Diodes

Carsten Wulff, *2022-11-20*, v0.1.1

***Abstract*—I try to explain how diodes work.**

For the source of this paper, see the markdown.

Why

Diodes are a magical [1] semiconductor device that conduct current in one direction. It's one of the fundamental electronics components, and it's a good idea to understand how they work.

If you don't understand diodes, then you won't understand transistors, neither bipolar, or field effect transistors.

A useful feature of the diode is the exponential relationship between the forward current, and the voltage across the device.

To understand why a diode works it's necessary to understand the physics behind semiconductors.

This paper attempts to explain in the simplest possible terms how a diode works [2]

Silicon

Integrated circuits use single crystalline silicon. The silicon crystal is grown with the Czochralski method which forms a ingot that is cut into wafers. The wafer is a regular silicon crystal, although, it is not perfect.

A silicon crystal unit cell, as seen in Fig. 1 is a diamond faced cubic with 8 atoms in the corners spaced at 0.543 nm, 6 at the center of the faces, and 4 atoms inside the unit cell at a nearest neighbor distance of 0.235 nm.

As you hopefully know, the energy levels of an electron around a positive nucleus are quantized, and we call them orbitals (or shells). For an atom far away from any others, these orbitals, and energy levels are distinct. As we bring atoms closer together, the orbitals start to interact, and in a crystal, the distinct orbital energies split into bands of allowed energy states. No two electrons, or any Fermion (spin of 1/2), can occupy the same quantum state. We call the outermost "shared" orbitial, or band, in a crystal the valence band. Hence covalent bonds.

If we assume the crystal is perfect, then at 0 Kelvin all electrons will be part of covalent bonds. Each silicon atom share 4 electrons with its neighbors. All the neighbors also share electrons, and nowhere is there an vacant state, or a hole, in the valence band. If such a crystal were to exist,

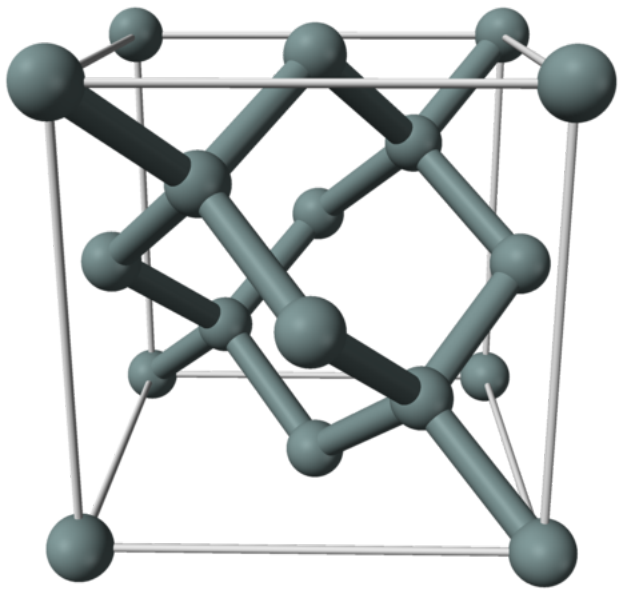

Fig. 1. Silicon crystal unit cell

it would not conduct any current, as the charges cannot move.

In a atom, or a crystal, there are also higher energy states where the carriers are "free" to move. We call these energy levels, or bands of energy levels, conduction bands. In singular form "conduction band", refers to the lowest available energy level where the electrons are free to move.

Due to imperfectness of the silicon crystal, and non-zero temperature, there will be some electrons that achieve sufficient energy to jump to the conduction band. The electrons in the conduction band leave vacant states, or holes, in the valence band.

Electrons can move both in the conduction band, as free electrons, and in the valence band, as a positive particle, or hole.

Intrinsic carrier concentration

The intrinsic carrier concentration of silicon, or how many free electrons and holes at a given temperature, is given by

$$n_i = \sqrt{N_c N_v} e^{-\frac{E_g}{2kT}} \tag{1}$$

where E_g is the bandgap energy of silicon (approx 1.12 eV), k is Boltzmann's constant, T is the temperature in Kelvin,

[1] It doesn't stop being magic just because you know how it works. Terry Pratchett, The Wee Free Men

[2] Simplify as much as possible, but no more. Einstein

N_c is the density of states in conduction band, and N_v is the density of states in the valence band.

The density of states are

$$N_c = 2\left[\frac{2\pi kTm_n^*}{h^2}\right]^{3/2} \quad N_v = 2\left[\frac{2\pi kTm_p^*}{h^2}\right]^{3/2}$$

where h is Planck's constant, m_n^* is the effective mass of electrons, and m_p^* is the effective mass of holes.

In [1] they claim the intrinsic carrier concentration is a constant, although they do mention n_i doubles every 11 degrees Kelvin. In BSIM 4.8 [2] n_i is

$$n_i = 1.45e10\frac{TNOM}{300.15}\sqrt{\frac{T}{300.15}\exp^{21.5565981-\frac{E_g}{2kT}}}$$

Comparing the three models in Fig. 2, we see the shape of BSIM and the full equation is almost the same, while the "doubling every 11 degrees" is just wrong.

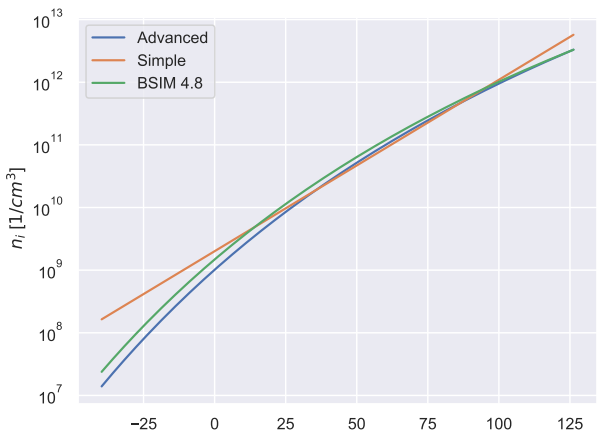

Fig. 2. Intrinsic carrier concentration versus temperature

At room temperature the intrinsic carrier consentration is approximately $n_i = 1 \times 10^{16}$ carriers/m^3.

That may sound like a big number, however, if we calculate the electrons per um^3 it's $n_i = \frac{1\times10^{16}}{(1\times10^6)^3}$ carriers/μm$^3 < 1$, so there are really not that many free carriers in intrinsic silicon.

But where does Equation 1 come from? I find it unsatisfying if I don't understand where things come from. I like to understand why there is an exponential, or effective mass, or Planck's constant. If you're like me, then read the next section. If you don't care, and just want to memorize the equations, or indeed the number of intrinsic carrier concentration number at room temperature, then skip the next section.

It's all quantum

There are two components needed to determine how many electrons are in the conduction band. The density of available states, and the probability of an electron to be in that quantum state.

For the density of states we must turn to quantum mechanics. The probability amplitude of a particle can be described as

$$\psi = Ae^{j(k\vec{r}-\omega t)}$$

where k is the wave number, and ω is the angular frequency, and $\vec{r}$ is a spatial vector.

In one dimension we could write $\psi(x,t) = Ae^{j(kx-\omega t)}$

In classical physics we described the Energy of the system as

$$\frac{1}{2m}p^2 + V = E$$

where $p = mv$, m is the mass, v is the velocity and V is the potential.

In the quantum realm we must use the Schrodinger equation to compute the time evolution of the Energy, in one dimension

$$\frac{1}{2m}\frac{\hbar}{j^2}\frac{\partial^2}{\partial^2 x}\psi(x,t) + V(x)\psi(x,t) = -\frac{\hbar}{j}\frac{\partial}{\partial t}\psi(x,t)$$

where m is the mass, V is the potential, $\hbar = h/2\pi$.

To compute "how many Energy states are there per unit volume in the conduction band", or the "density of states", we start with the three dimensional Schrodinger equation for a free electron

$$-\frac{\hbar^2}{2m}\Delta^2\psi = E\psi$$

I'm not going to repeat the computation here, but rather paraphrase the steps. You can find the full derivation in Solid State Electronic Devices. I'm not sure why the complex parts of the Schrodinger equation dissapeared, but I would assume it's some form of simplification of the real world.

The derivation starts by computing the density of states in the k-space $N(dk) = \frac{2}{(2\pi)^p}dk$ Where p is the number of dimensions (in our case 3).

Then uses the band structure $E(k)$ to convert to the density of states as a function of energy $N(E)$. The simplest band structure, and a approxmiation of the lowest conduction band is

$$E(k) = \frac{\hbar^2k^2}{2m^*}$$

where m^* is the effective mass of the particle. It is within this effective mass that we "hide" the complexity of the actual three-dimensional crystal structure of silicon.

The effective mass is defined as

$$m^* = \frac{\hbar^2}{\frac{d^2E}{dk^2}}$$

as such, the effective mass depends on the localized band structure of the silicon unit cell, and depends on direction of movement, strain of the silicon lattice, and probably other things.

In 3D, once we use the above equations, one can compute that the density of states per unit energy is

$$N(E)dE = \frac{2}{\pi^2}\frac{m^*}{\hbar^2}^{3/2} E^{1/2}dE$$

In order to find the number of electrons, we need the probability of an electron being in a quantum state, which is given by the Fermi-Dirac distribution

$$f(E) = \frac{1}{e^{(E-E_F)/kT}+1} \tag{2}$$

where E is the energy of the electron, E_F is the Fermi level or checmical potential, k is Boltzmann's constant, and T is the temperature in Kelvin.

Fun fact, the Fermi level difference between two points is what you measure with a voltmeter.

If the $E - E_F > kT$, then we can start to ignore the $+1$ and the probability reduces to

$$f(E) = \frac{1}{e^{(E-E_F)/kT}} = e^{(E_F-E)/kT}$$

A few observiation on the Fermi-Dirac distribution. If the Energy of a particle is at the Fermi level, then $f(E) = \frac{1}{2}$, or a 50 % probability.

In a metal, the Fermi level lies within a band, as the conduction band and valence band overlap. As a result, there are a bunch of free electrons that can move around. Metal does not have the same type of covalent bonds as silicon, but electrons are shared between a large part of the metal structure. I would also assume that the location of the Fermi level within the band structure explains the difference in conductivity of metals, as it would determined how many electrons are free to move.

In an insulator, the Fermi level lies in the bandgap between valence band and conduction band, and usually, the bandgap is large, so there is a low probability of finding electrons in the conduction band.

In a semiconductor we also have a bandgap, but much lower than an insulator. If we have thermal equilibrium, no external forces, and we have an un-doped (intrinsic) silicon semiconductor, then the fermi level E_F lies half way between the conduction band edge E_C and the valence band edge E_V. The bandgap is defined as the $E_C - E_V = E_g$, and we can use that to get $E_F - E_C = E_C - E_g/2 - E_C = -E_g/2$. This is why the bandgap of silicon keeps showing up in our diode equations.

The number of electrons per delta energy will then be given by $N_e dE = N(E)f(E)dE$, which can be integrated to get

$$n_e = 2\left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} e^{(E_F-E_C)/kT} \tag{3}$$

For intrinsic silicon at thermal equlibrium, we could write

$$n_0 = 2\left(\frac{2\pi m^* kT}{h^2}\right)^{3/2} e^{-E_g/(2kT)} \tag{4}$$

As we can see, Equation 4 has the same coefficients and form as the computation in Equation 1. The difference is that we also have to account for holes. At thermal equilibrium and intrinsic silicon $n_i^2 = n_0 p_0$.

I've come to the realization that to imagine electrons as balls moving around in the silicon crystal is a bad mental image.

For example, for a metal-oxide-semiconductor field effect transistor (MOSFET) it is not the case that the electrons that form the inversion layer under strong inversion come from somehwere else. They are already at the silicon surface, but they are bound in covalent bonds.

What happens is that the applied voltage at the gate shifts the energy bands close to the surface (or bends the bands in relation to the Fermi level), and the density of carriers in the conduction band in that location changes, according to the type of derivations above.

To make matters more complicated, an inversion layer of a MOSFET is not in three dimensions, but rather one must compute for two dimensions, as the density of states is confined to the silicon surface.

Once the electrons are in the conduction band, then they follow the same equations as diffusion of a gas, Fick's law of diffusion. Any charge concentration difference will give rise to a diffusion current given by

$$J_{\text{diffusion}} = -qD_n\frac{\partial\rho}{\partial x} \tag{5}$$

where J is the current density, q is the charge, ρ is the charge density, and D is a diffusion coefficient that through the Einstein relation can be expressed as $D = \mu kT$, where mobility $\mu = v_d/F$ is the ratio of drift velocity v_d to an applied force F.

Careful with the mobility μ though, since the inversion layer of a MOSFET is a two dimensional electron gas, so will have a different μ than in three dimensional bulk silicon.

Doping

We can change the property of silicon by introducing other elements, something we've called doping. Phosphor has one more electron than silicon, Boron has one less electron. Injecting these elements into the silicon crystal lattice changes the number of free electron/holes.

These days, we usually dope with ion implantation, while in the olden days, most doping was done by diffusion. You'd paint something containing Boron on the silicon, and then heat it in a furnace to "diffuse" the Boron atoms into the silicon.

If we have an element with more electrons we call it a donor, and the donor concentration N_D.

The main effect of doping is that it changes the location of the Fermi level at thermal equilibirum. For donors, the Fermi level will shift closer to the conduction band, and increase the probabilty of free electrons, as determined by Equation 2.

Since the crystal now has an abundance of free electrons, which have negative charge, we call it n-type.

If the element has less electrons we call it an acceptor, and the acceptor concentration N_A. Since the crystal now has an abundance of free holes, we call it p-type.

The doped material does not have a net charge, however, it's the same number of electrons and protons, so even though we dope silicon, it does remain neutral.

The doping concentrations are larger than the intrinsic carrier concentration, from maybe 10^{21} to 10^{27} carriers/m^3. To separate between these concentrations we use $p-, p, p+$ or $n-, n, n+$.

The number of electrons and holes in a n-type material is

$$n_n = N_D \ , \ p_n = \frac{n_i^2}{N_D}$$

and in a p-type material

$$p_p = N_A \ , \ n_p = \frac{n_i^2}{N_A}$$

In a p-type crystal there is a majority of holes, and a minority of electrons. Thus we name holes majority carriers, and electrons minority carriers. For n-type it's opposite.

PN junctions

Imagine an n-type material, and a p-type material, both are neutral in charge, because they have the same number of electrons and protons. Within both materials there are free electrons, and free holes which move around constantly.

Now imagine we bring the two materials together, and we call where they meet the junction. Some of the electrons in the n-type will wander across the junction to the p-type material, and visa versa. On the opposite side of the junction they might find an opposite charge, and might get locked in place. They will become stuck.

After a while, the diffusion of charges across the junction creates a depletion region with immobile charges. Where as the two materials used to be neutrally charged, there will now be a build up of negative charge on the p-side, and positive charge on the n-side.

The charge difference will create a field, and a built-in voltage will develop across the depletion region.

The magnitude of the built-in voltage can be computed from Fermi-Dirac distribution, stating that the average number of fermions in a single-particle state j is given by

$$n_j = \frac{1}{e^{(E_j-\mu)/kT}+1}$$

where E_j is the energy of the single-particle state, μ is the chemical potential (or the Fermi Level).

Assuming the exponential is much larger than 1, and taking the ratio number of free electrons on the n-side, and p-side, we get

$$\frac{n_n}{n_p} = \frac{e^{(E_{n_p}-\mu)/kT}}{e^{(E_{n_n}-\mu)/kT}} = e^{\frac{E_{n_p}-E_{n_n}}{kT}}$$

where $E_{n_p} - E_{n_n}$ is the energy difference between electrons on the p-side and n-side. This energy difference is equivalent to $q\Phi_0$ where Φ_0 is the built-in voltage. As a result, and inserting for n_p, we get

$$\frac{N_A N_D}{n_i^2} = e^{\frac{q\Phi_0}{kT}}$$

or

$$\Phi_0 = \frac{kT}{q} ln\left(\frac{N_A N_D}{n_i^2}\right)$$

Current

As mentioned before, we continuously have electron/hole pairs generated by the temperature. In addition, we can have electron/hole pairs generated by for example photons (photo diodes), or impact ionization (charges at high speed, like radiation). Those electron/hole pairs that come into existence in the depletion region, or happen to wander into the depletion region before recombining will be swept across the depletion region due to the electric field. The electron will drift to the n-type, and holes will drift to the p-type. This drift creates a leakage current in the diode.

To estimate the leakage current we would need to know how many electron/hole pairs are generated per second, and how many reach the depletion region before recombining. Not at all a trivial calculation.

If we apply a voltage in the forward direction, opposite to the field, the current will be

$$I_D = I_S(e^{\frac{V_D}{V_T}} - 1)$$

where V_D is the voltage across the diode, $V_T = \frac{kT}{q}$ and

$$I_S = qAn_i^2\left(\frac{D_n}{L_nN_A} + \frac{D_p}{L_pN_D}\right)$$

where A is the area of the diode, D_n,D_p is the diffusion coefficient of electrons and holes and L_n,L_p is the diffusion length of electrons and holes.

Forward voltage temperature dependence

We can rearrange I_D equation to get

$$V_D = V_T \ln\left(\frac{I_D}{I_S}\right)$$

and at first glance, it appears like V_D has a positive temperature coefficient. That is, however, wrong.

First rewrite

$$V_D = V_T \ln I_D - V_T \ln I_S$$

$$\ln I_S = 2\ln n_i + \ln Aq\left(\frac{D_n}{L_nN_A} + \frac{D_p}{L_pN_D}\right)$$

Assume that diffusion coefficient [3], and diffusion lengths are independent of temperature.

That leaves n_i that varies with temperature.

$$n_i = \sqrt{B_cB_v}T^{3/2}e^{\frac{-E_g}{2kT}}$$

where

$$B_c = 2\left[\frac{2\pi km_n^*}{h^2}\right]^{3/2} \quad B_v = 2\left[\frac{2\pi km_p^*}{h^2}\right]^{3/2}$$

$$2\ln n_i = 2\ln\sqrt{B_cB_v} + 3\ln T - \frac{V_G}{V_T}$$

with $V_G = E_G/q$ and inserting back into equation for V_D

$$V_D = \frac{kT}{q}(\ell - 3\ln T) + V_G$$

Where ℓ is temperature independent, and given by

$$\ell = \ln I_D - \ln\left(Aq\frac{D_n}{L_nN_A} + \frac{D_p}{L_pN_D}\right) - 2\ln\sqrt{B_cB_v}$$

From equations above we can see that at 0 K, we expect the diode voltage to be equal to the bandgap of silicon. Diodes don't work at 0 K though.

Although it's not trivial to see that the diode voltage has a negative temperature coefficient, if you do compute it as in vd.py, then you'll see it decreases.

The slope of the diode voltage can be seen to depend on the area, the current, doping, diffusion constant, diffusion length and the effective masses.

Fig. 3 shows the V_D and the deviation of V_D from a straight line. The non-linear component of V_D is only a few mV. If we could combine V_D with a voltage that increased with temperature, then we could get a stable voltage across temperature to within a few mV.

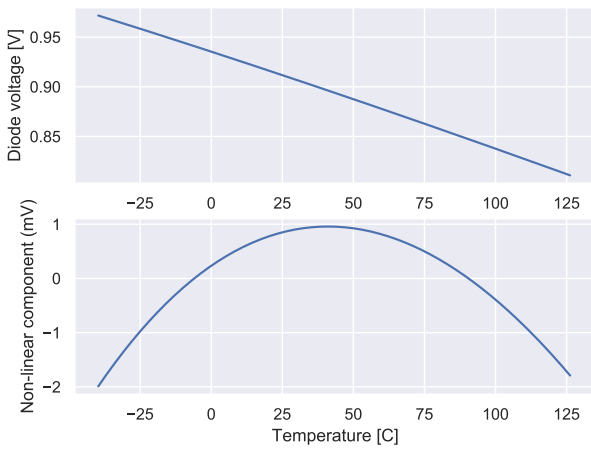

Fig. 3. Diode forward voltage as a function of temperature

Bandgap references

Assume we have a circuit like Fig. 4 Here we have two diodes, biased at different current densities. The voltage on the left diode V_{D1} is equal to the sum of the voltage on the right diode V_{D2} and voltage across the resistor R_1. The current in the two diodes are the same due to the current mirror. A such, we have that

$$I_Se^{\frac{qV_{D1}}{kT}} = NI_Se^{\frac{qV_{D2}}{kT}}$$

Taking logarithm of both sides, and rearranging, we see that

$$V_{D1} - V_{D2} = \frac{kT}{q}\ln N$$

[3] From the Einstein relation $D = \mu kT$ it does appear that the diffusion coefficient increases with temperature, however, the mobility decreases with temperature. I'm unsure of whether the mobility decreases with the same rate though.

Or that the difference between two diode voltages biased at different current densities is proportional to absolute temperature.

In the circuit above, this ΔV_D is across the resistor R_1, as such, the $I_D = \Delta V_D / R_1$. We have a current that is proportional to temperature.

If we copied the current, and sent it into a series combination of a resistor R_2 and a diode, we could scale the R_2 value to give us the exactly right slope to compensate for the negative slope of the V_D voltage.

The voltage across the resistor and diode would be constant over temperature, with the small exception of the non-linear component of V_D.

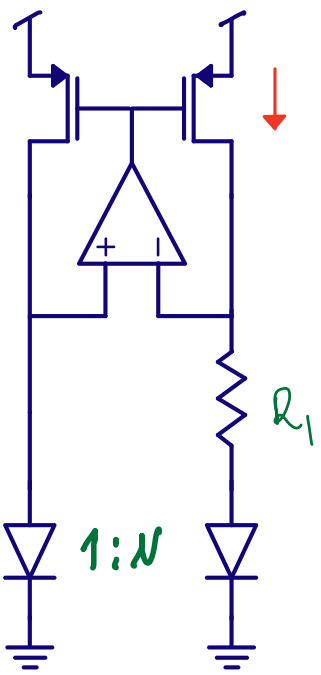

Fig. 4. Circuit to generate a current proportional to kT

References

[1] T. C. Carusone, D. Johns, and K. Martin, *Analog integrated circuit design*. Wiley, 2011 [Online]. Available: https://books.google.no/books?id=1OIJZzLvVhcC

[2] Berkeley, "Berkeley short-channel IGFET model." [Online]. Available: http://bsim.berkeley.edu/models/bsim4/