific amount of energy, it can promote the atom into an energy state where it will oscillate with a given frequency that obeys this equation. (As a side note, note that $n = 0$ is an allowed solution, which is the so-called "zero-point energy".) Figure 6.1 shows the atom and spring system, the potential, and the energy eigenvalues: We can use this description to explain phonons a bit more qualitatively. If we have a lattice of atoms that are bonded together, then we may expect the solutions to the movement of these atoms to take on a similarly quantized formulation. This can be thought of as a collective motion of the atoms

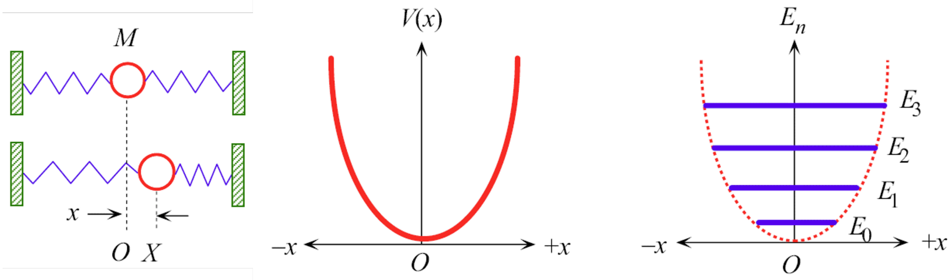


Figure 6.1: The quantum harmonic oscillator. Treating the bond as spring with a constant restoring force, the potential that can act on the atom as a function of distance from its equilibrium position is a quadratic function. The solutions to this equation give evenly spaced energy eigenvalues. Not shown here: The solutions look similar to a finite potential well, but the wavefunctions are described using Hermite polynomials. A more advanced treatment will go through this derivation. You may also find this in a chemistry version of quantum mechanics, as it can be helpful when thinking about spectroscopy. Adapted from Kasap.

that we will call phonons. Conversely, it additionally means that if we want the system to oscillate at a specific frequency, we could find a way of giving the crystal a specific amount of energy. This can be thought of as creating a phonon with a specific energy.

For this model, we will assume that we have a 1-dimensional line of atoms that are spaced by a given distance a. Each atom has the same mass M and the bonds are all of the same strength β. The position of atom r in this chain is given as $x_r = ra$ when there are no vibrations. The value of r goes from 0 to $N-1$. When the atoms are perturbed by some distance, we will refer to this as the displacement u. For the atom in position r, we would refer to the displacement u_r. Likewise, for the atoms that are nearest neighbors we would have to define u_{r-1} for the displacement of the atom in position $r-1$ and u_{r+1} for the atoms that are in position $r+1$. We note this because it will become important as we solve the equations of motion. This is represented in figure 6.2.

Before we walk through the derivation of this model, we can make a few comments about what we expect the equations to give us. First, we note that the behavior is a little different than the single atom model we discussed. Perhaps most importantly, each atom is bonded to another object that is able to move. This will mean that we need to think about the collective motion of the system rather than just the movement of one atom; the relative position of the nearest neighbors will also impact the restoring force. Second, we must note that the displacement can take place in any combination of directions. If the displacement is along the chain of atoms, this is referred to as a longitudinal wave. If the atoms are oscillating perpendicular to the bond, these are called transverse waves. There can be two different transverse waves - if the chain is along the x direction, then a displacement u_r in the x direction would be longitudinal. If the displacement were along the y or z directions, then the wave would be transverse. This is shown schematically in figure 6.3:

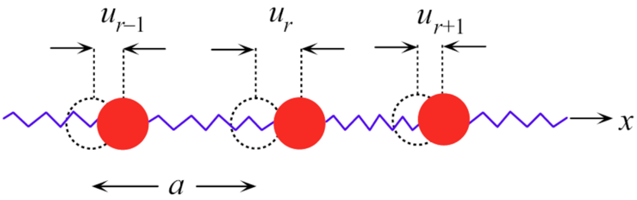


Figure 6.2: Coupled masses on springs in a crystal. here, we have a 1-dimensional chain of identical atoms spaced a unit cell a apart, with identical springs of spring constant β connecting them. From this, we can derive the position of any atom in position r through the lattice constant. We will write down the displacement of the atom from its equilibrium position as u_r, where r is the lattice site. Adapted from Kasap.

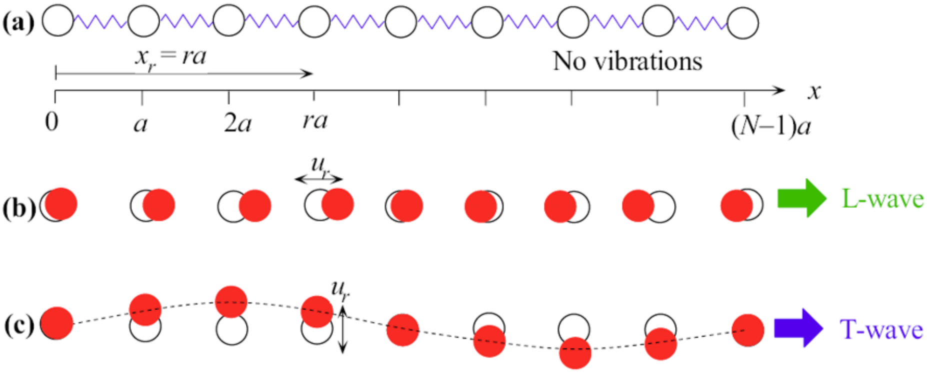


Figure 6.3: Longitudinal and transverse phonons. If the displacement from equilibrium is in the direction of propagation for the wave, then the phonon is called a longitudinal phonon. If the atoms are displaced perpendicular to this, they are transverse. Because we live in a 3 dimensional world, this means that there can be two transverse phonon modes and one longitudinal mode for a given wave. Adapted from Kasap.

One can show that the allowed eigen solutions to this problem will take the form of a traveling wave, not unlike the allowed solutions for free electrons in a solid. More explicitly, we can write that the displacement of the atom at position r can be written as:

$$u_r(x,t) = A\exp\left(i\left(Kx_r - \omega t\right)\right) \tag{6.9}$$

where K is the wave vector, x_r is the position in the x direction, ω is the oscillation frequency, and A is the amplitude. This is the form of a traveling wave, but Kx_r is discrete, as the atoms exist within a specific direction within the crystal. These are nominally referred to as "lattice waves" that have a period $\Lambda = 2\pi/K$ before u_r repeats itself. The displacement of the atom will oscillate in time and take a time period of $2\pi/\omega$ before u_r repeats itself. NOTE: This is the convention used by Kasap. There is no functional difference between K used for the wavevector here and k to define the wavevector for an electron. The difference is here just to make it clear when you are talking about phonons or electrons (as you will see later, electron-phonon interactions can help us to understand the mean free path of an electron).

To justify this, let's write down the net force of an atom that is displaced at some instantaneous point in time. The net force acting on the atom in position r will be given by Hooke's law for the bond with the $r-1$ atom and the $r+1$ atom. Since each of these atoms can be displaced by some amount relative to their equilibrium value, the net difference in the displacements will determine the net force. If the atoms at position $r-1$ and r are displaced by the same amount but in the same direction, then the bond is effectively not stretched at that moment in time. Therefore, we can write down that the net force is:

$$F_r = \beta(u_{r-1} - u_r) + \beta(u_{r+1} - u_r) = M\frac{d^2 u_r}{dt^2} \tag{6.10}$$

where F_r is the net force for the atom at position 4, β is the bond strength, and M is the mass. We have written down that the force will impart a net acceleration. One can show (which may be a good idea to prove to yourself) that the traveling wave solution we wrote down can be a solution to this equation. If we insert this proposed solution into this equation, then we can solve for the evolution of the displacement u_r with time. Note that this time harmonic solution has a straightforward solutions - each time derivative will return the same function, but multiplied by $-i\omega$. Two time derivatives therefore yields the same function with a factor of $(-i\omega)^2 = -\omega^2$ such that the solution to the equation of motion is:

$$F_r = M\frac{d^2 u_r}{dt^2} = -M\omega^2 u_r = \beta(u_{r-1} - u_r) + \beta(u_{r+1} - u_r) \tag{6.11}$$

Further, if this traveling wave solution is a solution for the atom at position r, then we can solve for the position of all of the other atoms in the chain, as the center position of the atoms is separated by a distance a. Because we are assuming that the solution to the oscillation is a traveling wave, we can write down the relationship between one atom and another in the lattice as having a phase difference, in the same way that we could for Bloch's theorem. If we know the displacement at position r at a time instant t, then we know that the position of any other atom at the same time as:

$$u_{r-1} = u_r \exp(iK(-a)) \tag{6.12}$$

We can insert our proposed solution into the equation for the force balance, and by factoring out u_r, we are left with only the $\exp(iKa)$ terms for the atoms that are in position $r+1$, 1 for atoms at position r, and $\exp(-iKa)$ for atoms at position $r-1$. Then we have:

$$F_r = M\frac{d^2 u_r}{dt^2} = -M\omega^2 u_r = \beta(u_{r-1} - u_r) + \beta(u_{r+1} - u_r) \tag{6.13}$$

transforms into:

$$-M\omega^2 u_r = \beta u_r \left(\exp(iK(-a)) - 1 + \exp(iKa) + 1\right) \rightarrow \omega^2 = \frac{2\beta}{M}\left(1 - \cos(Ka)\right) \tag{6.14}$$

where we have used a trig identity to convert the exponentials to the cosine term. We can additionally turn this into a slightly more compact form by transforming it again as:

$$\omega^2 = \frac{4\beta}{M}\sin^2\left(\frac{Ka}{2}\right) \rightarrow \omega = \sqrt{\frac{4\beta}{M}}\left|\sin\left(\frac{Ka}{2}\right)\right| \tag{6.15}$$

Where we have used another trig identity to get to this form, which is a bit easier to interpret in the form of the band structure. Here, we can immediately see that the minimum of the bandstructure frequency solutions will occur when $k = 0$ and the maximum will occur when $K = \pi/a$. The solutions to the band structure for this crystal is in figure 6.4:

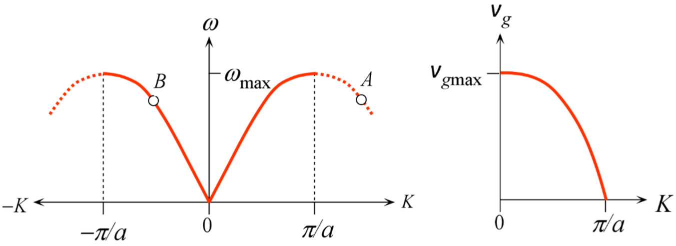


Figure 6.4: Phonon band structure. Like electrons, the band structure describes the allowed modes of the system and represents the phonon energy-momentum relationship. From this we can extract a group velocity, which is the derivative of this function, and determine how quickly phonons can propagate in the crystal. Note that phonons have a maximum frequency defined by the bond strength and atomic mass. Adapted from Kasap.

Let's discuss a few details about this band structure. First, we note that the same band structure rules apply here as they did for the electron - we can define a Brillouin Zone that has all of the meaningful information for the problem. In addition, the number of modes that exist along the wavevector axis is evenly spaced, so we can also determine an analogous density of states $g(w)$ for the crystal. The eigenmodes look like a series of waves with half wavelength increments as k increases and the number of nodes increases. Note that in general K can take on any quantized value from 0 to π/a, which we can write down as $K = \frac{q\pi}{Na}$, where $q \leq N$ (the number of atoms). If we insert this value of K into the solution for the wave, which then the form $\exp(iq\pi/Na)$ (at some instant in time t we can drop that term for illustration), we find that the period of this function decreases as the alue of q increases. Here this is even more apparent if we plot the solutions to the phonon band structure, as in figure 6.5: Essentially we can only draw waves using the positions of the atoms as points on the plot. At very low frequencies, the wavelength of the Bloch mode is very long, taking the entire crystal to make one half wavelength of the mode. As we increase in momentum, the number of nodes increases until we get to the Brillouin Zone edge, where every atom is oscillating in the opposite direction from its neighbors. From, this, and from the expression for the dispersion, we can find that the maximum frequency is defined to be

$$\omega_{max} = 2\left(\frac{\beta}{M}\right)^{1/2} \tag{6.16}$$

which are related directly to the bonding and the material. In addition, we can define the energy of a phonon as:

$$E_{phonon} = \hbar\omega \tag{6.17}$$

and the phonon's momentum is defined as:

$$p_{phonon} = \hbar K \tag{6.18}$$

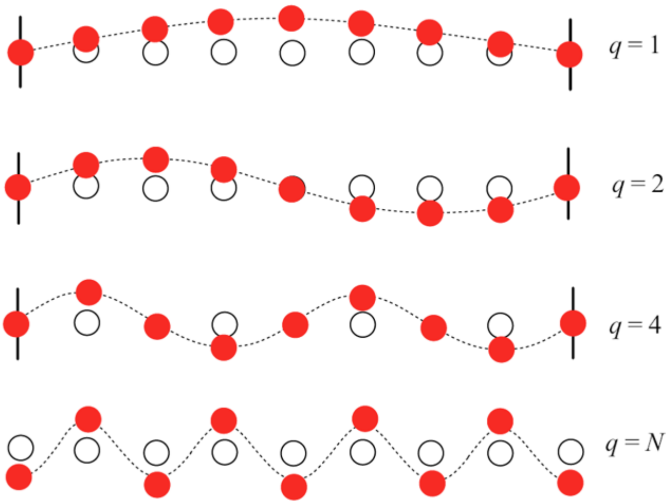


Figure 6.5: Phonon modes. As q increases, the effective wavelength of the phonon goes down and the number of nodes increases. Once q reaches the total number of atoms in the crystal lattice, we find that every other atom is oscillating completely opposite of its neighbors. We cannot go to higher spatial frequencies than this, as we no longer have enough atoms in the crystal to represent a wave with higher spatial momentum. That is, any value outside of the first Brillouin zone is essentially re-writing the modes that are already in the first brillouin zone. Adapted from Kasap.

Finally, note that a crystal in 3 dimensions can easily generalize this description. The wave vector K can now be a vector in 3 dimensions, and the Brillouin Zone edge will end at π/a, where a is the lattice constant in the specified direction. The same behavior can be extracted as in the electronic band structure.

The group velocity is an important factor for many physical processes. It is simply defined as:

$$v_g = \frac{d\omega}{dk} = \left(\frac{\beta}{M}\right)^{1/2} a\cos\left(\frac{1}{2}Ka\right) \tag{6.19}$$

(Note that the definition is general, but the equation on the right side is specific to this dispersion). We can see that the group velocity also depends on the strength of the bonding and the mass of the atoms in the lattice. The group velocity represents how quickly the wave can propagate within the crystal (and is distinct from the phase velocity $v_p = \omega/k$, which describes how quickly the wave can acquire phase in the crystal). At low frequencies, we can write down that the group velocity is essentially:

$$v_g = a\sqrt{\beta/M} \tag{6.20}$$

Noting that the Young's modulus $Y = \beta/a$ and that the density of the material is approximately $\rho = M/a^3$, we find that we can re-write the group velocity as:

$$v_g = \left(\frac{\beta}{M}\right)^{1/2} a = \sqrt{\frac{\beta a^2}{M}} = \sqrt{\frac{Ya^3}{m}} = \sqrt{\frac{Y}{\rho}} \tag{6.21}$$

This is quite powerful - we can essentially relate the phonon band structure (at low frequencies) to physical properties of the material - the elastic modulus and the density! In this brand, we also therefore can relate the speed of sound to the phonon dispersion.

6.4 Heat capacity

The heat capacity of a material can change with temperature, which does not follow from classical theories. We will find that a simplified model of the phonon band structure is actually a reasonable approximation to determine the temperature-dependent heat capacity, which matches well for many crystalline materials. In addition, it follows the classical theory well at high enough temperatures.

The classical method to compute the heat capacity is to use the equipartition theorem, which states that an object at thermodynamic equilibrium will evenly split its energy amongst all of its degrees of freedom. This means that each degree of freedom will add $1/2k_BT$ to the overall energy. For example, an electron can have kinetic energy, but this energy can be in any direction. This translational kinetic energy in the x, y, ad z directions means that the average energy of an electron is $3/2k_BT$. For a solid, we must additionally consider the potential energy that is stored in the form of bonds. For a 3-dimensional material, this would then mean that the total average energy at a given temperature is $3k_BT$. The heat capacity per atom is defined as:

$$C = \frac{d(KE + PE)}{dT} = \frac{dU}{dT} \tag{6.22}$$

which gives us the classical result of $C = 3k_B$. This is known as the Dulong-Petit heat capacity. This works qiute well for many materials - Silver, Copper, Germanium, Mercury, and Tungsten all have very similar heat capcities that are essentially this value. However, materials like Silicon, Beryllium, Diamond, and many polymers have values that are much lower than this, and violate this classical rule. As we will find out, this is related to the phonon band structure and the fact that these materials are composed of light elements (and hence the low M in the denominator of the constant term) or have strong bonds (the β term in the numerator).

The solution to this (or at least at the level of this course) is to use the Debye model, which assumes that the phonon bandstructure is linear within the Brillouin Zone. This is clearly incorrect, but also very clearly makes the problem easier to solve (the final result is already gross without there being extra terms to try and integrate). This essentially says that all phonons can propagate with the same group velocity throughout the crystal. The Debye model will also have a maximum frequency in the Brillouin Zone ω_D, which will become important later. Finally, consider the fact that we have frequently interchanged frequency and energy in this course. Since we know that these particles can have a frequency ω and have a defined energy $\hbar\omega$, we can also consider the effective temperature of this particle by mapping the energy to thermal energy, which would essentially define a $T_D = \hbar\omega_D/k_B$ by equating $k_BT = \hbar\omega$.

In the quantum picture, we are considering all of the thermal energy that a crystalline solid (like diamond) has will go to phonons. These are quantized particles (and therefore there is a countable and finite number of them in a solid at a given temperature). In order to determine the heat capacity, we need to understand what this distribution of phonons looks like at different temperatures. The heat capacity essentially defines how much energy it takes to increase the crystal temperature by 1 degree. Here, we can understand this by computing the internal energy of the system and look at the temperature dependence of this by counting up the phonons in the crystal.

As alluded to earlier, phonon occupation statistics obey the Bose-Einstein relationship, which is:

$$P(E) = \frac{1}{\exp\left(\frac{\hbar\omega}{k_B T} - 1\right)} \tag{6.23}$$

The energy of a phonon is defined as $\hbar\omega$. Noting that the difference between frequency and energy in these calculations is a factor of $\hbar$, we can write down a general description for the internal energy of the system as:

$$U_m = \int_0^{\omega_{max}} \hbar\omega g(\omega) P(\omega) d\omega \tag{6.24}$$

This is essentially fancy accounting - the understand the total energy of the system, we need to add up all of the different places that energy can go. Each phonon frequency has a different energy, and there can be a certain distribution of modes in the system defined by the density of states - at some frequencies there may be more modes than others. Finally, we need to know what the probability will be to have a phonon at that given frequency. To sum everything together, we integrate over all frequencies to find the total energy. By multiplying the energy for a phonon at a frequency ω, the density of states, and the probability distribution, we are computing the average expected energy contained in each frequency. By integrating through all possible frequencies, we are then summing all possible sources of energy.

We will not derive the density of states for the Debye model, but sketch out the rationale and provide the end result. Here, we must conserve the total number of modes that could exist for a crystal with a defined number of atoms. Just like in the linear combination of atomic orbitals, we must have an equal number of modes for the number of atoms. In the Debye model, the dispersion (the ω vs K diagram or the band structure) is entirely linear, which makes this computation straightforward. For a linear dispersion, we know that the density of states is:

$$g(\omega) = \frac{3V}{2\pi^2}\frac{\omega^2}{v^3} \tag{6.25}$$

where V is the volume of the crystal, v is the average velocity of the phonon waves. The maximum frequency in this model is different than the actual band structure and is given by:

$$\omega_{max} = v\left(\frac{6\pi^2 N_a}{V}\right)^{1/3} \tag{6.26}$$

where N_a is the number of atoms. We can define the Debye temperature from this, where

$$T_D = \frac{\hbar\omega}{k_B} \tag{6.27}$$

The average speed of the phonons will be defined as the phase and group velocity in this case (and generally the case in a material with linear dispersion), which will follow the low frequency slope of the actual band structure from above.

figure 6.6 shows the density of states as a function of frequency for experimentally derived values of Copper (through neutron scattering) and the 3D debye model. As you can see, the overall behavior is quite different, but we can still use this in this model. More accurate predictions can be made by numerically integrating the actual density of states for a material if a more accurate result is required.

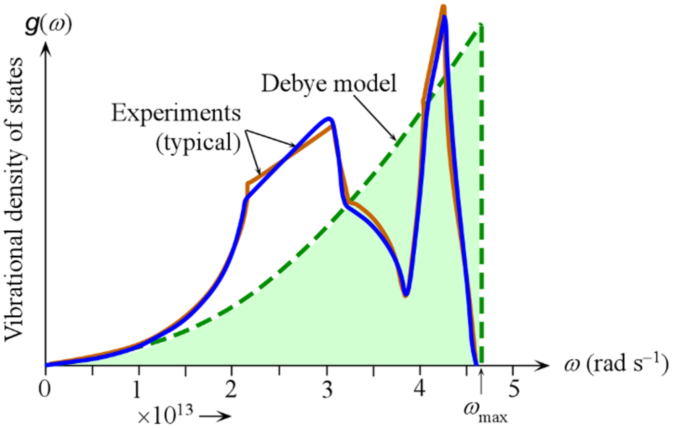


Figure 6.6: The Debye model density of states as a function of frequency compared to experimentally derived values from neutron scattering. Clearly they are different, but the low frequency behavior matches as expected. Here, the curve is scaled so that the total density of states is conserved. Note that the Debye temperature of Copper is inferred from this model to be $T_D = 344K$, right around room temperature. Adapted from Kasap.

Putting all of this together and taking into account, we have that:

$$U = \int_0^{\omega_{max}} \frac{\hbar\omega}{\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1} \frac{3V}{2\pi} \frac{\omega^2}{v^3} \tag{6.28}$$

From which we can determine a molar heat capacity by solving for U, then taking the derivative of the energy with respect to temperature and taking into account that it is a molar quantity rather than a total quantity. This is a complicated function that is a bit of a mess to solve. The most straightforward method to do this is to non-dimensionalize with respect to the Debye temperature T_D to get (a still gross equation):

$$C_m = 9k_B \left(\frac{T}{T_D}\right)^3 \int_0^{T_D/T} \frac{x^4 e^x}{(e^x - 1)^2} dx \tag{6.29}$$

We see that the equation still contains an integral and is therefore not the most helpful. however, there are some limiting cases that are worth discussing. First, we can plot the result and see how things look. Figure 6.7 plots the molar heat capacity normalized by the Dulong-Petit value of $3k_B$, as a function of temperature normalized by the Debye temperature:

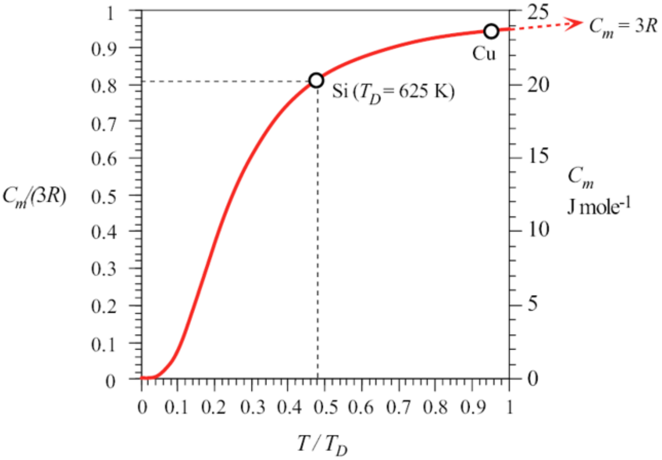


Figure 6.7: The heat capacity of a solid as a function of temperature according to the Debye model. At low temperature there is a nonlinear dependence that follows T^3. At high temperatures the heat capacity tends towards the classical value. Note that Kasap uses R instead of k_B. Adapted from Kasap.

There are a few conceptual things to unpack here. First and foremost we should note that this model, even simplifying the dispersion to be linear, captures a large portion of crystalline materials. At low temperatures it gets the T^3 dependence found in experiments. At high temperatures, it gets to the Dulong-Petit value. In addition, the phonon dispersion for light element materials like Diamond also make sense - because the slope of their dispersion is quite a bit higher, their effective Debye temperature is higher. Since the function is monotonically increasing, this means that light element materials will have a value lower than the classical limit if their Debye temperature is substantially above room temperature. Indeed, the estimated Debye temperature for Diamond is well over 1000K, which would put the heat capacity for the material very low compared to the classical limit (on the lower end of the x axis).

Let's interpret what this may mean. At very low temperature, only a subset of the phonon energies can be accessed. As the temperature increases, more phonons are being created and the average energy of the phonons can also increase, as higher energy phonons can be created. The phonon concentration increases as t^3 and the average energy is linear in T, which yields the T^3 dependence for the heat capacity. (Note: This is true for several crystalline materials, but not true for others. Amorphous materials are a well known exception that have a different scaling relationship). The Debye temperature is related to the maximum frequency of phonons. When the temperature is near or above this value, we find that the total number of phonons increases and the phonon energies are in this energy range, but the scaling relationships are different. Here, the total internal energy of the system scales linearly, which leads to a near constant heat capacity.

This deviation of the heat capacity from the classical value is a direct consequence of the existence of phonons and the quantization of energy associated with them.

6.5 Heat transport and thermal conductivity

As a final note, we will consider some basics of the thermal conductivity of a solid and how one might be able to engineer the thermal conductivity of a material. Similar to an electron, we can take a quasi-classical approach to try and understand what is going on. Here, we can treat the phonon as a particle that propagates with a certain speed for a characteristic time before it scatters off of a defect, grain boundary, etc. in the crystal to re-direct it. In this way, we can quantify the speed with which an object will travel. Knowing its energy, we can then understand the rate of change of heat energy inside of a material with there is a temperature difference, which will lead to the diffusion of heat in a crystal. Here, we will consider an insulating material like Diamond, Sapphire, or Magnesium Oxide, which only have phonons that can conduct heat. In general, electrons can also conduct heat, but the way in which insulators transmit heat is through the phonon contribution. In metals, the thermal conductivity can be dominated by the electronic thermal conduction rather than the vibrational.

6.5.1 Thermal conductivity from phonons

Quantitatively, we can define the thermal conductivity as a differential equation that relates the heat flux Q' traveling through a cross-sectional area A to the temperature gradient dT/dx in a material through the thermal conductivity κ, which provides a differential equation not unlike diffusion:

$$\frac{Q'}{A} = \kappa \frac{dT}{dx} \tag{6.30}$$

Here, the temperature gradient provides a diriving force for thermal energy to diffuse from the hot side of a material to the cold side of the material. The rate at which this can happen is related to the thermal conductivity, in the same way that diffusion of electrons is related to the conductivity of a material. We can treat the thermal conducitvity similar to an electron and give a quasi-classical picture, where:

$$\kappa = \frac{1}{3} C_V v_{ph} l_{ph} \tag{6.31}$$

where κ is the thermal conductivity of phonons, C_V is the constant volume heat capacity, v_{ph} is the velocity of the phonons, and l_{ph} is the mean free path before scattering. This is quite analogous to an electron - there will be some characteristic speed we can associate with the particle as well as a characteristic time before it is re-directed due to scattering off of an object. The mean free path l_{ph} is related to the velocity of the particle (the phonon in this case) and the characteristic time between scattering events τ as:

$$l_{ph} = v\tau \tag{6.32}$$

To a rough approximation we can assume that the speed of the phonons is given by the speed of sound in the material (which is reasonable in the linear part of the dispersion diagram, but perhaps inaccurate for higher frequency waves). The characteristic time can be based on the average distance to a defect, dislocation, grain boundary, etc. in a polycrystalline material. In addition, phonons can interact with each other and scatter off of each other thorugh phonon-phonon interactions. This is a complicated thing to compute. Bosons are typically non-interacting - unlike an electron and other fermions, bosons don't really "feel" the presence of other bosons, which is part of the reason that their occupation statistics are different (and follow the Bose-Einstein relationship rather than the Fermi-Dirac one). However, materials that have an anharmonic potential well that describes their bonding can have phonon-phonon scattering due to the inherent nonlinearity associated with this. This is beyond the scope of the course, but the key thing to think about is the fact that a standard Hooke's law spring will have a quadratic potential well associated with it, which is symmetric. Any material that has an anharmonic potential is not perfectly symmetric - say, a Lennard-Jones potential. This potential well must be described with a functions that contains higher order terms than the quadratic, which can enable the coupling between different phonons. We will not derive this, but this can point to the fact that thermal expansion coefficients can be measured to help understand the rate at which phonons will interact (as the thermal expansion coefficient α is a a measure of the bonding anharmonicity). All of this is to say that phonons can scatter off of other phonons, which means that we will expect a temperature dependence in the thermal conductivity. This is shown schematically in figure 6.8:

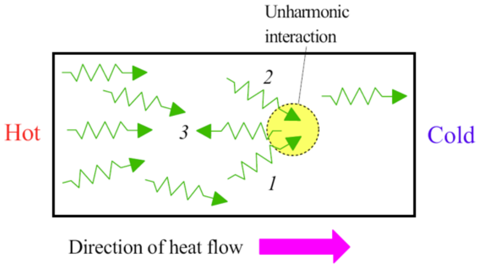


Figure 6.8: Phonon-phonon interactions. Here, the anharmonic potential well allows phonons to scatter off of others. In addition, the collision of two higher energy phonons can create phonons of different energies if momentum matching is conserved. This can lead to unintuitive results like the one shown in the figure, in which two phonons traveling to the right interact and form a third wave that propagates in the opposite direction. This is nominally called Umklapp scattering and is due to the fact that any mode with momentum outside the Brillouin Zone can be folded back into the Brillouin Zone. Adapted from Kasap.

If the velocity doesn't change (a rough approximation) but the density of phonons increases linearly with increasing temperature (reasonable above the Debye temperature), then we would expect an inverse relationship with temperature. For pure crystals with few defects, we indeed observe this in experiment, with:

$$\kappa_{crystal} \approx 0.38 \frac{\bar{M}}{(NV^2\gamma^2)^{1/3}} \frac{v^3}{T} \tag{6.33}$$

where $\bar{M}$ is the average mass of the atoms in the crystal, N is the number of atoms, V is the crystal volume, v_s is the speed of sound (and therefore the speed of phonons), T is temperature, and γ is the grüneisen parameter, which is related to the bulk modulus B of the crystal, the thermal expansion coefficient α, and the heat capacity C as:

$$\gamma = \frac{B\alpha}{C} \tag{6.34}$$

Note that this is generally an approximation and valid for several materials, but not all. This works well for pure crystals like MgO and Sapphire.

Glass/amorphous materials have a very different thermal conductivity scaling relationship. As you may expect, the phonon picture works well in highly crystalline solids, and the density of defects is relatively low to allow a characteristic scattering time that can depend on the phonon concentration. In an amorphous material, there are vacancies and general disorder in the material such that scattering off of a defect is quite common. In general, this would mean that the mean free path of a phonon in an amorphous material is quite short, which will additionally be roughly temperature independent, as scattering off of defects will be dominating over scattering off of other phonons. Therefore, we expect there to be a different scaling relationship and we indeed find this to be the case. The thermal conductivity scales roughly as:

$$\kappa_{glass} \approx 1.2 \frac{k_B v_s}{V^{2/3}} \tag{6.35}$$

which is temperature independent. For crystals with very large unit cells or with complicated structures, the conductivity can also have a weaker temperature dependence. In general these materials will also have a lower thermal conductivity. For example, the thermal conductivity of sapphire around room temperature is approximately $\kappa \approx 1000 W/(m \cdot K)$, whereas for glass it is closer to $\kappa \approx 1 W/(m \cdot K)$. Clearly, this coherent thermal vibration/wave approach is a reasonable way to think about the transport of heat in these materials.

At lower temperatures, the scaling relationships will be different. Because the phonon concentration is low, the limiting factor is the mean free path before colliding with defects in the crystal. At this temperature range the heat capacity scales as T^3, so we expect the thermal conductivity to do so as well, since the velocity and the mean free path don't change significantly with temperature.

6.5.2 Thermal conductivity from electrons

Electrons can additionally carry heat, and will scatter in a similar way to phonons. Because electrons are particles that can carry heat and charge, it is reasonable to believe that the thermal conductivity of an electron will be related to the electrical conductivity. This is approximately true, with the following relationship holding for a metal:

$$\frac{\kappa_{electron}}{\sigma_{electron}} = LT \tag{6.36}$$

where we have explicitly defined these quantities as the electronic contribution. This is known as the Wiedemann-Franz law, and L is the Lorenz number equal to:

$$L = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2 = 2.44 \times 10^{-8}\ \mathrm{V^2K^{-2}} \tag{6.37}$$

Note that this is an empirical relationship, but nominally stands up to reason - if electrons can flow more easily and carry both charge and heat, then you may expect that more thermal conduction from electrons will result. Therefore, if we can determine the electrical conductivity and how this scales with temperature then we will have a reasonable understanding of how much heat will be conducted by these particles as a function of temperature. Note that, in general, thermal conducitivity will be a result of both phonons and electrons, and different materials will have different dominating components - Diamond is a great thermal conductor, driven entirely by phonons. In metals, the thermal conductivity of electrons is the dominant mechanism for their thermal conductivity.

Consider the fermi velocity of an electron, which would represent the speed of the conduction electrons in a metal. These will have a velocity given as:

$$v_F = \sqrt{2E_F/m_e} \tag{6.38}$$

(more strictly the effective mass may be more appropriate here). We can follow the same analysis and come up with a nearly classical representation for the scattering of the electron in the material in the same way we did earlier in the quarter. However, we can be a bit more prescriptive in our analysis of the scattering time with temperature and explain the transport behavior of many classes of materials. Previously, we discussed that there's an approximate analysis for the mobility of an electron related to the size of an atomic core that an atom can scatter off of. We said that this scaled with the temperature and that the average size of the atoms appeared to increase as the temperature increased, as they vibrated with a greater amplitude with more thermal energy. Since we have discussed thermal energy in the crystal as a collection of phonons, we can actually think of this as a particle-particle interaction and think about the scaling relationships between the momentum of the phonons, the concentration of phonons, and therefore understand the scaling relationship with temperature. The electron-phonon interaction is described in figure 6.9:

When an electron interacts with a phonon, we must conserve momentum. We can essentially think of this as an inelastic collision in which the final momentum of the particle. The phonon momentum is quite small and so we can assume that the electron has nearly the same magnitude after the collision. This re-direction can modify the conductivity. As more of these collisions take place, the electron movement will eventually be randomized. If it takes N collisions to randomize the electron momentum (and therefore reduce the conductivity), we can posit that:

$$\sigma_{electron} \propto N\tau \tag{6.39}$$

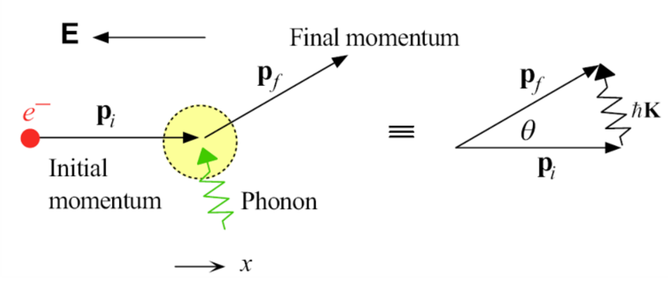


Figure 6.9: Electron-phonon scattering can re-direct electrons through the absorption of a phonon. The momentum is very similar to the initial momentum, as the electron wavelength is much shorter than the phonon wavelength. This re-direction of momentum will reduce the current flow in the initial propagation direction, asthe projection of the momentum vector onto that direction has been reduced. Adapted from Kasap.

where N is the number of collisions and τ is the average timescale before scattering events. From this, we can determine the temperature scaling relationships for the conductivity of electrons in a metal. Electrons will scatter off of defects, dislocations, and grain boundaries. They will also scatter off of phonons, which scale with the temperature. Above the debye temperature, the energy and momentum of phonons is large and few scattering events are required to randomize the momentum of an electron. The number of phonons increases linearly with temperature. Because the number of collisions required is small, then we expect that:

$$\sigma_{electron,highT} \propto \tau \propto \frac{1}{n_{ph}} \propto \frac{1}{T} \tag{6.40}$$

Where the high temperature conductivity $\sigma_{electron,highT}$ is inversely proportional to the concentration of phonons n_{ph}, which gives an inverse dependence with temperature. This scaling relationship also gives the thermal conductivity scaling for electrons at this temperature.

Below the Debye temperature, we have a different scaling relationship. As we saw in the Debye model, the phonon concentration follows a different scaling relationship, with $n_{ph} \propto T^3$ AND the average energy of a phonon increases linearly with T. Since the energy and momentum are directly related in the Debye model (i.e. the ω vs k diagram is linear), this also means that the momentum of the phonon increases linearly with temperature. Because the momentum is lower, more collisions will be required to randomize an electron. Each collision will impact the momentum in a random process, such that averaging the effect is important. Combining this with the temperature dependence of the phonon concentration below the Debye temperature yields:

$$N \propto \frac{1}{T^2} \rightarrow \sigma \propto N\tau \propto \frac{N}{n_{ph}} \propto \frac{1}{T^5} \tag{6.41}$$

Which is indeed the temperature dependence of a metal like Copper below the Debye temperature. Note that this is true for pure materials - if there are a significant number of dislocations, defects, or impurities, the scaling relationship may not be quite the same.

7 Bonus content

7.1 Introduction

In this last chapter, we will include a few vignettes that go over some aspects of the course material and synthesize them in an applications-driven format. This year, we will discuss Thermoelectrics, magnetic recording technologies, and piezoelectric materials. These are by no means a comprehensive analysis but are intended to review the course content and provide context for engineering challenges or optimization of material behaviors.

7.2 Thermoelectrics

A thermoelectric is a device that can heat or cool with an applied voltage, or where a temperature difference can be used to induce a voltage. From this, a variety of useful devices can be created, included Peltier coolers and thermoelectric generators. These are essentially devices that can interconvert thermal and electrical energy. These can be useful to control the temperature of an object without moving parts, as it requires running current through a device rather than liquid or fans. Conversely, for stable power generation, a thermoelectric generator can use temperature differences to generate electricity. NASA, for example, will use thermoelectrics for space missions, as solar cells and other power generating objects are not sufficient far away from Earth.

Consider the following thought problem: A heavily doped semiconductor that has a temperature gradient within it will additionally lead to the net flow of electrons. This can be for a few reasons - first, the thermal generation of electrons across the band gap can create a gradient of carriers that will then want to diffuse to regions of lower concentration of electrons. In addition, the thermal energy of the electrons on the hotter side will give them a higher velocity than electrons on the colder side. As electrons drift or diffuse across this temperature gradient, there will be a steady state voltage that will be created across the semiconductor. We can define a value that quantifies the voltage difference created as a function of the temperature coefficient and call this the Seebeck coefficient as:

$$S = -\frac{\Delta V}{\Delta T} \tag{7.1}$$

More concretely, we can define the fraction of current density driven from temperature gradients through a new term to the current density as:

$$J = -\sigma \nabla V - \sigma S \nabla T \tag{7.2}$$

where the last term is a measure of the temperature induced current flow. In hte limit of no current flow (i.e. J = 0), we recover the Seebeck coefficient equation as:

$$\nabla V = -S \nabla T \tag{7.3}$$

This equation means that we can construct systems such that energy can be generated from this potential difference OR we can use a voltage difference to drive a temperature gradient.

7.2.1 Peltier Cooling

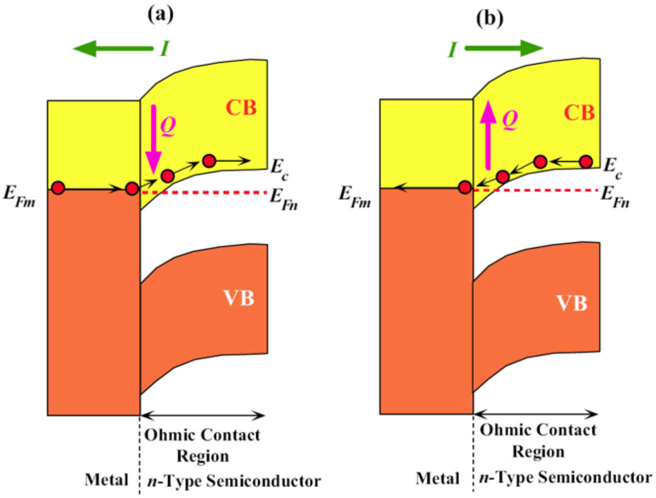


Figure 7.1: Current flow across an ohmic contact can locally heat or cool. If the current flow leads to an increase in the electron potential energy, the local heat flux is negative as heat is removed from the system. Conversely, reversing the flow of electricity will lead to local heating across the junction. Adapted from Kasap.

Consider the band structure of an n-type semiconductor forming an ohmic contact with a metal. Figure 7.1 shows an ohmic contact and the flow of electrons depending on the current. We saw previously that this led to a buildup of electrons at the metal-semiconductor junction. Because of this, the limiting factor for charge transport in this material system was the bulk resistivity of the semiconductor, not the metal-semiconductor junction. The band bending means that there is are energy levels in the semiconductor that are lower near the metal interface. When current flows across this interface, this means that the potential energy of the electrons must change, as the only states that are available for conduction are in the conduction band of the semiconductor. As electrons move from the metal into the semiconductor, the potential energy of the electrons increases. This energy must be taken from somewhere; in fact, the electron will take thermal energy from around it in order to increase their thermal energy. This means that the local area must cool down to balance the energy.

We can quantify the amount of energy a single electron will lose or gain by traversing the ohmic contact region, which we call the Peltier coefficient. Here, we can define the peltier coefficient Π simply by looking at the energy difference between the fermi level of the metal and the conduction band away from the junction. We find then that:

$$\Pi = \frac{1}{e}\left[\left(E_C - E_{Fn} + \frac{3}{2}k_B T\right)\right] \tag{7.4}$$

From which we can measure the heat flux Q' as

$$Q' = \pm \Pi I \tag{7.5}$$

where the current direction can determine whether the junction is heating up or cooling down.

7.2.2 Thermoelectric power generation

We can put these components together to form a power generation or thermoelectric cooling component. This effect will work for either a p-type or n-type semiconductor, depending on the flow of current. The direction of current for cooling in a p-type semiconductor will be the opposite of an n-type semiconductor. Consider the case in figure 7.2. If both metal contacts on a semiconductor are the same metal, then current flow across the entire device will lead to one junction cooling and the other heating. If we combine p-type materials and n-type materials into a structure, then connecting them in the appropriate manner can lead to current flow such that cooling occurs on one side of the device, and heating occurs on the other. In this way, a cool body can be connected to a warm body and heat can be removed from the cooler body, leading to refrigeration.

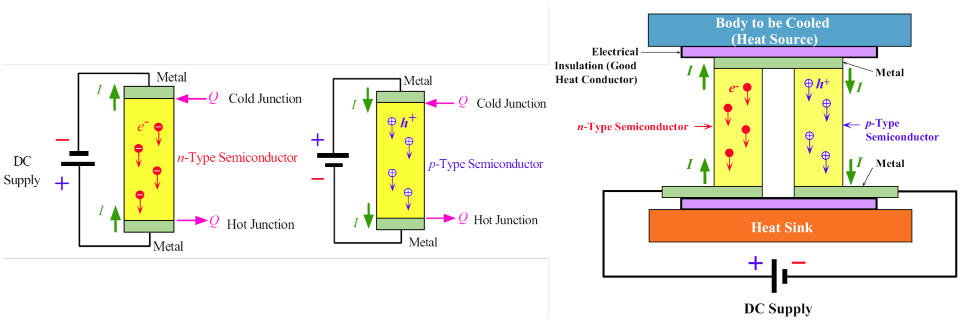


Figure 7.2: Current flow and heating/cooling in semiconductors. Depending on the direction of current flow, heat can be removed or added to a junction in the material. Connecting these to cooled or warming bodies can lead to a thermoelectric cooler, where electrical energy will move heat from the cooler to the warmer body. Adapted from Kasap.

Conversely, connecting a source of thermal energy to a cooler body through a thermoelectric material can lead to useful work that can be extracted and stored, enabling thermoelectric power generators to convert thermal energy into electronic energy.

7.2.3 Figure of merit and tradeoffs

This sort of device combines the thermal transport characteristics and the semiconducting properties of materials, which means that there will inherently be tradeoffs. The figure of merit for thermoelectrics is typically defined as zT, which relates the thermal and electrical conductivity of a material through:

$$zT = \frac{\sigma S^2}{\kappa} T \tag{7.6}$$

where a material with low thermal conductivity (which may maximize a ΔT in a material, high electrical conductivity (such that charge can easily flow) and high Seebeck coefficient can increase the efficiency of these materials. In fact, one can show that a thermoelectric power generator can act as a heat engine that has an efficiency defined as:

$$\eta = \frac{\text{Energy to the load}}{\text{heat absorbed at the hot junction}} = \frac{T_H - T_C}{T_H} \frac{\sqrt{1+z\bar{T}}-1}{\sqrt{1+z\bar{T}}+\frac{T_C}{T_H}} \tag{7.7}$$

where $z\bar{T}$ is the average value between the hot and cold temperature ranges. Here, we can see that the first term $\frac{T_H - T_C}{T_H}$ is the Carnot efficiency of a heat engine. As $z\bar{T}$ increases, the efficiency of the system increases. In the limit of extremely high $z\bar{T}$, this system approaches the Carnot efficiency. For low $z\bar{T}$, the efficiency decreases.

As we have previously discussed, the thermal conductivity of a material can depend both on the phonon contribution and the electron contribution. In addition, the thermal and electrical conductivity themselves are also strong functions of temperature, which means optimizing the figure of merit zT is challenging and will depend on the fundamental material properties and their microstructure. Even at a specific temperature, the doping concentration can modify the properties of the material such that there is a strong variation in the thermoelectric response. One such example is plotted in figure 7.3, where the doping concentration, conductivities, seebeck coefficient, and overall zT is plotted as a function of doping.

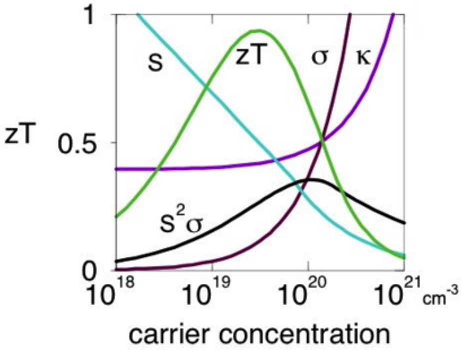


Figure 7.3: zT of a small bandgap material as a function of doping concentration. There is an optimal range for this material to act as a thermoelectric at this specific temperature. Adapted from Prof. Snyder's website.

7.3 Magnetic recording

7.3.1 Soft vs Hard magnets

Here we will briefly highlight some of the factors for ferromagnetic materials in modern technology. As we discussed, the hysteresis curves are a hallmark of a ferromagnetic materials. The size of this hysteresis curve is a measure of a few important quantitites. First, the coercive field is the field required to demagnetize a magnetized piece of material and is therefore a measure of how good of a permanent magnet you will have. In addition, the remnant magnetization B_r is a measure of how much net magnetism remains in the material when there is no magnetizing field on it. These components essentially specify the intersection values on the $B - H$ curve. For a very good permanent magnet, we would want the remnant magnification to be large and the coercive field to be large. This will lead to a strong magnet that is hard to demagnetize, and can be used in motors, speakers, headphones, tools, and permanent data storage in magnetic recording technologies. Essentially, a hard magnet will have as large and wide of a hysteresis loop as possible. However, the space contained within the hysteresis loop is also a measure of the energy lost in the magnet during each cycle. The large and open hysteresis loop then means that the cycle the magnetization requires a lot of energy. This makes intuitive sense, as having a large remnant magnetization and coercive field indicates that the potential energy barrier required to flip the sign of the spin direction is large. This is undesirable for many applications where you may want to cycle the magnetic field quickly, as it means that a lot of energy is lost due to the hysteresis losses.

In contrast, a soft ferromagnet has essentially the opposite properties. The coercive field is quite low and the remnant magnetization is also quite low. This means that they would make poor permanent magnets because the magnet strength without an applied field is quite low. In addition, it would take very little magnetic field to demagnetize the material. This is good in cases where you want to quickly apply magnetic fields for some purpose. For example, high speed electronics may have inductors that are enhanced by using magnetic fields. Transformers and other structures will create magnetic fields that oscillate at least at 60 Hz, so minimizing the hysteresis loss is valuable for these devices. They would make poor devices where you need a strong permanent magnet. These differences are highlighted again in figure 7.4:

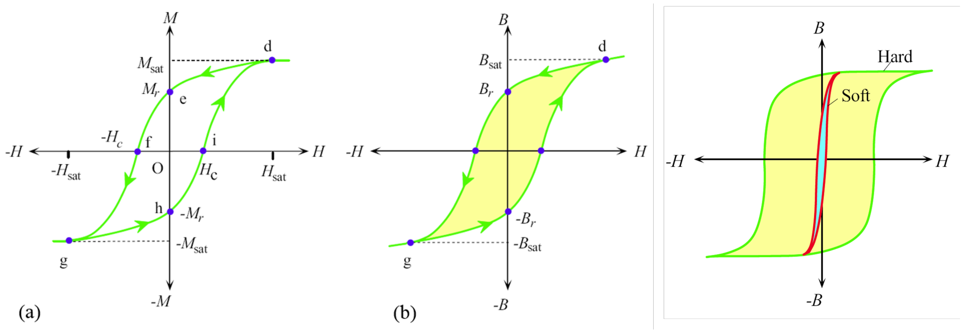


Figure 7.4: M vs. H of a magnet. Adding in the magnetizing field, the y axis can be computed as $B = M + \mu_0 H$ (middle). The area within the hysteresis loop describes the energy loss per cycle. Finally, the relative area, coercive field, etc. can help to classify magnetic materials as "soft" or "hard", and may find different uses in modern technologies. Adapted from Kasap.

7.3.2 Magnetic recording

Magnetic recording is another useful example of using multiple kinds of ferromagnets. Figure 7.5 shows a schematic of how a magnetic hard drive recording device functions. At the top, a disc will spin about an axis and a recording head (that looks like a record player). The disk has a thin film of magnetic nanoparticles on it. The magnetic nanoparticles are hard magnets, but are small enough that there are no ferromagnetic domains on them. The tip of the write and read head has a series of soft magnetic devices that can allow for direct writing of the magnetization direction. write head has a soft magnet with an electrical coil around it. When current flows through the coil, the gap in the magnet will have a magnetic field with a direct specified by the coil current direction. When the write signal is on, the magnetic field in and around the gap is strong enough to flip the magnetization direction of the hard drive. When the disk is spinning at a given velocity and the input signal is timed appropriately, a "bit" of information can be stored in the magnetic domain orientation. Several nanoparticles within a patch will have a magnetization that is aligned in a specific direction, which can form the "1" or "0" in the hard drive. These particles are smaller than the write head such that several will form a single bit.

To read the data, a different soft magnet can be passed over. The magnetic field from each bit will induce a voltage that can be read out on a current loop. When this is amplified, the voltage signal can then be a readout of the magnetic field orientation, which then can be interpreted as a data stream.

There are some challenges associated with this technology. Because the magnetic nanoparticles are small, there are no domain walls to minimize the overall energy. Each particle then has a volume of spins that can be flipped from an external magnetic field OR from thermal energy. The magnetization direction is set by the magnetic field and is essentially a local equilibrium orientation for the nanoparticles. The nanoparticles will have their spins aligned along an easy direction of the crystal,

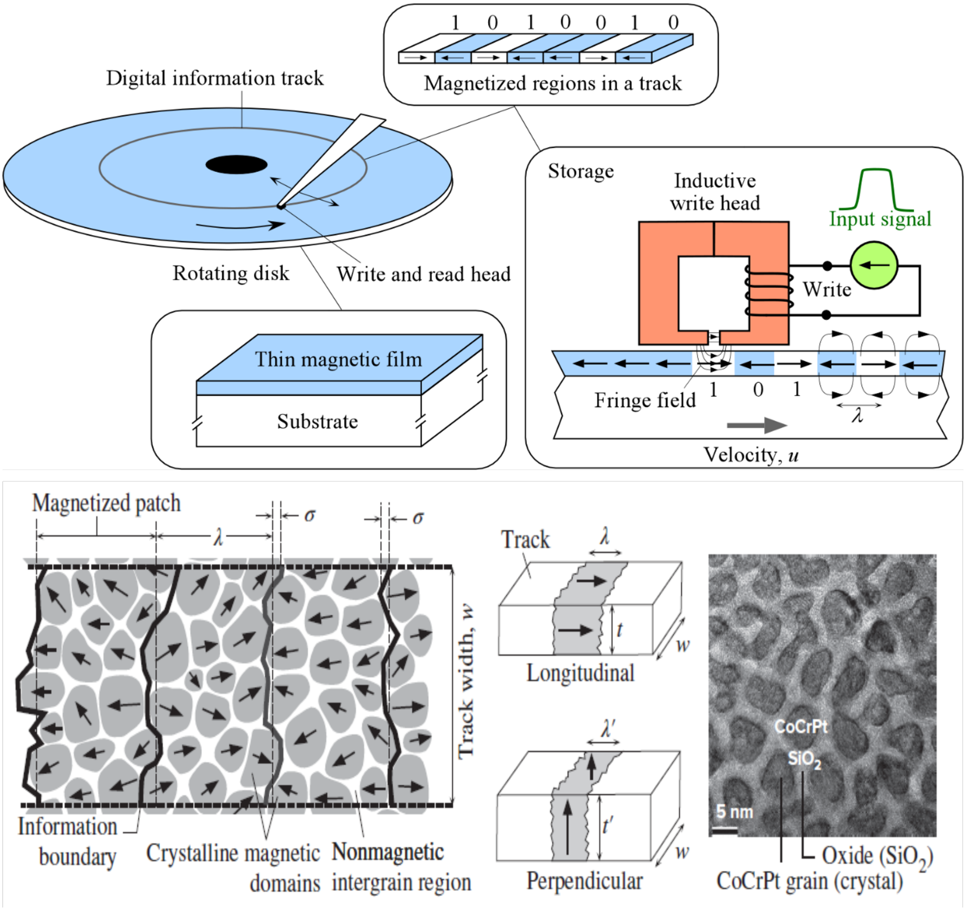


Figure 7.5: Magnetic recording. Bits of magnetic tape or magnetic material on a disk can be flipped in plane to represent “1” and “0”. An inductive write head can have an input signal encoded through a current signal through a coil, which induces a magnetic field in the gap that can locally flip magnetic nanoparticle domain orientations. Adapted from Kasap.

as it is harder to magnetize them along the hard direction. The energy difference, as quantified by the magnetetocrystalline anisotropy, essentially means that the particles are in a local equilibrium position. As the spin orientation varies as a function of angle, one would find a potential energy barrier in other directions due to this anisotropy. If the potential barrier to overcome to flip the magnetic field orientation is low, then thermal energy could be enough to overcome this barrier quickly. If it is high, then it is unlikely that thermal fluctuations can cause this to flip quickly. We can quantify this in an effect timescale that represents the average time that these nanoparticles will be oriented the way that they were set during the writing step. This time τ is related to the magnetocrystalline anisotropy energy K_u and the particle volume V and the thermal energy $k_B T$ as:

$$\tau \propto \exp\left(\frac{K_u V}{k_B T}\right) \tag{7.8}$$

Where the energy to flip the magnetization direction of the entire nanoparticle is related to the anisotropy energy and the entire volume of spins, and is normalized against the thermal energy. This therefore poses a problem only when the magnetic anisotropy is very low or if the particle volume is small. For very high hard drive densities, the volume of these particles can therefore become a problem.

7.4 Dielectric Breakdown

This final section will discuss dielectric breakdown briefly. We discussed dielectric materials as perfectly insulating materials where no conduction could occur. This is not strictly true. In addition, we have seen throughout this class that applying an electric field can increase the potential energy of the electrons in the material with respect to others. The electric field in a dielectric is linear between the electrodes, so one side of the dielectric will have electrons at a higher potential energy than the other. A free electron in the conduction band can move in response to the electric field to minimize its potential energy, which will be converted to kinetic energy. If this kinetic energy is enough to ionize an electron (i.e. of order the band gap energy) from an atom from a scattering event, this can create a cascade effect that can lead to bond breakage, a massive increase in the conductivity locally, and breakdown of the device permanently. If the average mean free path is of order 50 nm and the band gap is about 50 nm, then the total energy required for breakdown is about $E_{br} \approx 1\text{MV cm}^{-1}$. Now that there are two electrons that can conduct, this effect will multiply and lead to catastrophic failure in the device.

This order of magnitude value also ignores defects, microstructure, and other details and are rarely achieved. The native oxide of SiO_2 formed in microelectronics may approach this value, but most other oxides will not. This effect is called electron avalanche.

7.4.1 Thermal breakdown

In addition to this, the finite conductivity of the material will lead to Joule heating that will locally heat up the material. The dielectric loss at high frequencies will also lead to local heating. If this heat cannot be dissipated easily (as is the case for many insulators, which can feature very low thermal conductivities), this increased heating can increase the conductivity of the material, which then heats the material more. This can start to age the dielectric and locally induce breakdown after extended time and use.

7.4.2 Electromechanical breakdown

Finally, we can consider the electromechanical breakdown. Simply put, a capacitor will create charges that are separated by a distance due to the low/nonexistence conductivity of the dielectric. These electrons feel a coulombic interaction through this gap, which leads to a force. This force, after extended periods of time and with sufficiently high fields, can lead to plastic deformation that can induce defects, dislocations, etc. that can induce other failure mechanisms. A sufficiently strong electric field can lead to shear stresses, crack formation, and failure in a mecehanisms known as electrofracture.

In general, these effects all place an upper bound on the electric field that can be applied to a material before it begins to fail. These will also age the material that can place timescales for the operational lifetime of these materials. This is visualized in figure 7.6, which shows both dielectric breakdown from a strong electric field, as well as some relevant field and time scales for these breakdown mechanisms. The point here is simply to say that a dielectric cannot have an infinite voltage applied before they fail.

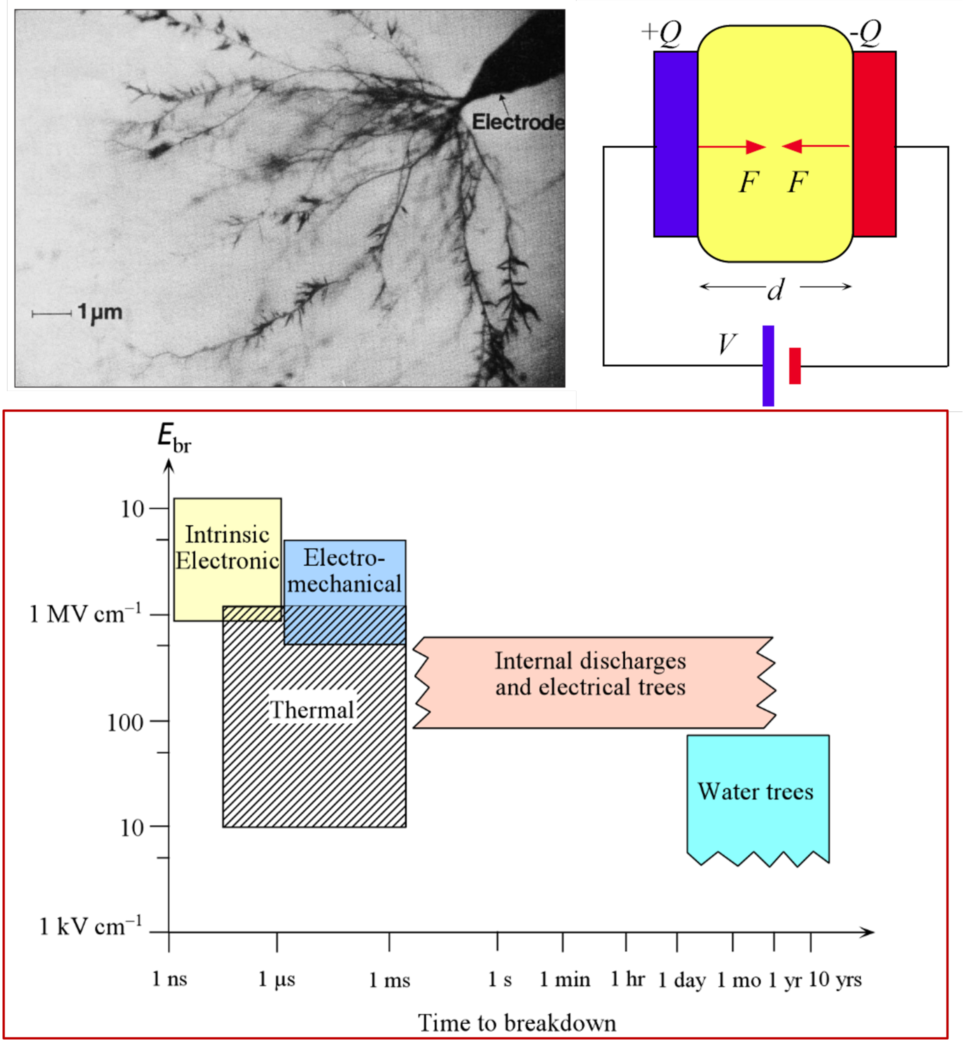


Figure 7.6: Dielectric breakdown examples. Left: Discharge from a high voltage applied to an insulator, with tracks of damage throughout material. Middle: electromechanical breakdown occurs when the coulombic force between the two places starts to induce damage to the dielectric in between. Right: Some fields and timescales for damage. note that not all of these mechanisms are discussed in this course. Adapted from Kasap.