# Diodes

Carsten Wulff, *2022-05-09*, v0.1.0

***Abstract*—I 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 an extremely 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.

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, it would not conduct any current, as the charges cannot 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. In the conduction band electrons are free to move. 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, by jumping from hole to hole.

[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

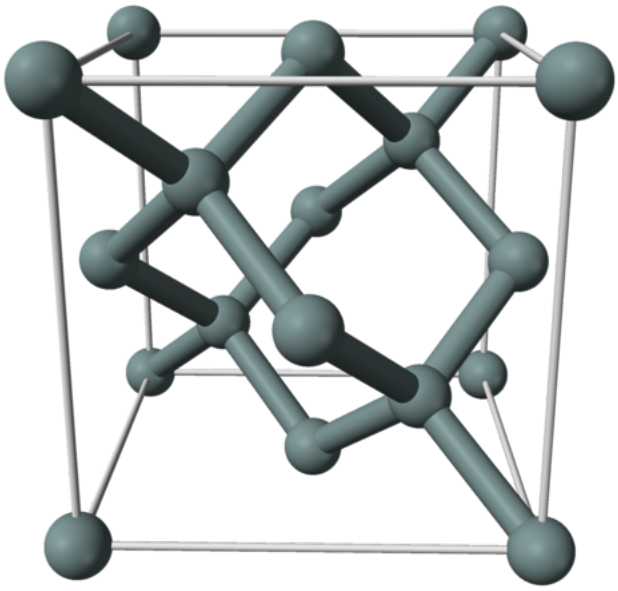

Fig. 1. Silicon crystal unit cell

The movement of electrons in solids seem to obey the mathematics of quantum mechanics. I don't know why electrons move as they do, but I do know that physicists tell me that no evidence has been found that contradicts the mathematics of quantum mechanics.

If you wanted to try an compute movement of electrons, then you must deal with the complex probability amplitudes, or the wave function, given by

$$\psi(x,t) = Ae^{j(kx-\omega t)}$$

where $k$ is the wave number, and $\omega$ is the angular frequency. To compute the time evolution of the wave function you could use the Schrodinger equation, or

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

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

Now I could say "one can easily see that", or "it could be shown that", but in all honesty, I don't understand how electrons move.

Assume an electron with a certain momentum passes close enough to a Boron atom with a vacant state such that the wave functions of the vacant state and the electron

interact. An electron can scatter off the energy boundary, or get stuck in the hole. We cannot say for certain what will happen for a particular electron, but in general we can provide the probability of events.

Drawing an analogy to the real world, imagine a golf ball rolling towards the edge of a large empty pool. The behavior of the electron would be like the golf ball sometimes would turn around at the edge of the pool and come back, and other times, it would fall in. That's just weird. But it is how quantum mechanics tells us the small world works.

Imagine now a 1 $\mu m^3$ piece of silicon. That has $(1\mu)^3/(0.543\text{n})^3 = 6.2\text{G}$ unit cells. Not all of them are perfect. Some might be missing an atom, or have a impurity atom, or maybe a missing unit cell. How do we compute the movement of electrons with wave function in such a system? I'm not sure we can, as in, I'm not sure anyone in the world actually knows how to do that. We have to make assumptions, we have to make simplifications, we have to ignore the fact that the world is not perfect.

Instead of trying to envision electrons as particles moving around in the crystal, try to think of electrons, and holes, as a gas. We don't know exactly how every atom of the gas will behave, however, we can say in general that if there is a density difference, then there will be a flow of gas atoms that follow Fick's law of diffusion. Same in silicon. Any charge concentration difference will give rise to a diffusion current given by

$$I_{\text{diffusion}} = -qD_n \frac{\partial \rho}{\partial x}$$

where $q$ is the charge, $\rho$ 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 $F$.

## Intrinsic carrier concentration

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

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

where $E_g$ is the bandgap energy of silicon (approx 1.12 eV), $k$ is Boltzmann's constant, $T$ is the temperature in Kelvin, $N_c$ 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. The effective mass of electrons and holes in silicon depend on direction of movement, strain of silicon, and I'm not entirely sure what is the correct number to use when computing density of states.

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.

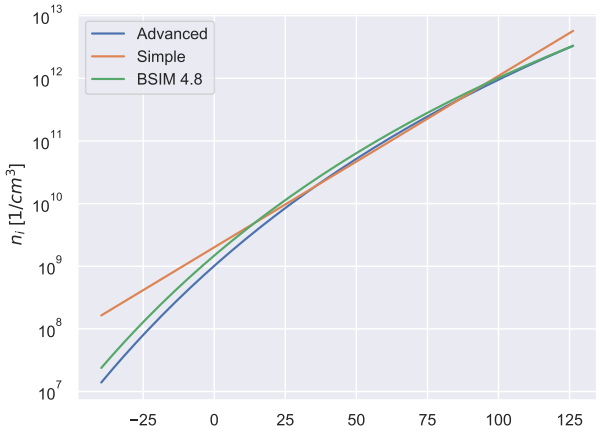


Fig. 2. Intrinsic carrier concentration versus temperature

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/$\mu$m$^3 < 1$, so there are really not that many free carriers in intrinsic 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 $N_D$. 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, $\mu$ is the chemical potential.

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 $\Phi_0$ 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

$$I_S = qAn_i^2\left(\frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D}\right)$$

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

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

$$\ln I_S = 2 \ln n_i + \ln Aq \left( \frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right)$$

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

That leaves $n_i$ that varies with temperature.

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

where

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

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

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 $\ell$ is temperature independent, and given by

$$\ell = \ln I_D - \ln \left( Aq \frac{D_n}{L_n N_A} + \frac{D_p}{L_p N_D} \right) - 2 \ln \sqrt{B_c B_v}$$

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.

[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.

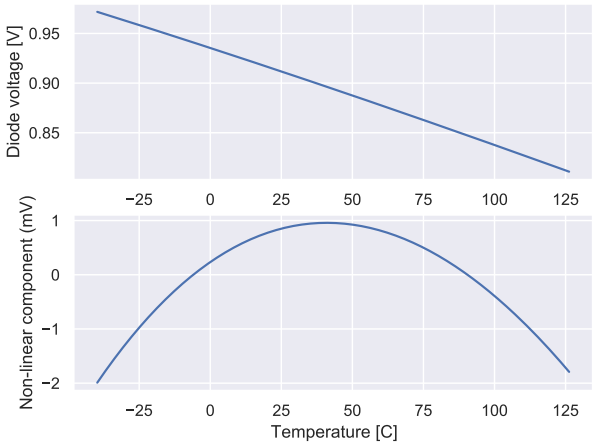


Fig. 3. Diode forward voltage as a function of temperature

## 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

$$I_S e^{\frac{qV_{D1}}{kT}} = N I_S e^{\frac{qV_{D2}}{kT}}$$

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

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

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

In the circuit above, this $\Delta V_D$ 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 $V_D$.

## 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

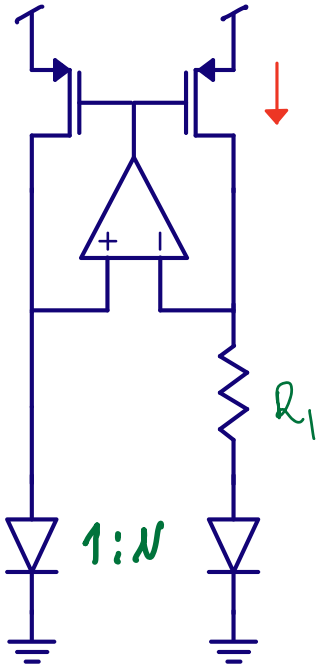


Fig. 4. Circuit to generate a current proportional to kT

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