HARVEY MUDD COLLEGE Department of Engineering

Introduction to Diodes

Matthew Spencer
Harvey Mudd College
E151 – Analog Circuit Design

1

In this video series we're going to be introducing diodes, which are the first non-linear element we're going to talk about in detail. Non-linear circuit elements violate all the assumptions we made in order to make the highly useful theories of linear circuit analysis, so you might be wondering if we'll be building some new mathematical machinery to analyze non-linear devices. We will soon! But here we're asking a more fundamental question: why is anything non-linear at all? We're going to go to the underlying semiconductor physics of diodes to learn that.

HARVEY MUDD COLLEGE Department of Engineering

What's a Diode?

Matthew Spencer
Harvey Mudd College
E151 – Analog Circuit Design

2

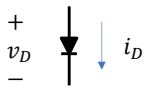
In this video we're going to talk about some of the basics of diodes: what are they and how do you make them?

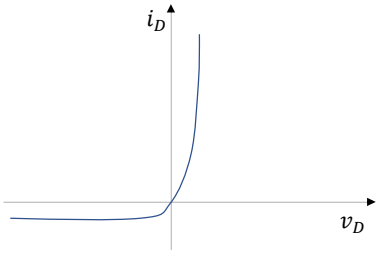
HARVEY MUDD COLLEGE

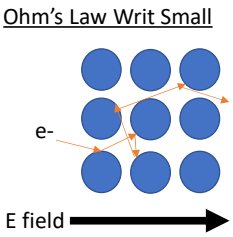
Department of Engineering

Definition of Diodes

- Diodes are:
  - Two terminal circuit elements that only let current flow one way
  - Two terminal circuit elements with an exponential I-V relationship
  - Metallurgical junctions between P and N type semiconductors







3

I'd like you to pause the video and write yourself a definition of diodes. If you've heard multiple definitions, then write them all down. For the record, I came up with three.

CLICK the three definitions I came up with are (1) that diodes are two-terminal circuit elements that let current flow one way, (2) that diodes are two terminal circuit elements with an exponential IV relationship, and (3) that diodes are metallurgical junctions between P and N type semiconductors.

CLICK the first two definitions both kind of say the same thing. That diodes are a circuit element with two terminals, and I've included the standard schematic for a diode above, and that the IV relationship in a diode is kind of surprising. I've drawn an exponential curve on an IV plane above, and you can see that the current on that curve is high and growing when vD is positive, and the current is small and constant when vD is negative. As a result, the claim that "diodes have an exponential IV curve" and "diodes only let current flow one way" are basically saying the same thing.

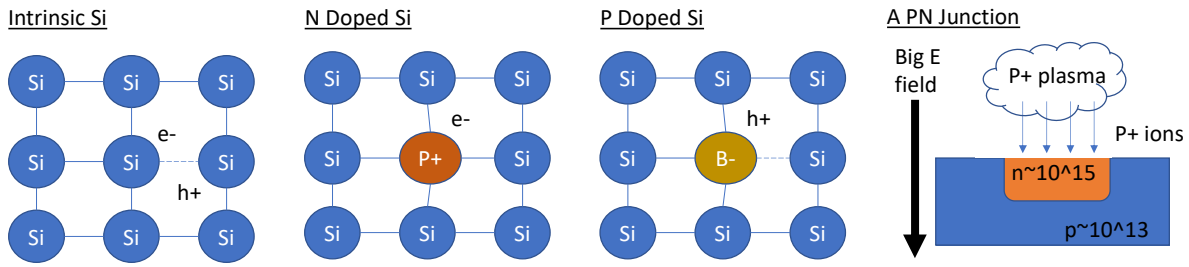
CLICK This exponential behavior is kind of weird though. My mental model for current conduction when I first ran into diodes is captured by this picture, which says that electrons are accelerated by an electric field, but that their maximum velocity is limited by bouncing off of atoms. That picture makes sense, and I can believe it will lead to a linear relationship

between I and V. However, the model says that materials are mute obstacle courses for atoms, and a diode is just made of some material. So how does the diode know which way to let current conduct? How do we back a sense of asymmetry into a material?

Our third diode definition gets at that. It is different than the first two definitions because it speaks to what a diode is rather than what a diode does. Somehow these P and N type semiconductors act differently than resistive materials, and we'll get to the bottom of that in these videos.

HARVEY MUDD COLLEGE Department of Engineering

Doping and physical construction



4

To start solving the mystery of diodes, we should figure out what P type and N type semiconductors are. We start with a picture of intrinsic silicon, which is a pure silicon crystal. Intrinsic silicon is made of silicon atoms that have covalent bonds with four nearby neighbors. I've drawn one of those lines as a dashed line, which indicates a bond that has been broken by thermal energy so that a few free charge carriers have been released into the crystal. Half of these charge carriers are the electrons that used to be tied up in the bond. The other half of the charge carriers are called holes, and they require a bit more description. Holes arise because the absence of the electrons is chemically unstable, so bonds in the crystal will shift around trying to plug that absence. This results in the absence of a bond appearing to move around the crystal, and we elect to treat that absence as a positively charged particle (technically it's a quasiparticle) to simplify modeling.

This is fine, but thermally broken bonds are pretty rare in silicon. The covalent bonds that hold a crystal together are pretty stable, so there aren't many charge carriers. For reference, there are about 10^10 carriers per cubic cm in silicon at room temperature, which is much less than the 5x 10^25 atoms per cubic cm. That means intrinsic semiconductors aren't terribly conductive because mobile charge carriers flowing in a material are the source of conductivity.

However, it's possible to introduce additional carriers to a silicon crystal by doping it, which is the process of deliberately introducing impurities to change a material's behavior. Injecting phosphorous atoms, which have five valence electrons, into a silicon lattice will result in them displacing silicon atoms and leaving an extra free electron behind to roam around. The fixed phosphorous ion that is left behind in the lattice has a net positive charge because it has lost one of its electrons. Silicon with sufficiently many phosphorous atoms in it is called N-doped, where the N stands for negative charge carriers. Similarly, Boron atoms have only three valence electrons, so they will steal an atom from their neighbor to integrate themselves into a lattice, which will result in a free hole and a net negative charge on the boron atom. This type of semiconductor is called P-doped where the P stands for positive charge carriers.

You can make a tight metallurgical junction between these materials, which just refers to two materials being mechanically stuck together, by driving lots of N-type dopants into P-type silicon. One way to do this is to strike a phosphorous plasma above a piece of P-type silicon, then apply a DC bias to the plasma so that ions are hurled into the P-type silicon. Once enough dopants accumulate, the N-type dopants overwhelm the already present P-type dopants resulting in an N-type region that is sometimes called a well or a tub. The interface between this region and the P-type region is called a PN junction. We'll look closer at the physics of a PN junction in the next few videos.

There are lots of details of this process that would be fun to talk about more if we didn't have other pressing business. The exact number of dopants at different depths is called a doping profiles, and real devices are made with more complex doping profiles than this cartoon. It's also worth noting that doping with different materials and dopants can make lots of PN diode-like devices, including common ones like Schottky diodes and Zener diodes, and weird ones, like Gunn diodes. Chase me down to talk more.

Summary

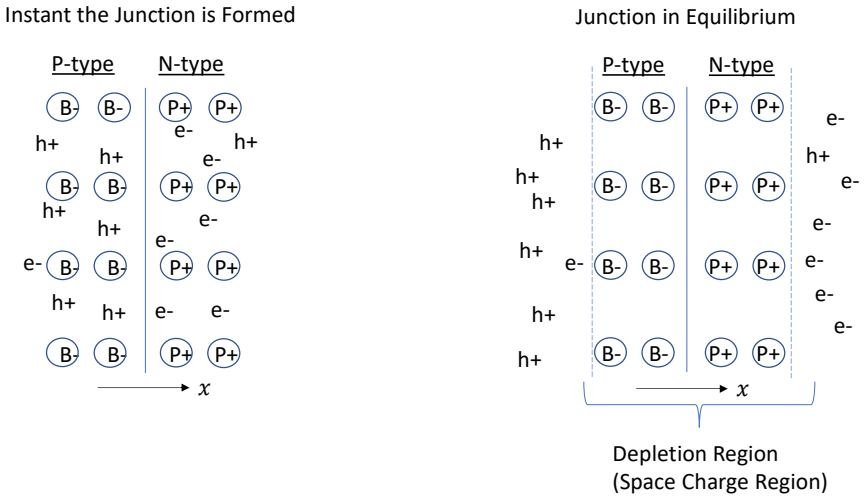
- Diodes have exponential IV relationships, unlike the linear R, C and L
- (Silicon) Diodes are metallurgical junctions of P and N type silicon
- You make P or N silicon with doping to generate more carriers
- You can make junctions by injecting ions into P-type wafers

In this video we're going to examine how charges behave around a PN junction, and we'll discover that they form an electrical bilayer called a depletion region. As a fun fact, this bilayer behavior occurs in lots of different contexts including battery electrodes and cell membranes!

HARVEY MUDD COLLEGE

Department of Engineering

PN Junctions Have Depletion Regions



7

On the left of this slide I've drawn P-type and N-type semiconductors at the very first instant that they make contact. This is a moment that never really exists, but it shows native P and N type semiconductors right next to one another. In each semiconductor region you can see the fixed dopant charges, which can't move because they're stuck in a lattice. They're surrounded by majority charge carriers that have been introduced by the dopants, so holes on the P-type semiconductor and electrons on the N-type semiconductor. In addition, each side has a few minority charge carriers that have been thermally generated. These minority carriers usually aren't long for this world because they electrons annihilate with holes when they meet, but random motion will keep producing new minority carriers.

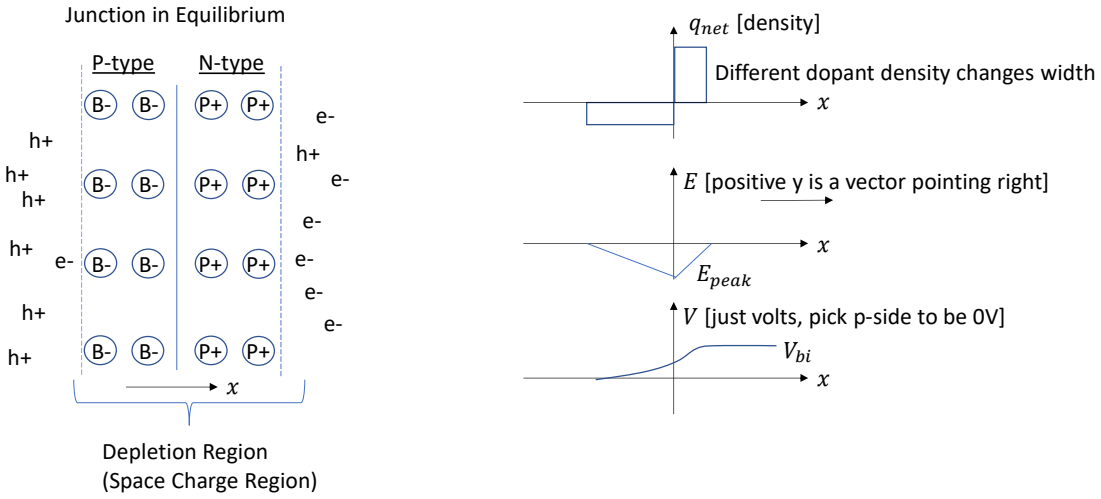
The right side of the picture shows what happens after the two materials settle into equilibrium. The big pile of holes in the P-type material will tend to randomly move into the N-type material, and vice versa for electrons in the N material. This process of random motion of carriers from high concentrations to low concentrations is called diffusion. This inter-diffusion of carriers results in lots of annihilation, which just leaves the fixed dopant charges behind. These dopant charges have an electric field inside of them pointing from the positively charged phosphorous dopants to the negative boron dopants, and that electric field eventually cancels out the effects of diffusion. We'll formalize the descriptions of fields, charges and carrier flows soon, but for now we'd just like to give a name to the

special region of fixed charges. Typically, this region is called the depletion region because it has been depleted of free charge carriers. It is also sometimes called the space-charge region.

HARVEY MUDD COLLEGE

Department of Engineering

Depletion Regions Create an Electric Field



8

We can be more specific about the electric field in the depletion region by making some assumptions about the charge distribution. It's common to assume that the doping profile in the region is uniform and infinitely sharp, which results in the net charge distribution shown in the upper right. This is called the full-depletion approximation. Outside of the depletion region there is no net charge because there are equal numbers of fixed dopants and free charge carriers. Inside the depletion region, the absence of charge carriers means that we have a net charge that matches the polarity of the fixed dopants, so the P-type semiconductor will have a negative net charge because of the fixed Boron dopants and the N-type will have a positive net charge because of the fixed Phosphorous dopants.

I've chosen to draw the height and width of the depletion region as different on either side of the y-axis. That is reflecting that different sides of the junction might have different dopant densities. However, the depletion region was formed by free charge carriers annihilating one-another, and each free charge carrier is associated with a single fixed dopant atom. That means the total charge on either side of the junction has to be the same, so the weakly doped side of the junction will be wider than the heavily doped side. Calculating the exact width is involved.

While we're counting charges, it's worth noting that even though we're drawing a 1D plot of net charge, we're actually talking about a 3 dimensional chunk of material. Representing

the actual total charge of the material would mean these plots would change with the y and z dimensions of the junction, so this plot usually has a charge density on the y-axis.

CLICK We can integrate the charge with respect to x to find the electric field in the depletion region, and it turns out to be this negative shaped triangle. Because we're integrating with respect to x, we'd expect the electric field to point in the positive x direction for positive values of E. However, E is always negative, which means the E field in the junction points in the negative x direction. That's consistent with physical expectations because we expect E fields to point from positive charges towards negative ones, and our positive charges are at higher x values than our negative charges.

Seeing this E field plot ought to excited you because it answers one of the fundamental questions we have about diodes, which is "why are they asymmetric". The E field has a direction to it, so it explains why current passes more easily one way in a diode than the other. It also explains why PN junctions are different our model of resistive materials. We assumed that the only E field acting on charge carriers in our resistive material came from outside, while PN junctions introduce some fields of their own.

The field strength in this junction is pretty high, usually a few megavolts/cm at its peak, which is right at the transition between P and N type material. Sharper doping profiles result in bigger built-in E fields and, as you might guess, if the E field get's big enough then things inside the semiconductor can break down. We'll talk more about that later.

CLICK Finally, we can integrate E field with respect to x, which gives us a voltage. You're always free to pick what you are measuring voltage relative to, and in this case we set the voltage on the P-type semiconductor to zero. We see that the built-in electric field results in a built-in voltage across the PN junction, labelled Vbi Vbi can be used in some sophisticated math to predict properties of the junction, but be careful not to oversimplify diode modeling and use Vbi for something it's not supposed to do. We are only going to use it once more in this class, and that's just to justify something in the next video.

HARVEY MUDD COLLEGE Department of Engineering

Diffusion and Drift Currents

Matthew Spencer
Harvey Mudd College
E151 – Analog Circuit Design

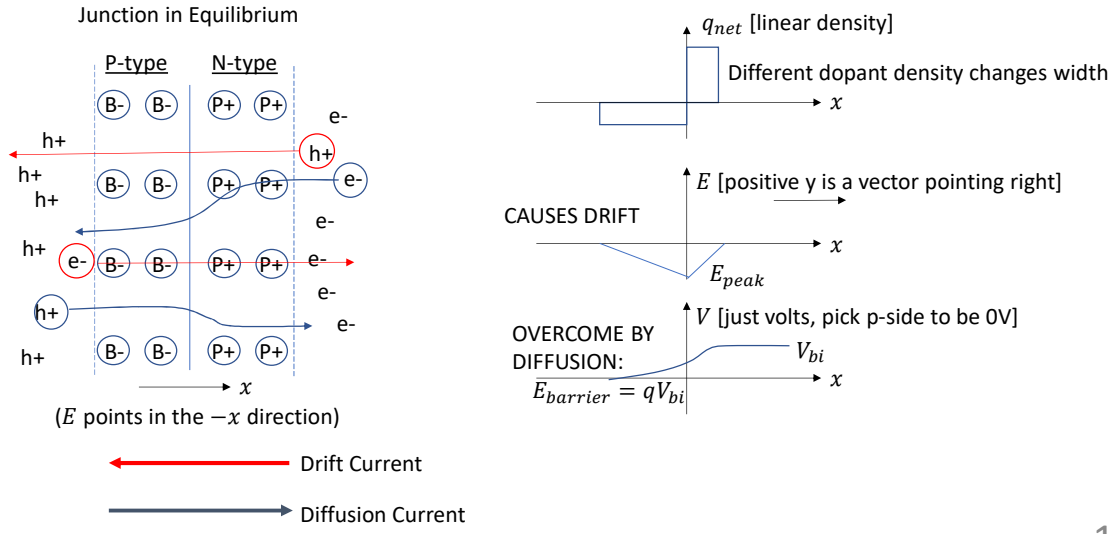
10

In this video we're going to talk about the currents that flow in a PN junction.

HARVEY MUDD COLLEGE

Department of Engineering

Two Types of Current Balance in PN Junctions



11

We need to examine current flows in a PN junctions because we're confronted with a paradox. Common sense says that diodes don't make current without some externally applied energy, so we can't hook up a diode to a flashlight and expect it to light up. However, we've got this E field inside of our diode that will cause carriers to move. Why doesn't that make current shoot out of the edges of the silicon? I'd like you to take a minute to guess, pause the video and write an explanation for yourself about how to resolve this paradox.

The resolution is that this E field isn't the only thing that makes charge carriers move around in PN junctions. There are actually two ways that charge carriers move around in junctions, and we call these two carrier flows by different names: diffusion and drift. The two carrier flows cancel out at equilibrium, which results in no net current flowing through the junction.

CLICK The first carrier flow is the one that got us into this paradox trouble in the first place. Minority carriers on either side of the junction see an electric field pointing towards the opposite side, so if a minority carrier stumbles into the junction then it is swept to the other side by an electric field. This phenomenon is called drift. It results in a current that points in the negative x direction, because holes move in the negative x direction and electrons move in the positive x direction. Because electrons have negative charge, having

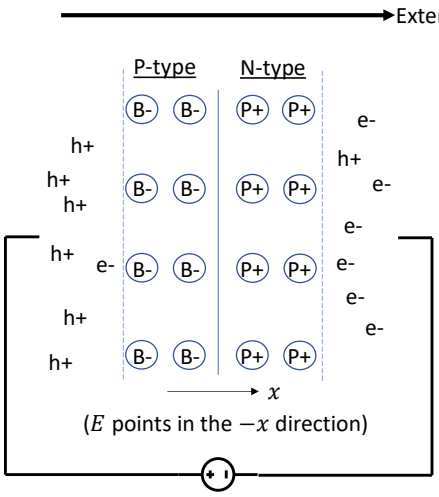
them move in the positive direction contributes to currents in the negative direction. Also note that the negative charge on electrons means they move in the opposite direction that the E field points in.

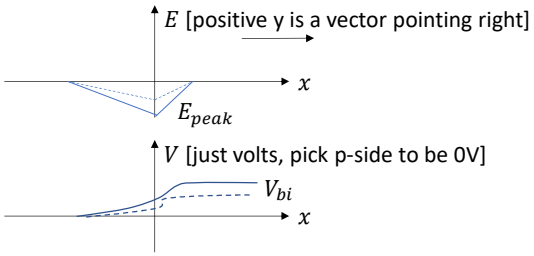
CLICK The second type of current is diffusion. It comes from the random motion of majority charge carriers, which will result in carriers preferentially moving from areas of high concentration to low concentration. Only energetic carriers can diffuse successfully across the depletion region because the E field is exerting a force in the opposite direction of diffusion. Alternately, you can conceive of the effect of the E field as an energy barrier that has a height of q*Vbi where q is the charge on an electron, and carriers with energy higher than that barrier are able to hop over it. Diffusion current points in the positive x direction because it causes holes to move in the positive direction and electrons in the negative direction.

HARVEY MUDD COLLEGE

Department of Engineering

Forward Bias Leads to Exponential Diffusion





A vast simplification: $p_{transit} \propto \exp\frac{qv_D}{nkT}$

Non-ideality factor

12

Apply a voltage to the diode unbalances the two currents by changing the total electric field in the PN junction. This process is complicated, and we're painting a vastly simplified picture on this slide for expediency's sake.

However, if we apply positive voltage to the P side of the PN junction, a condition referred to as forward bias, then an external E field pointing in the opposite direction as the built-in field will be present in the diode. These two fields will cancel out because E fields add, though the dynamics are complicated because the external field also pushes some charges around. However, the upshot is that the external field results in a reduction of the height of the energy barrier that electrons need to cross.

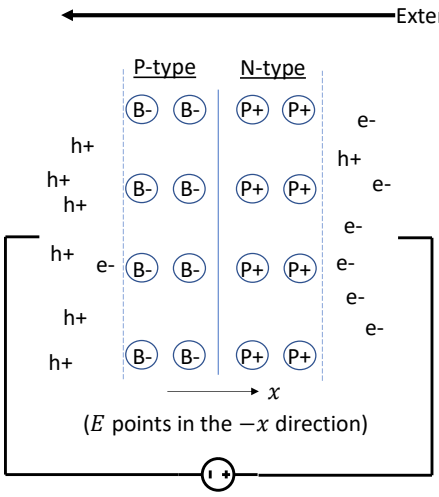
To figure out how many more charge carriers diffuse across the junction when the energy barrier height is changed, we're going to make an analogy with chemical reaction rates. There are similarities between carriers crossing this energy barrier and chemical reactions, in the sense that both involve thermally distributed particles overcoming an energy barrier. As a result, the probability of an electron transiting across the energy barrier can be expressed as an exponential comparing a change in energy barrier height – q times vD – against thermal energy kT where k is the Boltzmann constant and T is temperature. You may note this looks a bit like an Arrhenius equation, and that's because that is also examining the thermal distribution of reactants and comparing it to an energy barrier.

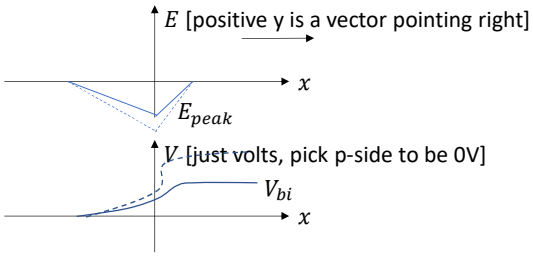
We don’t include Vbi in this equation because we’re measuring changes in our barrier height that are induced by vD. It turns out that Vbi it’s baked into a constant called Is that indicates the baseline amount of diffusion in an equilibrium system. We’ll see Is in action a little later.

We do include one new parameter: a non-ideality factor, n, that captures various imperfections in the PN junction that reduce the total number of charge carriers that transit successfully across the junction.

Department of Engineering

Reverse Bias Leads to a Constant Drift Current





Essentially shuts off diffusion.

Drift limited by # of carriers, no effect

13

Applying a reverse bias to the diode increases the size of the built-in E field and raises the energy barrier. By the same logic as the last slide, this exponentially decreases the probability of transit for diffusing carriers, so drift becomes the dominant method of conduction in reverse biased diodes. However, increasing the electric field doesn't actually increase the amount of drift current because drift is limited by the small number of minority carriers. Those minority carriers were already getting swept across the junction by the small built-in field, so increasing the field doesn't result in a changing current.

However, if you increase the field high enough then carriers in the depletion region start moving really fast. If they move fast enough, then they can knock other electrons loose when they collide with atoms, and those electrons are also accelerated by the high e resulting in a chain reaction of free carriers called avalanche breakdown. Having a high Epeak can cause this chain reaction, so reverse bias is what leads PN junctions to avalanche.

Summary

- Minority charge carriers are pushed across the PN junction by the built-in field, this is called drift.
- Energetic majority charge carriers can randomly jump across the PN junction, this is called diffusion.
- Drift and diffusion balance in equilibrium.
- Diffusion exponentially increases under forward bias (or decreases under reverse bias) because an energy barrier is lowered (or raised).
- Drift is limited by the number of minority carriers and unaffected by the strength of the electric field in the depletion region.

HARVEY MUDD COLLEGE Department of Engineering

The Diode Equation

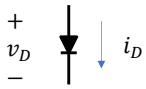
Matthew Spencer
Harvey Mudd College
E151 – Analog Circuit Design

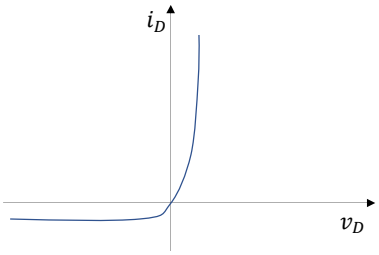
15

In this video we're going to combine the facts we observed about current in the last video to write a single equation that describes current in diodes.

Department of Engineering

Diode Equation Combines Diffusion and Drift





The Diode Equation

$$i_D = I_S\left(\exp\frac{qv_D}{nkT} - 1\right)$$

Dark Current

Diffusion Term

Drift Term

A simplification

$$i_D = I_S\left(\exp\frac{v_D}{n\phi_{th}} - 1\right)$$

The thermal voltage, kT/q
≈ 26 mV at room temp

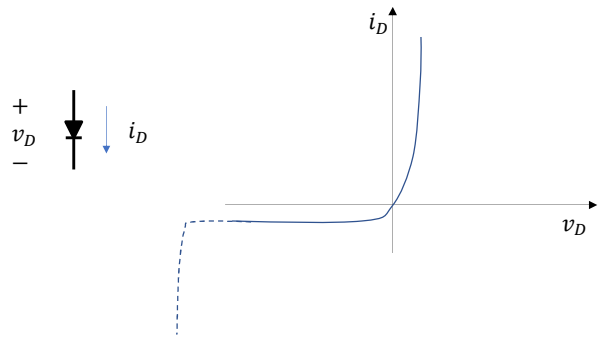
16

This slide has replicated the circuit model and IV curve that we had on the very first slide of this lecture. That IV curve is consistent with the diode equation written on the right of the slide, which combines the exponential probability of transit that is associated with diffusion, a constant to represent drift that is unaffected by external biases, and a constant called the dark current. The dark current is a device parameter that is affected by doping profile, and calculating it relies on the built-in voltage. It indicates the amount of diffusion and drift that are present with no external bias. It's called 'dark current' because photons that impinges on PN junctions can also create free charge carriers, which by the way is how solar cells and photodiodes work, so this dark current is differentiating itself from optical effects.

CLICK There's one common simplification to the diode equation that we're going to use liberally in this class. It's common to define a constant called the thermal voltage, which we'll abbreviate as phi_th to avoid confusion with another constant called a threshold voltage, that is equal to kT/q. This thermal voltage is about 26mV at room temperature, which is easy to remember and compare against measurements in lab.

Department of Engineering

Reverse Bias Causes Avalanche Breakdown



- Caused by chain reaction of e- breakaway at critical E_{peak}
- Surprisingly, it's reversible
- Deliberately used in some diodes

17

The IV curve we've drawn up until this point doesn't capture breakdown behaviors, so I've added avalanche breakdown to the curve on this slide. When reverse bias causes the peak electric field in a PN junction to rise to the point that minority carriers dislodge other minority carriers during their transit, then you get a chain reaction that causes a rapid increase in current in the diode. Remarkably, this process is reversible as long as the high current doesn't light the diode on fire with Joule heating. Avalanche and other related breakdown processes are reliable enough that they are deliberately designed into some diodes. Most notably, Zener diodes use Zener breakdown to make reliable voltage references.

Summary

- The diode equation combines diffusion and drift.

$$i_D = I_S\left(\exp\frac{v_D}{n\phi_{th}} - 1\right)$$

- The thermal voltage simplifies the diode equation, and it's about 26mV at room temperature.

$$\phi_{th} = kT/q$$

- Avalanche breakdown happens at large reverse bias.