Lecture 3 - Notes

Rahul Arya

January 2019

1 Overview

In the first lecture, we viewed transistors as "magic" components, with the position of a switch between two terminals controlled by the voltage at a third terminal. In the second, we saw that transistors in reality have an internal resistance and capacitance, and studied their effects on the behavior of digital circuits. Now, we will investigate the physical construction of transistors and other semiconductor-based components.

2 Atomic Models

First, though, we will consider various models of the atom. The classical model of an atom is the *planetary model*, with negatively charged, light particles known as *electrons* orbiting around a heavy, positively charged central body known as the *nucleus*, made up of positively charged *protons*. Both protons and electrons have the same magnitude of charge, which we will denote as $\pm e$.

However, this model fails to explain several physical observations - in particular, atomic absorption and emission lines. These lines can be observed by exciting a gas formed of a single type of atom, and measuring the frequency of light emitted. We observe that this emitted light is composed of light of only a few different frequencies - for instance, excited hydrogen gas only emits visible light of the frequencies 656 nm, 486 nm, 434 nm, 410 nm, 397 nm, and 364 nm. Similarly, when monochromatic light at (or very near) these frequencies is shone into hydrogen gas, it is absorbed to a far greater extent than light of other frequencies [1].

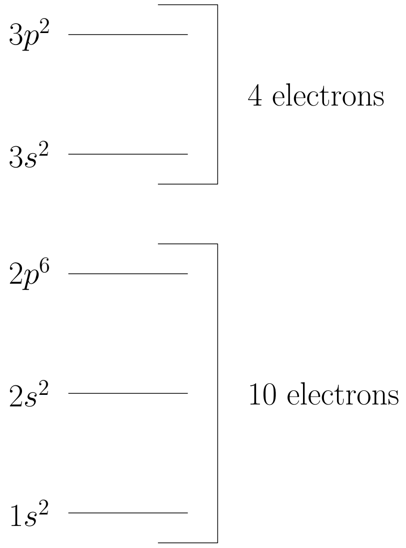
These observations become more interesting when we consider the photon model of light. Here, light is viewed to be composed of small packets of energy known as *photons*. The energy of each packet is dependent solely on its frequency - specifically, $E = hf$, where E is the energy of the photon, h is Planck's constant, and f is the frequency. Thus, we find that atoms emit and absorb photons only

[1] IIRC, the reason that light of other frequencies is also absorbed is due, in part, to the frequency shift of incoming light from the perspective of moving atoms within the gas

with particular frequencies, a phenomenon that is entirely unexplained by the planetary model.

The Bohr model of the atom resolves this issue very elegantly. Its key assumption is that of the quantization of electron *angular momentum* - that is, the angular momentum of the electrons orbiting the nucleus cannot take on arbitrary values, but rather must be a multiple of the constant $\hbar = \frac{h}{2\pi}$. It can be shown that this, in turn, quantizes the energy levels of these electrons, forcing them into discrete *orbitals* around the nucleus, with the orbitals increasing in energy the further they are from the nucleus. When an atom is excited, an electron moves from one energy level E_2 to another of higher energy E_1. When this atom then returns to a ground state, this electron drops back to its original energy level, emitting a photon with energy $\Delta E = E_1 - E_2$, and frequency $f = \frac{E_1 - E_2}{h}$. Clearly, the discrete nature of these energy levels constrains ΔE to a discrete set of frequencies, which form the emission spectrum of the atom.

One property of these atomic orbitals is known as the *Pauli exclusion principle*. For our purposes, we may interpret this property as stating that each energy level can only hold a finite number of electrons. Silicon, which is a semiconductor, has at ground state 14 electrons distributed across the shells $1s^2$, $2s^2$, $2p^6$, $3s^2$, and $3p^2$, as shown:

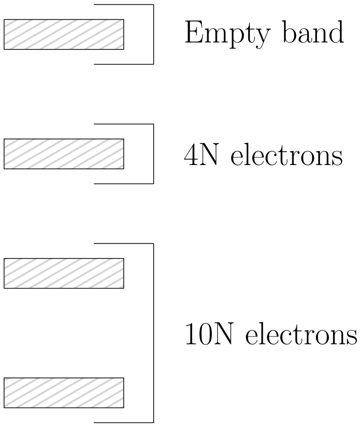


(note that the superscript denotes the number of electrons contained in each energy level).

Note that atoms tend towards lower energy states, which is why energy levels higher than shown are unoccupied. However, the electrons cannot all occupy the lowest energy state, and so form this configuration, which minimizes their total energy while obeying the Pauli exclusion principle. In addition, it is important to note that the 4 electrons at the highest energy levels (known as valence electrons) can move around more freely within a silicon crystal, whereas the lower 10 electrons remain firmly bound to their atom's nucleus.

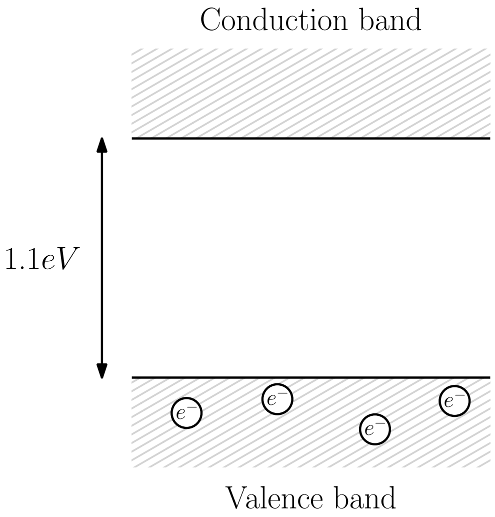
3 A Silicon Crystal

In a silicon crystal of N electrons (where N is very large), additionally (unlike in a single atom), we can view the composition of these discrete energy levels from all the different silicon atoms in the crystal to form *energy bands* as the number of atoms rises, as shown:



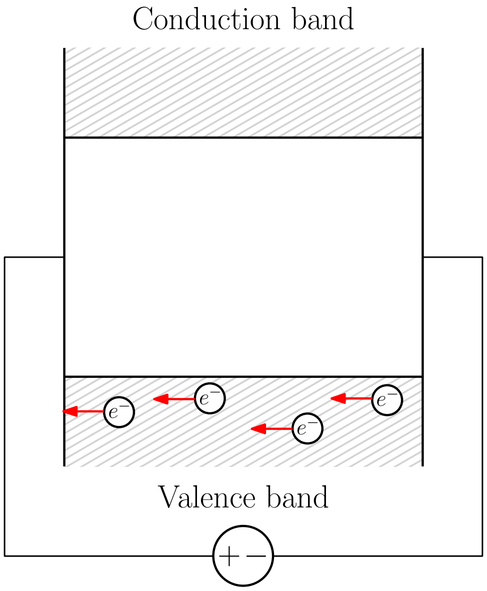
In particular, note that all the valence electrons lie in the highest nonempty energy band - known as the *valence band* - which is fully occupied. Since electrons in lower energy bands don't move between atoms, we are no longer interested in them. Also observe that we have drawn the energy band just above the valence band - this band is known as the *conduction band*, and will shortly be of interest. While electrons may occupy what is effectively a continuous interval of energy levels within energy bands, it is impossible for electrons to remain between these energy bands.

First, we will redraw our picture to focus on just these two bands:



On the left, we have labeled the energy gap between these two energy bands, known as the *band gap*. This represents the minimum energy required for an electron to jump from the valence band to the conduction band or, equivalently, the energy released when an electron drops from the conduction band to the valence band. Note that we label higher energy bands as vertically above lower energy bands, but this is done purely for visual clarity - in reality, electrons of differing energy levels all live around the nucleus of their atom, with electrons further from the nucleus occupying higher energy levels, on average.

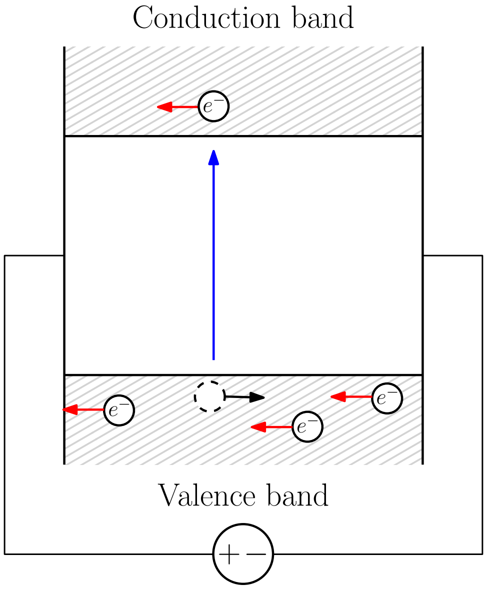
Now, we will apply a voltage across a silicon crystal, and observe what happens to the electrons:



The voltage source produces an electric field within the crystal directed to the right. Thus, the electrons (being negatively charged) experience a force to the left. Despite this force, as the valence band has no empty space, the electrons are unable to move. Thus, when all the electrons are in the valence band, the silicon crystal acts as an insulator.

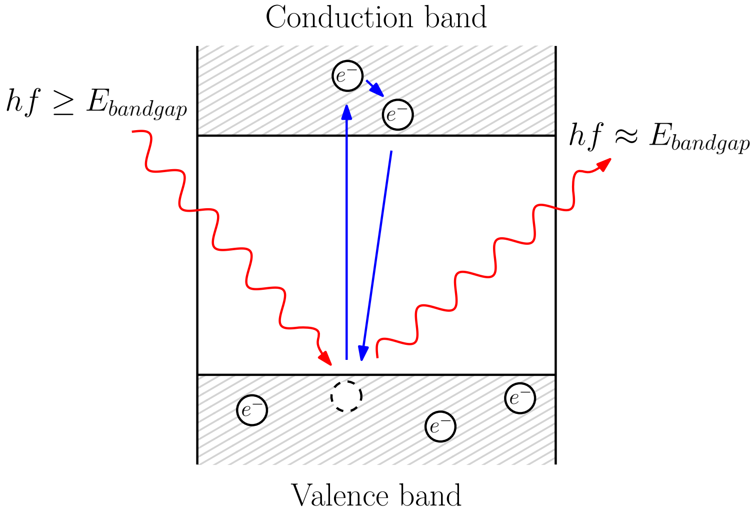
4 Effects of Heat and Light

However, due to thermal excitation, some of the electrons in the valence band may "jump" to the conduction band, leaving a *hole* in the valence band. Alternatively, if the crystal is exposed to light, then electrons may again be excited and jump to a higher energy band. If this is the case, then the excited electron will lie in an almost entirely empty energy band, meaning that it can freely move in response to the electric field, meaning that an current is induced, known as the *electron current*. Less obviously, the hole left in the valence band will appear to move rightwards with the electric field, inducing a further positive current to the right (known as the *hole current*). This mechanism is summarized in the following figure:



Of course, the hole is not a physical object, and so does not experience any forces directly. Rather, the surrounding electrons, driven by the electric field, move to fill the hole, "pushing" the hole to the right. This mechanism is analogous to the motion of bubbles rising in a drink - while they may appear to move upwards, it is really the surrounding drink's motion downwards that creates the bubbles' apparent motion.

As mentioned previously, photons can excite electrons, pushing them into the conduction band. Over a very short duration of time, these electrons will lose energy in the conduction band, falling continuously to the lower bound of the conduction band. As their energy continues to dissipate, they will drop abruptly to the valence band, emitting a photon of a very precise frequency (that corresponding to the band gap), even when the initial photon had energy greater than the band gap, as shown below:



By redirecting and collimating the emitted photons, we are able to produce a tight, coherent, monochromatic beam of light. This process is known as “light amplification by stimulated emission of radiation” - the absorption and reemission of photons being the “stimulated emission of radiation” in the above phrase. Of course, this technology is more commonly known as a *laser*.

Before we move further with our study of semiconductors, it is useful to reflect on why silicon, unlike metals, is considered to be a semiconductor. In a metal, the valence energy band may not be full, meaning that current can flow even when no electrons are excited. Alternatively, the valence and conducting energy bands may overlap, meaning that electrons in the valence band can easily move to and from the conducting band. Again, this means that current can flow in all cases. While this property of metals is (clearly!) extremely useful in producing pure conductors, it is less useful for our purposes.

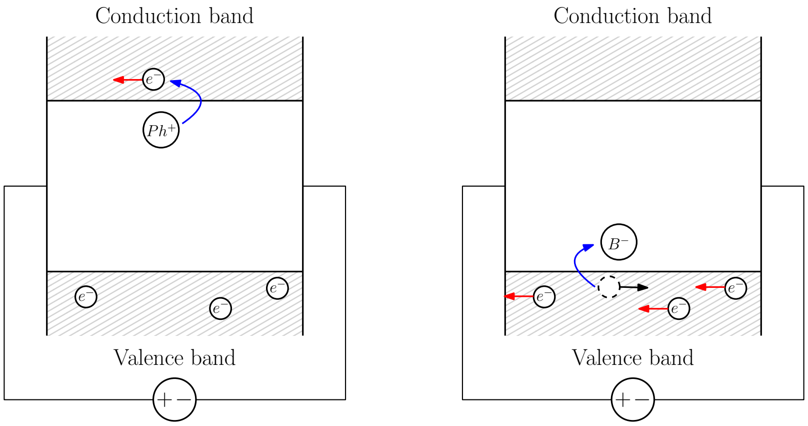
5 Doping

So far, we have seen that semiconductors can conduct a current if their electrons are excited, either by thermal processes or by the absorption of a photon. However, both of these methods are noisy, and very susceptible to changes in the environment. We will now aim to move electrons out of the valence band or into the conduction band in a controlled, stable manner. To do so, we will introduce slight impurities (less than one part per million) in our silicon crystal, in a process known as *doping*.

In one form of doping, we introduce small impurities of phosphorus into the silicon crystal. Phosphorus is one above silicon in the periodic table, and has the property that it can “donate” an electron to the conduction band, leaving behind a static unit of positive charge. Thus, this process enables negative charge carriers to move in a semiconductor - thus, we call it *N-type* doping.

In a similar manner, boron can readily accept an electron from the valence band. Thus, boron impurities create holes in the valence band. As we have seen, these holes behave very much like positive charge carriers, so we term this process *P-type* doping.

It is important to note that, in both of these processes, only one type of charge carrier is created, unlike the previous process of electron excitation, where both electrons and holes formed. The process of doping is summarized in the following figure, with a voltage source applied in order to illustrate the motion of charge carriers in an electric field:

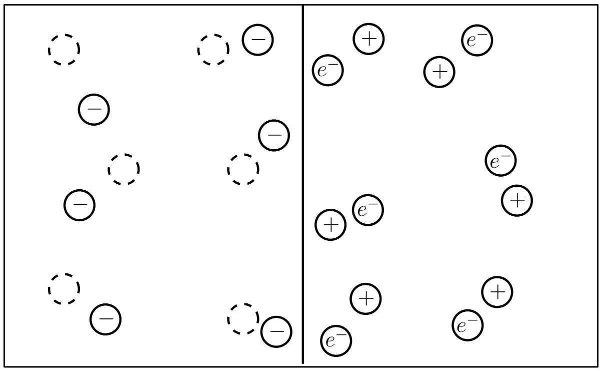


(a) N-type doping.

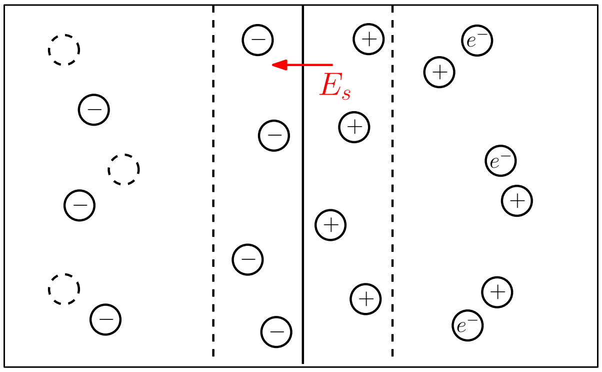
(b) P-type doping.

6 The P-N Junction

On their own, doped semiconductors don’t seem very interesting. However, this changes when we put them together. Consider a component (known as a *p-n junction*) formed by placing a p-type doped silicon crystal (known as the *P-side*) adjacent to an n-type doped crystal (known as the *N-side*), as shown:



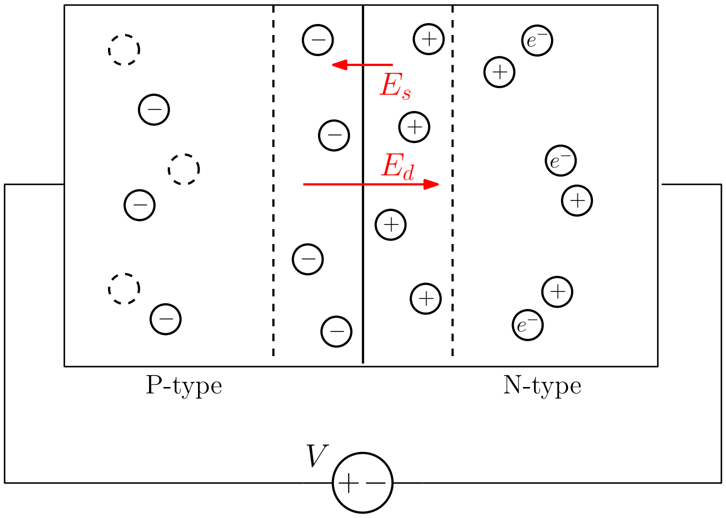
Note that we are no longer drawing the energy bands explicitly, but are instead simply showing the charge carriers - positively charged holes in the P-side, and negatively charged electrons in the N-side. Note that the ions in both regions are static, and so are not charge carriers. As it turns out, at the junction of these regions (known as the *depletion region*), there is a sort of "diffusion" of charge carriers, where some positively charged holes drift over to merge with negatively charged electrons, and vice versa. Thus, at the center of this junction, no charge carriers will be present, as shown:



A good question to ask at this point is the following: why don't *all* of the charge carriers cross over and merge, leaving just the static ions? To resolve this, notice that at the middle of the interface, there are a large number of negative charges to the left and positive charges to the right, producing the electric field E_s to the left. Thus, positively charged holes moving into this region will be pushed back to the P-side, and negatively charged electrons will be pushed back into the N-side.

7 Diodes

Now, consider what happens when we apply a voltage V across this junction, as shown:



If $V_d < 0$ (so the electric field driven by the voltage E_d within the semiconductor goes from from right to left), we simply strengthen the electric field (since E_s and E_d point in the same direction) near the middle of the interface, making it even less likely for charges to flow across the barrier. However, if $V_d > 0$, then we weaken the electric field at the center of the interface (as E_d opposes E_s), allowing charges to rush across the junction, inducing a large current. Thus, we see that this component effectively allows current to flow in one direction - in other words, it is a *diode*.

The above analysis is a simplified form of the ideal behavior of a p-n junction diode. We will now present, without proof, the I-V relation for this diode:

$$I_D = I_S(e^{V_D/V_{th}} - 1).$$

Here, I_D is the diode current, V_D is the voltage applied across the diode (as shown above), I_S is a constant of the material geometry, and V_{th} is the *thermal voltage* of the junction. It is the case that

$$V_{th} = \frac{k_B T}{e},$$

where T is the ambient temperature, and k_B is the Boltzmann constant. Notice that for values of $V_D < 0$, I_D can be negative, approaching I_S for very large negative values of V_D. The mechanism for this effect is out of scope.

However, it turns out that this equation behaves very similarly to our previous model. Substituting in known values for e and k_B, and letting the diode be at room temperature $T = 300\,\mathrm{K}$, we find that

$$V_{th} \approx 25.85\,\mathrm{mV},$$

which is an extremely small value.

For typical material geometries, we find that I_S is of the order of 10^{-12} amperes. Thus, it may appear, since $I_D \propto I_S$, that I_D will also be very small. However, this fails to take into account the exponential term in the coefficient of I_S. For $V_D = 1\,\mathrm{V}$, a reasonably voltage, we find that

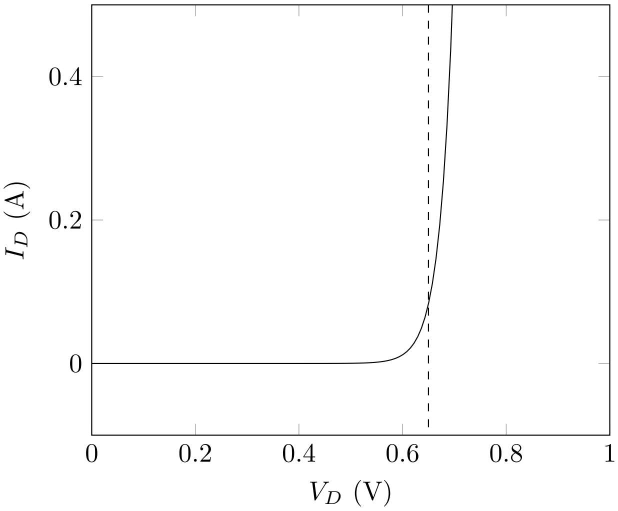
$$\begin{aligned} I_D &= I_S(e^{V_D/V_{th}} - 1) \\ &\approx I_S(e^{38.68} - 1) \\ &\approx I_S(6.28 \times 10^{16} - 1) \\ &\approx I_S(6.28 \times 10^{16}) \\ &\approx 62\,800\,\mathrm{A}, \end{aligned}$$

which is an extremely large current for most lab applications. Thus, due to the small magnitude of V_{th} and the exponential dependence in the equation for I_D, we are able to offset the effects of the small value of I_S.

We will now consider the case where $I_D = 1\,\mathrm{A}$. Solving for V_D,

$$\begin{aligned} && I_D &= I_S(e^{V_D/V_{th}} - 1) \\ \implies && e^{V_D/V_{th}} &= \frac{I_D}{I_S} + 1 \\ \implies && V_d &= V_{th} \ln \frac{I_D}{I_S} + 1 \\ && &\approx V_{th} \ln \frac{I_D}{I_S} \\ && &\approx 0.714\,\mathrm{V}. \end{aligned}$$

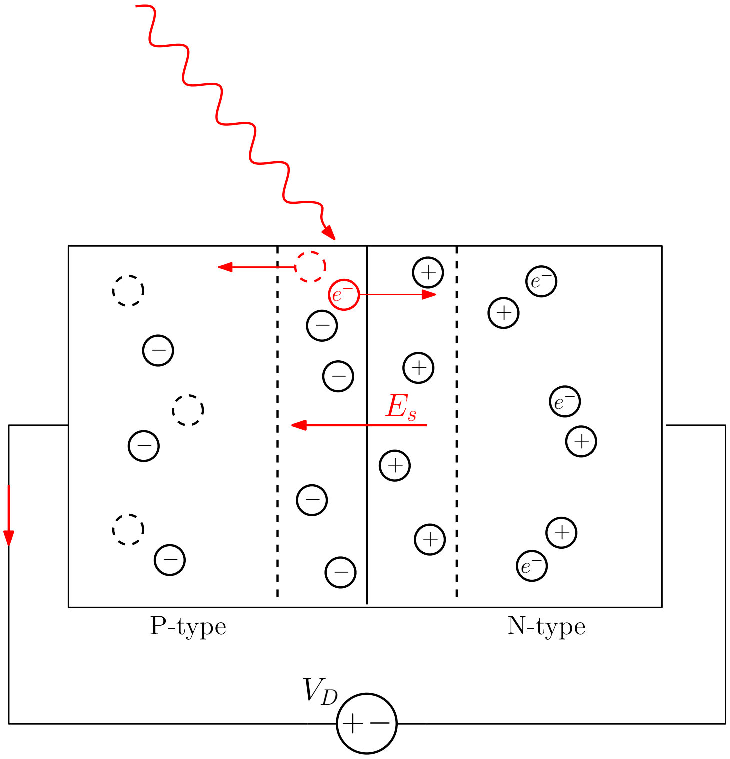
Notice that a small increase in V_D, from just above $0.7\,\mathrm{V}$ to $1\,\mathrm{V}$, gives rise to an over 10000-fold increase in current. Plotting I_D against V_D, we obtain:



We see that the current appears to rapidly increase for $V_D > 0.65\,\mathrm{V}$. This threshold is known as the *knee voltage*, and is between $0.6\,\mathrm{V}$ and $0.7\,\mathrm{V}$ for silicon diodes. Note that changing the scale of the graph will affect the position of the knee voltage, but the values seen here are valid for typical scale levels.

8 Solar Cells and Photodiodes

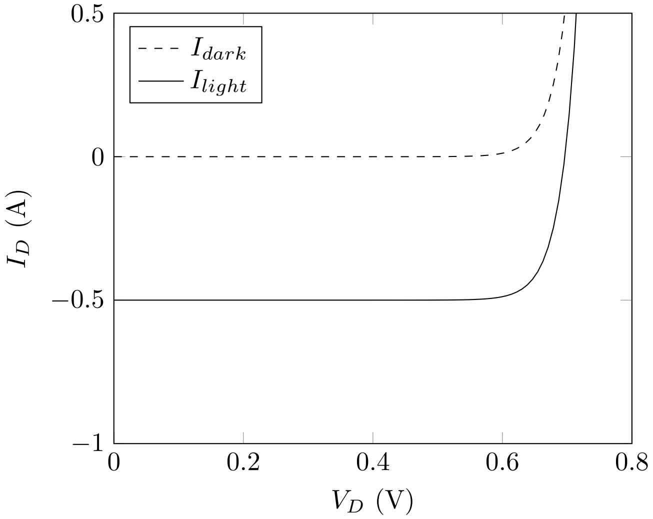
Recall that electrons within a semiconductor are excited by the impact of a photon. We will now use this property in order to create a solar cell, which generates power when exposed to light. Consider the impact of a photon in the depletion region of a p-n junction, as shown:



When the photon impacts, it excites an electron, producing a hole and an electron. The electric field E_s pulls the electron into the N-side and the hole into the P-side. This induces a current against the voltage supply, which generates energy (since this current goes against the voltage supply, we term it *negative current*). Note that in the absence of light, the voltage supply drives a positive current through the junction, wasting energy.

When attempting to measure levels of light precisely, this positive current is undesirable. To eliminate it almost entirely, we can reverse the direction of the voltage supply, setting $V < 0$. Here, in the absence of light, no current will flow. However, when a photon impacts the depletion region, it will drive current through the power supply, but this time the current will be from a high voltage to a low one (since $0 > V$). Thus, negative values of V_D cause energy to be consumed, not generated, in the presence of light - in this case, the component is called a *photodiode*, not a solar cell.

The following plot illustrates the movement of the I-V curve of a solar cell in the presence of light:



Notice that I_{light} is negative, since the current moves against V_D. To maximize the power generated by a solar cell, we must select a V_D that maximizes the product $P = -I_D V_D$, which lies somewhere on the curve I_{light}.

9 LEDs

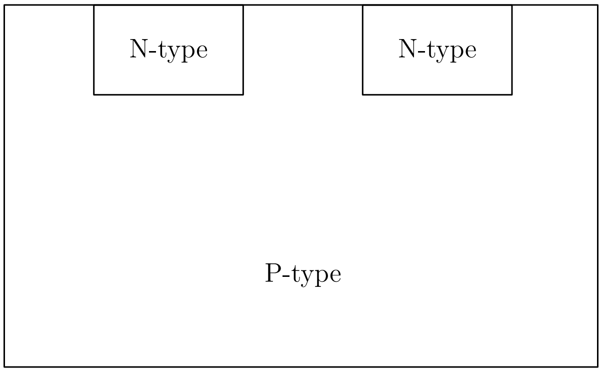
Observe that when we drive current forward through a p-n junction, we waste energy. The energy is dissipated when an electron joins with a positive hole, as the electron must then drop from the conduction energy band down to the valence energy band. Normally, this energy is dissipated as heat. However, it is possible to construct diodes where this energy is instead radiated outwards as a photon, whose energy corresponds to the band gap between the energy bands. Such a light emitting diode is known as a *light emitting diode*[2], or LED. By constructing our LED carefully, we can reduce the energy dissipated as heat to almost 0, making LEDs extremely efficient light sources.

One property of LEDs is that they emit functionally monochromatic light (light of approximately a single frequency), since the frequency of their emitted light depends solely on the band gap of the diode, which is a constant. To change the color of the light, we must use different semiconducting materials - for instance, to produce green light, we must make our LED out of a gallium nitride crystal. As it turns out, we cannot make LEDs out of silicon, since the photon whose energy corresponds to the band gap of silicon lies in the infrared spectrum, not the visible spectrum.

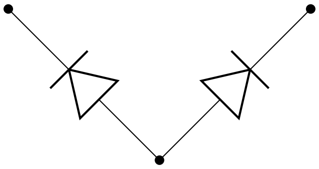
[2] super original, right?

10 Bipolar Junction Transistors

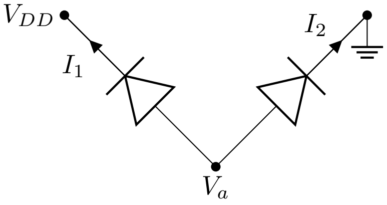
Finally, we will discuss how semiconductors enable the creation of transistors. Consider the following configuration of N-type and P-type doped regions:



We can imagine this configuration as being two diodes connected at a common terminal, as follows:

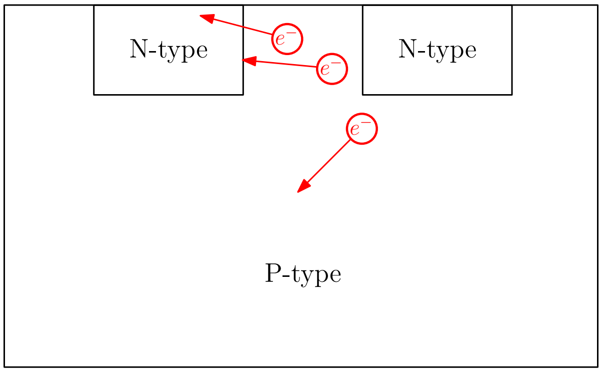


We will attach voltage sources to the three terminals of this component, as follows:



We will set $V_{DD} \approx 10\,\text{V}$, and vary V_a from $0\,\text{V}$ to about $0.7\,\text{V}$ (the knee voltage of our diodes).

Observe that as we raise V_a, the right diode will become forward biased and start to conduct, but the left diode will be reverse biased, so $I_1 \approx 0$. As we reach the knee voltage, I_2 will rise to a relatively large value. Recall from our understanding of p-n junctions that, in diodes conducting current, electrons will flow out of the N-type region, into the P-type region. However, if the two N-type regions are placed together in a suitable manner, most of these electrons will encounter the other N-type region, and be pulled by V_{DD} through the diode. Thus, the left diode will become conducting as well. This mechanism is summarized in the below figure:

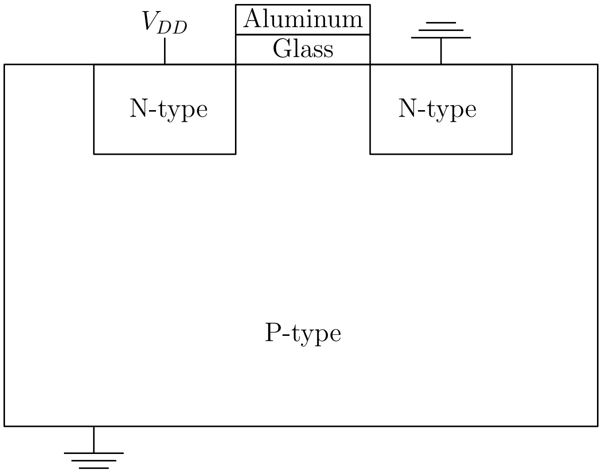


We call the rightmost N-type doped region the *emitter*, and the leftmost N-type doped region the *collector*, due to this motion of electrons. We call the point where the P-type region is connected to V_a the *base*. Thus, we see that when we raise the voltage at the base past some threshold voltage, the component begins to conduct between the collector and the emitter. Thus, this component is a transistor - known as a *bipolar junction transistor*.

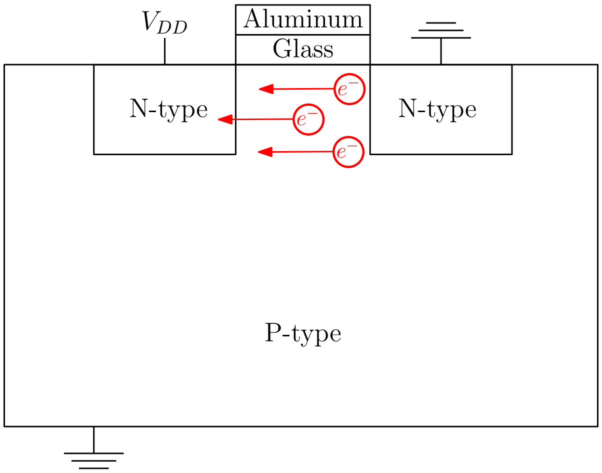
11 MOSFETS

However, one issue with the previous transistor is the fact that, for values of V_a greater than zero but less than the knee voltage, current will be drawn from the base to the emitter. This is undesirable, and to address this we will construct another type of transistor, where no current is ever drawn from the control node.

We will make a small modification to our bipolar transistor, by placing a sheet of glass on the outside of the P-type region, near the two N-type regions. On top of this sheet of glass we will place a sheet of aluminium, as shown:



We will refer to the sheet of aluminium as the *gate*, and apply a positive voltage V_a to it. As this voltage passes some threshold V_{th}, it repels positive holes in its vicinity. We assert, without proof, that the portion of the P-type region in the vicinity of the gate will begin to behave more like a N-type region, allowing electrons in the conduction bands of the adjacent true N-type regions to flow freely. Due to the voltages applied, electrons will be drawn from the rightmost N-type region by the voltage V_{DD}, draining into the leftmost N-type region. Thus, we will refer to the rightmost N-type region as the *source*, and the leftmost N-type region as the *drain*. This mechanism is summarized below:



This component clearly behaves like the transistors we're used to - specifically, an NMOS *MOSFET*. Swapping the P-type and N-type doping will produce a PMOS MOSFET.

Observe that, as glass is an insulator, no current is being drawn from the gate. However, it has some capacitance with the P-type region, which we have previously observed and analyzed. This completes our discussion of semiconductor physics.