Topic 10: PN Junction Introduction

Preface:

This set of notes will introduce many of the concepts needed to analyze devices like diodes, BJTs, and MOSFETs. The mathematics in this section is light compared to others. However, it is imperative that you understand the several assumptions and observations made because they will form the basis for our device analysis.

Device Types and Definition:

Devices seem to be one of the most misused terms within semiconductor physics. When we hear the word devices, we think of electronics. Things like our phones, laptops, computers, and such are marketed as electronic devices. But the origin of devices came from things that rely on the electrical properties of semiconductors [1]. There are other definitions, but the one above from Wikipedia summarizes it well.

The devices we will discuss are:

1. PN Junctions (Diodes & LEDs)
2. Bipolar Junction Transistors (BJT)
3. Metal-Oxide-Semiconductor Systems and Field Effect Transistors (MOS Capacitors & MOSFETS)

This brief list of devices aims at iterating the PN junction is the most fundamental building block of semiconductor devices. Diodes, LEDs, solar cells, BJTs, MOSFETs, thyristors, and many more all rely on the PN junction. Hence, we will take our time over the next sets of notes to understand their behavior comprehensively.

PN Junction:

The PN junction is the simplest of the devices. The junction involves the sandwiching of a p-type and an n-type material together. Note that a PN junction is not fabricated by mechanically sandwiching two semiconductors together. Instead, this metaphor serves as an easy visualization of the junction. Our study so far has focused on either an n or p-type material. We have exhausted our study on single-type semiconductors implying that combining regions seems to be the next natural step. Imagine we have an n and p-type material separate from one another. In an instant, the two regions are brought in perfect contact. The following figures show before and after the combination.

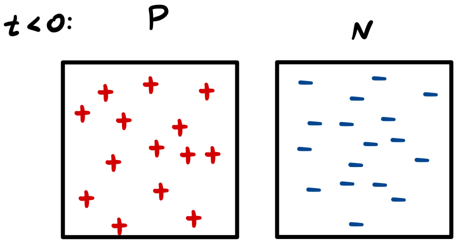

Figure 1: N and p-type before contact.

We should ask ourselves, once the two materials come into contact what would happen? We can visualize the material as one semiconductor that is somehow doped with both types simultaneously. Moments after contact is made diffusion will begin. The majority holes in the p-type will diffuse to the n-type.

[1] https://en.wikipedia.org/wiki/Semiconductor_device

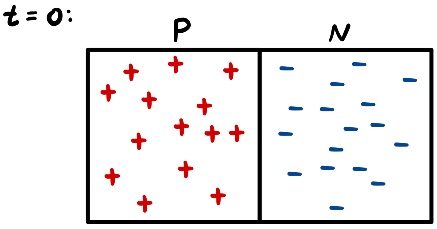

Figure 2: N and p-type when contact is made.

Similarly, the majority electrons in the n-type will diffuse to the p-type. **Electrons exist in the p-type and holes exist in the n-type before contact is made.** It was not drawn in figures 1 and 2, but each material has a nonzero minority carrier concentration. The relatively small minority concentration is responsible for the diffusion of each carrier across the junction.

Let the red and blue arrows represent carrier diffusion across the junction. Figure 3 visualizes diffusion across the junction.

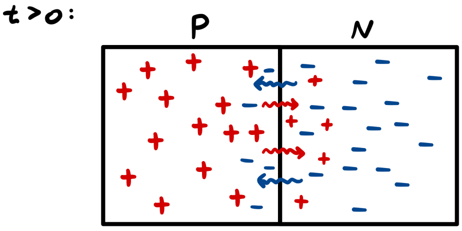

Figure 3: N and p-type when contact is made.

The next question we must ask is if diffusion will stop, and if so, when? Recall our discussion of recombination, what is the physical process and what will increase the recombination rate? As holes enters the n-type region, they will eventually recombine with the majority electrons. Remember that the bottleneck in recombination is the minority carrier concentration; more minority carriers increase the recombination rate. Therefore we will have rampant recombination in both materials relatively near the junction. Eventually, diffusion across the junction must stop. Why? If diffusion could continue infinitely then a net current would be generated from nothing. We also know that each minority carrier has a respective lifetime. Assuming the junction is not infinitesimally small, it is almost guaranteed that each carrier will recombine before reaching the end of the device.

When carriers diffuse across the junction what happens to the parent, or host atom? Recall that phosphorous in the n-type becomes a positive ion when the donor electron is removed, and Boron becomes negative. As the atoms near the junction are stripped of their carriers an ion is left behind. We call these ions **space charge**. Quite literally, they are charges due to ions in the space near the junction. This region near the junction is aptly named the **space charge region**, or SRC. By definition, the space charge region does not have free carriers. The free carriers have diffused across the junction **leaving only space charge**. The figure below visualizes the final state of the semiconductor.

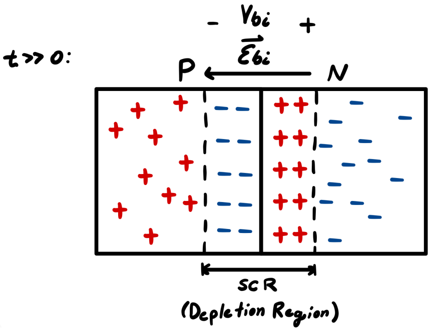

Figure 4: PN junction after diffusion has halted

Space charge region and depletion region are identical. These two terms refer to the same region of the semiconductor, and historically space charge has been the dominant terminology. However, we will use these two terms interchangeably. Take a minute to observe Fig. 4. In the depletion region, we have charges separated by some distance. The first thing that should come to mind are capacitors and electric fields. The space charge induces an electric field within the depletion region. An alternate notation is that a potential exists across the depletion region which forms an electric field. The subscript *bi* is short for built-in. The system has no internal source, so the resulting electric field is built into the device without any intervention from us.

Aside from the formed fields, we need to determine the net charge of each region. We know that a piece of semiconductor must be neutral. Therefore, the regions outside of the depletion region must be neutral. **The neutral region is commonly called the bulk**. The presence of space charge by definition means that the region between the bulk and junction is negative in the p-type and positive in the n-type. Using the conservation of charge we know that the space charge in the p-type must equal the space charge in the n-type. To understand the remainder of our device discussion visualizing the electric field, potential, and conservation of space charge is necessary.

Reiterating some of the above definitions:

- **Space Charge**: Dopant ions present after their carrier has been stripped away.
- **Depletion Region**: The region surrounding the junction which lacks mobile carriers.
- **Built-in Electric Field, $\mathscr{E}_{bi}$**: The built-in electric field due to diffusion.
- **Built-in Voltage, V_{bi}**: The built-in voltage due to diffusion.
- **Thermal Equilibrium**: The time after diffusion has stopped.

Let's take a moment to re-frame the diffusion process. The presence of a built-in electric field is opposite to the diffusion velocity. Holes diffuse to the right, but the built-in electric field acts against this motion. Another way we can then view the arrival at thermal equilibrium is when the built-in field induces a force equal to that diffusing carriers across the junction. Therefore, the built-in electric field and potential will be very involved in our analysis. Once diffusion has stopped, we say the device is in thermal equilibrium.

External Voltage:

The semiconductor in figure 4 is nice, but it is in steady state. The system will not change unless an excitation is applied. We can attach a voltage source in one of two ways: one with the positive connected to the p-type and one with the positive connected to the n-type. Before we reason through the consequences of introducing a voltage source, we must formally state two assumptions.

- **Quasi Neutrality** : The bulk of each region is neutral. Any applied voltage to the device will produce a negligible voltage across the neutral region of each material.
- **Depletion Approximation**: The space charge region does not include mobile carriers.

The two assumptions above were mentioned earlier in this set of notes. We want to give both assumptions a name for easier reference. We will elaborate on these two assumptions later in the notes. However, the assumptions are made to simplify the analysis. Quasi-neutrality eliminates the need to consider the behavior of the neutral region. Any applied voltage across the device appears only across the junction. Similarly, the depletion approximation allows us to ignore any carriers in the space charge region. If we needed to consider the occasional free charge which wanders into the depletion region, the analysis would become much more difficult.

For now, we shift our focus to a high-level walk-through of what will happen under the influence of an external voltage source. The positive terminal is attached to the p-type, and the negative terminal to the n-type. An electric field is formed from the p-type to the n-type region. This external electric field opposes the built-in field and reduces the effect of the built-in field. The smaller built-in field reduces the size of the space charge region. Inversely, if the positive terminal is attached to the n-type, the external field will be aligned with the built-in field. The net result is an increase in the built-in field, and the space charge region increases in size.

Another way to frame this argument is that any applied voltage is dropped purely across the space charge region. Quasi neutrality prevents a substantial voltage drop across the bulk, implying the external source either increases or decreases the built-in potential. Either argument is valid. Visualizing each of these conditions, we get the figures below.

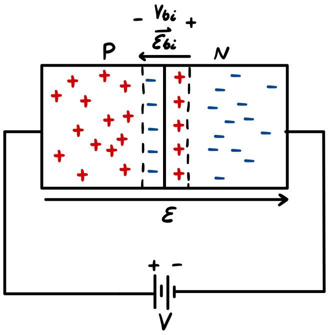

Figure 5: Forward biased PN junction.

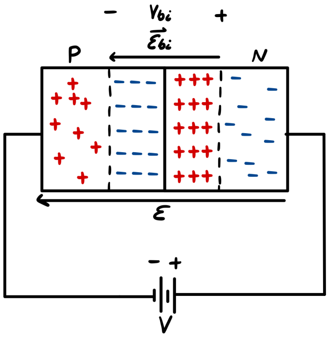

Figure 6: Reverse biased PN junction.

The two figures above are very qualitative. If we apply the external voltage with the positive terminal on the p-type we **forward bias** the junction, and the built-in electric field is reduced. Placing the positive terminal on the n-type **reverse biases** the junction by increasing the built-in field. If we were to apply a forward bias greater than the built-in voltage then diffusion could continue and current would flow. Ideally, a reverse biased PN junction will block any current passing through the device. In reality, the electric field across the entirety of the junction produces drift through the device. The drift current is much much less than that of the forward bias due to minority carriers controlling drift.

Reiterating, under forward bias diffusion is dominant and under reverse bias drift is dominant. The majority carriers from the n and p-type diffuse becoming minority carriers. Whereas reverse bias does not allow the majority carrier to traverse across the junction. A hole in the p-type sees an electric field pushing it back into the p-type. Hence, only holes in the n-type, the minority carrier, can contribute to drift current. This observation is incredibly important as it will simplify our analysis later.

Take a moment to realize the switch-like behavior of the PN junction. If we forward bias the junction it seems like a short circuit. Inversely if we reverse bias the junction we see an open circuit. Hence, the PN junction forms a voltage controlled switch.

Simulation

Before we delve into the analysis NanoLab has an incredibly thorough interactive simulation of a PN junction. The simulation allows you to tweak with the junction and produce curves like the built-in electric field, built-in potential, capacitance versus voltage curves, and many more. The website requires you to log in with UCF credentials. The process is simple and only takes a moment or two. The simulation allows you to play with parameters like doping concentration, size of each junction, material type and is truly an incredible tool. NanoHub PN Junction link: https://nanohub.org/resources/pnjunctionlab

PN Junction Model:

Before we begin our analysis of the PN junction, we should formalize some of our observations regarding the current through the junction. We have studied linear elements such as resistors, capacitors, and inductors. These components have linear current-voltage characteristics. For example, Ohm's law gives us the classic $I = \frac{V}{R}$. The slope of the I-V curve is $\frac{1}{R}$. The PN junction does not adhere to a linear I-V curve. We have shown that for a bias voltage below the built-in potential a small, negative current flows through the device. For biases above the built-in potential diffusion takes over and drastically increases the current. Hence, the I-V curve for a diode is exponential. We will prove this behavior in the next set of notes, however, the general form of the diode current is:

$$I = I_s \left(e^{\frac{qV_{bias}}{kT}} - 1 \right) \tag{1}$$

Let's substituting various voltages for V_{bias} and see if it agrees with our current assumption. For negative bias voltages, the exponential term becomes zero leaving a negative I_s term. I_s is called the **reverse saturation current** and represents the drift current mentioned earlier. The more negative the bias voltage, the closer the current approaches the reverse saturation value. It is helpful to remember that I_s is the max reverse bias current for the junction. For a large, positive, bias voltage the current is dominated by the exponential term. A large exponential compared to -1 implies the -1 term becomes negligible. Therefore, Eq. 1 can be used without knowing if the diode is forward or reverse biased.

Switching our focus to circuits, we need to comment on the complexity associated with non-linear equations. Using techniques like mesh, nodal, Thevenin equivalence, Norton Equivalence, and Two-port networks are all linear models. Two-port networks can be used to characterize non-linear systems, but they reduce to sets of linear equations. For example, the simple voltage divider can be expressed by writing out all the equations and reducing a matrix, relying on the voltage divider formula, or writing KVLs and using Ohm's law. Either way, the equations can be easily manipulated and solved for every unknown. Say we replace one resistor with a diode. The exponential term in the current equation immediately increases the complexity of the system and makes a closed-form often impossible by hand. For example, $x = e^x$ is not a closed-form expression because we cannot solve for x on one side of the equation. We must rely on either graphical or iterative techniques to determine the intersection of the curves. Another option is to guess and check the answer. Solving nonlinear systems of equations are non-issues for computers but pose a significant challenge to hand-analysis. These techniques will be discussed in Electronics 1, but we mention them now because nonlinear equations are rife throughout semiconductor devices.

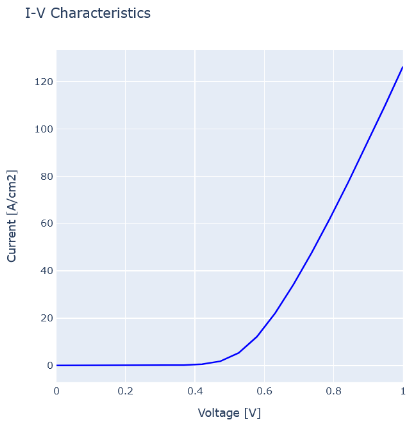

Figure 7: PN Junction I-V characteristic courtesy of [1].

The PN junction I-V curve above was generated from the NanoHub simulator. We can approximate the built-in voltage to be 0.45 V. Below 0.45 V the current is approximately zero and grows to approximately 120 amps in a few hundred millivolts. Note the current density is given in the plot, this does not make a substantial difference.

Analysis:

For our analysis, we begin with a PN junction at thermal equilibrium. Both the n and p-type bulk are shorted to ground, and the depletion region is present.

Energy Band Diagram:

We know how the energy band diagram should look for both an n and p-type before the two materials are connected. The Fermi level is near the valence band in the p-type, and near the conduction band in the n-type. The question we must answer is what happens within the space charge region.

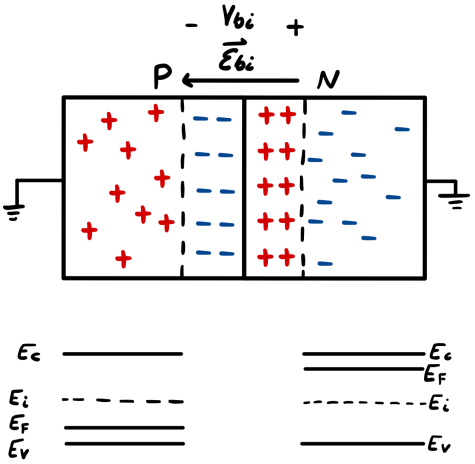

Figure 8: Energy band diagram of PN junction bulk at thermal equilibrium.

We know whenever an electric field is present some drift will occur. It is not uncommon for free carriers to make their way to the depletion region and get swept away by the electric field. Since some component of drift is present we can theorize that the energy bands must bend. Recall that the electric field is encoded within the slope of the energy bands, $\mathscr{E} = \frac{1}{q}\frac{dE_i}{dx}$. If we have a negative electric field that points to the left, then $\frac{dE_i}{dx}$ must be negative. The fundamental charge is a strictly positive constant implying the slope must be negative. A negative slope in the energy bands means they decrease from the p-type to the n-type. Another way of remembering this is that the energy band will always increase in the direction of the electric field.

Hence, we can shift the band diagram on the n-type side downward and draw a straight line connecting the conduction, valence, and intrinsic levels.

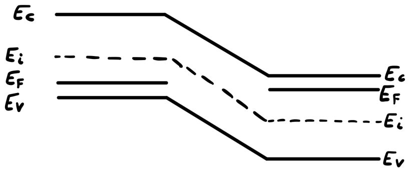

Figure 9: Energy band diagram with influence of built-in electric field.

However, we do not yet know what to do with the Fermi level. We know that the conduction, valence, and intrinsic level must bend under an electric field, but what about the Fermi level? To dig deeper, we need to make the observation that zero net current flows through the device. The lack of current is due to diffusion being blocked by the built-in electric field and some minority carriers drifting through the junction. **A zero net current does not imply all current components are zero**. Since we know that some drift will occur, it must be the case that the drift and diffusion components cancel each other out for both holes and electrons. For holes we have,

$$\begin{aligned} J_{total} = 0 &= J_{diffusion} + J_{drift} \\ &= -qD_p\frac{dp}{dx} + qp\mu_p\mathscr{E}_{bi} \end{aligned} \tag{2}$$

$$= D_p \left(-q\frac{dp}{dx} + \frac{q^2 p}{kT}\mathscr{E}_{bi} \right) \tag{3}$$

The Einstein relation was used to simplify from Eq. 2 to Eq. 3. In the p-type bulk, we know the hole concentration is given by the intrinsic and Fermi level. We can take the derivative with respect to x and plug it into Eq. 3.

$$\begin{aligned} p &= n_i e^{\frac{E_i - E_F}{kT}} \\ \frac{dp}{dx} &= \frac{n_i}{kT} e^{\frac{E_i - E_F}{kT}} \left(\frac{dE_i}{dx} - \frac{dE_F}{dx} \right) \\ \frac{dp}{dx} &= \frac{p}{kT} \left(\frac{dE_i}{dx} - \frac{dE_F}{dx} \right) \end{aligned}$$

The last step took advantage of substituting p back into the equation to remove the need to rewrite the exponential term. The last bit of background is to observe that we know that $\frac{dE_i}{dx} = q\mathscr{E}_{bi}$. Both of these expression may be substituted into Eq. 3 and some terms will simplify. First dividing the D_p term we have,

$$\begin{aligned} 0 &= -q\frac{dp}{dx} + kTp\mathscr{E}_{bi} \\ &= -\frac{qp}{kT}\left(q\mathscr{E}_{bi} - \frac{dE_F}{dx} \right) + \frac{q^2 p}{kT}\mathscr{E}_{bi} \\ &= \frac{qp}{kT}\frac{dE_F}{dx} \end{aligned} \tag{4}$$

The derivation above is only one way of reaching Eq. 4. You could have factored out the mobility term or used the equation for electrons instead. Either way, you will arrive at some constant being multiplied by the derivative of the Fermi level. We know that q, p, k, and T, are all positive and non-zero constants. Implying that if the current density is zero then the **Fermi level must be flat.** This observation is the last we needed to fill in the band diagram.

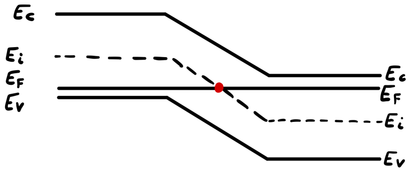

Figure 10: Complete band diagram at thermal equilibrium.

We cannot understate how important the flat Fermi level is. This observation allows us to analyze more complex devices and even extend to heterogeneous semiconductor systems. It will always be our first attempt at trying to draw the band diagram for any mystery semiconductor. Before moving on I want to make a small note of the red dot in Figure 10. We know that $E_i > E_F$ implies the region is a p-type. Inversely, $E_F > E_i$ implies the region is an n-type. A question we should ask is what happens at the intersection between the two curves. Well, this intersection precisely represents the junction between the n and p-types. A good question would be to give you an arbitrary energy band diagram and ask where is the junction? What region is n or p-type? What direction are the built-in electric field and potential? Which side of the junction is more heavily doped? There is no need for practical values, the diagram tells you everything.

Conclusion

I am ending this set of notes before the heavy mathematical analysis of PN junctions. Not only to keep the file as short as possible, but to introduce concepts like thermal equilibrium, depletion region formation, built-in field & potential, and most importantly the flat Fermi level is all prerequisites to the fun analysis. The next set of notes will look solely at the PN junction analytically and two common ways they are doped. We will also derive some of the curves in the NanoHub simulation, such as the built-in electric field and built-in potential.

References

[1] Klimeck Daniel Mejia & Gerhard. *PN Junction Lab*. NanoHub https://nanohub.org/resources/pnjunctionlab. 2021.