Diodes

Carsten Wulff, carsten@wulff.no

Keywords: Silicon, Intrinsic Carriers, Density of States, Fermi-Dirac, Doping, PN Junction, Built-in Voltage, Diode Equation, Temperature, Leakage

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 an 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 Figure 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.

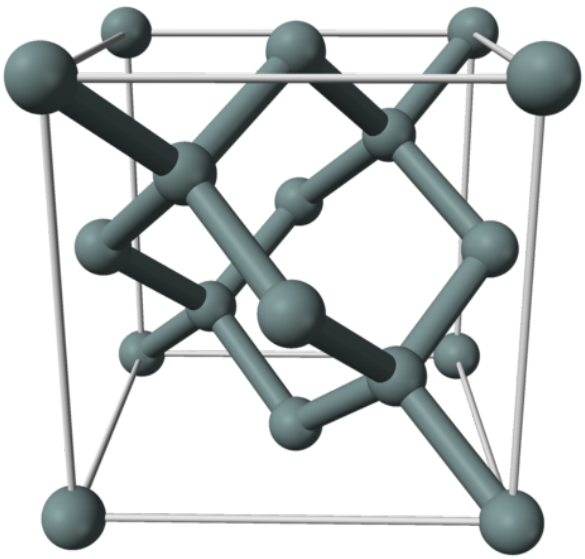

Figure 1: Silicon crystal unit cell. Image: Ben Mills, public domain, via Wikimedia Commons

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" orbital, 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. What we really mean when we say "share 4 electrons" is that the wave-functions of the outer orbitals interact, and we can no longer think of the orbitals as belonging to either of the silicon nuclei. All the neighbors atoms "share" electrons, and nowhere is there a vacant state, or a hole, in the valence band.

If such a crystal were to exist, where there were no holes in the valence band, and a net neutral charge, the crystal could not conduct any drift current. Electrons would move around continuously, swapping states, but there could be no net drift of charge carriers.

In an 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. Both bands can support drift and diffusion currents.

INTRINSIC CARRIER CONCENTRATION

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

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

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

[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

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.

Leave it to engineers to simplify equations beyond understanding. Equation (1) is complicated, and the density of states includes the effective mass of electrons and holes, which is a parameter that depends on the curvature of the band structure. To engineers, this is too complicated, and n_i has been simplified so it "works" in daily calculation.

Through engineering simplification, however, physics understanding is lost.

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] the intrinsic carrier concentration 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 Figure 2, we see the shape of BSIM and the full equation is almost the same, while the "doubling every 11 degrees" is just wrong.

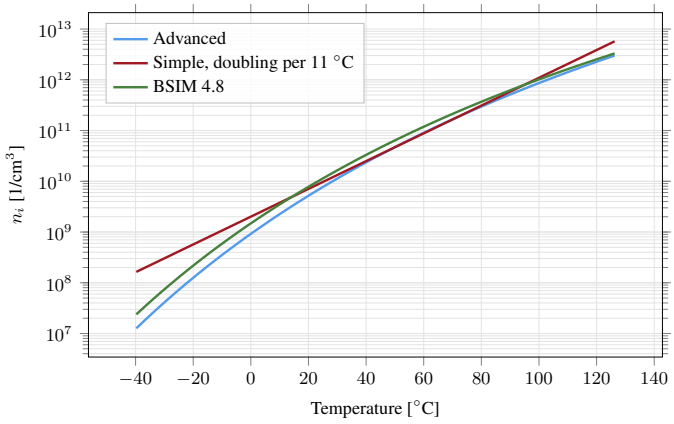

Figure 2: Intrinsic carrier concentration versus temperature

At room temperature the intrinsic carrier concentration 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.

From Figure 2 we can see that n_i changes greatly as a function of temperature, but the understanding "why" is not easy to get from "doubling every 11 degrees". To understand the temperature behavior of diodes, we must understand Eq (1).

So where does Eq (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

$$\psi = Ae^{i(k\mathbf{r}-\omega t)}$$

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

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

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 space dimension

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

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

We could rewrite the equation above as

$$\widehat{H}\psi(x,t) = i\hbar\frac{\partial}{\partial t}\psi(x,t) = \widehat{E}\psi(x,t)$$

where $\widehat{H}$ is sometimes called the *Hamiltonian* and is an operator, or something that act on the wave-function. In Feynman's Lectures on Physics Feynman called the Hamiltonian the *Energy Matrix* of a system. I like that better. The $\widehat{E}$ is the energy operator, something that operates on the wave-function to give the Energy.

We could re-arrange

$$[\widehat{H} - \widehat{E}]\psi(r,t) = 0$$

This is an equation with at least 5 unknowns, the space vector in three dimensions, time, and the energy matrix $\widehat{H}$.

The dimensions of the energy matrix depends on the system. The energy matrix further up is for one free electron. For an atom, the energy matrix will have more dimensions to describe the possible quantum states.

I consider all energy matricies as infinite dimensions, but most state transitions are so unlikely that they can be safely ignored.

I was watching Quantum computing in the 21st Century and David Jamison mentioned that the largest system we could today compute would be a system with about 30 electrons.

We know exactly how the equations of quantum mechanics appear to be, and they've proven extremely successful, we must make simplifications before we can predict how electrons behave in complicated systems like the silicon lattice with approximately 0.7 trillion electrons per cube micro meter. You can check the calculation

$$\left[\frac{1\ \mu\text{m}}{0.543\ \text{nm}}\right]^3 \times 8 \text{ atoms per unit cell} \times 14 \text{ electrons per atom}$$

Density of states

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

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

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.

The derivation starts by computing the density of states in the k-space, or momentum space,

$$N(dk) = \frac{2}{(2\pi)^p}dk$$

Where p is the number of dimensions (in our case 3).

The band structure $E(k)$ is used to convert to the density of states to a function of energy $N(E)$. The simplest band structure, and an approximation of the lowest conduction band is

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

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 when we compute the density of states is

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

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

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

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 chemical 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 observations on the Fermi-Dirac distribution. If the Energy of a state is at the Fermi level, then $f(E) = \frac{1}{2}$, or a 50 % probability of being occupied.

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 energy 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

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

For intrinsic silicon at thermal equilibrium, we could write

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

As we can see, Equation (3) 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 $n_i^2 = n_0 p_0$

How to think about electrons (and holes)

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 somewhere else. They are already at the silicon surface, but they are bound in covalent bonds (there are literally trillions of bound electrons in a typical transistor).

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.

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 density concentration difference will give rise to a diffusion current given by

$$J_{\text{diffusion}} = -qD_n \frac{\partial \rho}{\partial x} \tag{4}$$

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.

To make matters more complicated, an inversion layer of a MOSFET is not in three dimensions, but rather a two dimensional electron gas, as the density of states is confined close to the silicon surface. As such, we should not expect the mobility of bulk silicon to be the same as the mobility of a MOSFET transistor.

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 [3]. 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 probability 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, as 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.

Built-in voltage

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

The density of free electrons in the conduction band is

$$n = \int_{E_C}^{\infty} N(E) f(E) dE$$

, where $N(E)$ is the density of states, and $f(E)$ is a probability of a electron being in that state (Equation (2)).

We could write the density of electrons on the n-side as

$$n_n = e^{E_{F_n}/kT} \int_{E_C}^{\infty} N_n(E) e^{-E/kT} dE$$

since the Fermi level is independent of the energy state of the electrons (I think).

The density of electrons on the p-side could be written as

$$n_p = e^{E_{F_p}/kT} \int_{E_C}^{\infty} N_p(E) e^{-E/kT} dE$$

If we assume that the density of states, $N_n(E)$ and $N_p(E)$ are the same, and the temperature is the same, then

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

The difference in Fermi levels is the built-in voltage multiplied by the unit charge.

$$E_{F_n} - E_{F_p} = q\Phi$$

and by substituting for the minority carrier concentration on the p-side we get

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

or rearranged to

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

Current

The derivation of current is a bit involved, but let's try.

The hole concentration on the p-side and n-side could be written as

$$\frac{p_p}{p_n} = e^{-q\Phi_0/kT}$$

The negative sign is because the built in voltage is positive on the n-type side

Assume that $-x_{p0}$ is the start of the junction on the p-side, and x_{n0} is the start of the junction on the n-side.

Assume that we lift the p-side by a voltage qV

Then the hole concentration would change to

$$\frac{p(-x_{p0})}{p(x_{n0})} = e^{q(V-\Phi_0)/kT}$$

while on the n-side the hole concentration would be

$$\frac{p(x_{n0})}{p_n} = e^{qV/kT}$$

So the excess hole concentration on the n-side due to an increase of V would be

$$\Delta p_n = p(x_{n0}) - p_n = p_n\left(e^{qV/kT} - 1\right)$$

The diffusion current density, given by Equation (4) states

$$J(x_n) = -qD_p \frac{\partial \rho}{\partial x}$$

Thus we need to know the charge density as a function of x. I'm not sure why, but apparently it's

$$\partial\rho(x_n) = \Delta p_n e^{-x_n/L_p}$$

where L_p is a diffusion length. I think the equation above, the exponential decay as a function of length, is related to the probability of electron/hole recombination, and how the rate of recombination must be related to the exceess hole concentration, as such related to Exponential decay.

Anyhow, we can now compute the current density, and need only compute it for x_n = 0, so you can show it's

$$J(0) = q\frac{D_p}{L_p} p_n\left(e^{qV/kT} - 1\right)$$

which starts to look like the normal diode equation. The p_n is the minority concentration of holes on the n-side, which we've before estimated as $p_n = \frac{n_i^2}{N_D}$

We've only computed for holes, but there will be electron transport from the p-side to the n-side also.

We also need to multiply by the area of the diode to get current from current density. The full equation thus becomes

$$I = qAn_i^2 \left(\frac{1}{N_A} \frac{D_n}{L_n} + \frac{1}{N_D} \frac{D_p}{L_p} \right) \left[e^{qV/kT} - 1 \right]$$

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.

Which we usually write as

$$I_D = I_S(e^{\frac{V_D}{V_T}} - 1), \text{ where } V_T = kT/q$$

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

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

[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.

Where ℓ 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.

From the diode voltage relation we can calculate the derivative (ask your favorite chatbot to explain it to you), and you'll get

$$\frac{dV_D}{dT} = \frac{k}{q}(\ell - 3 \ln T - 3).$$

and we can see the slope is negative with a factor of $\ln(T)$, which turns out to be very close to linear for the temperature range we're interested in (-40C to 125C)

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.

Figure 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.

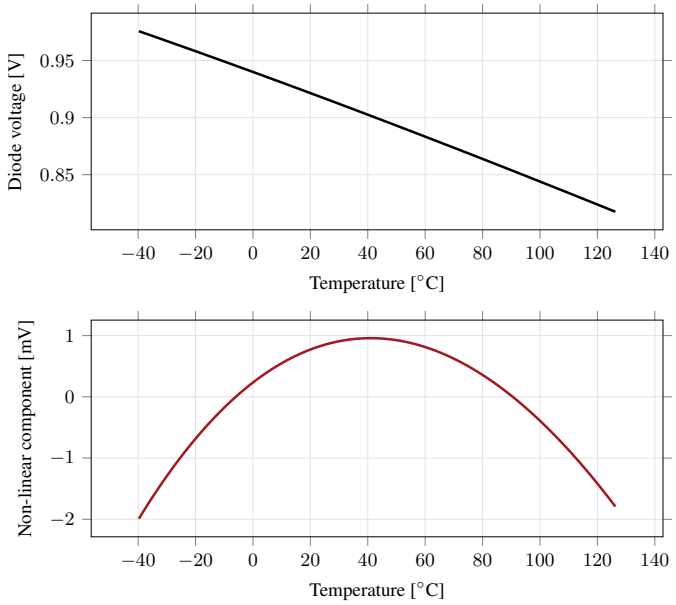

Figure 3: Diode forward voltage as a function of temperature

Current proportional to temperature

Assume we have a circuit like Figure 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 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.

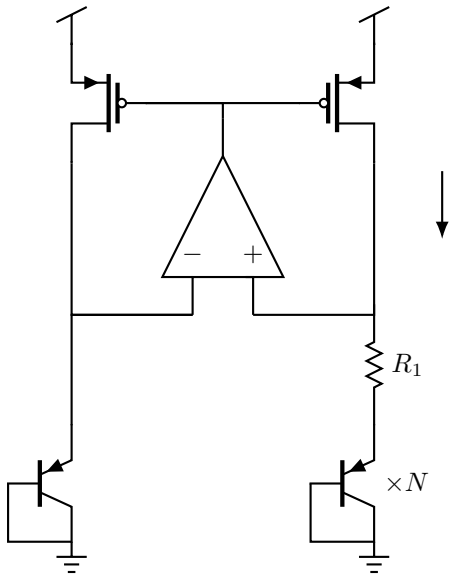

Figure 4: Circuit to generate a current proportional to kT

REVERSE LEAKAGE

So far we have only looked at the diode in forward bias, where the exponential term dominates and the current is set by the diffusion saturation I_S. In integrated circuits we equally often care about what happens when the diode is reverse biased, with $V_D < 0$. The exponential term is then negligible, and you might expect the current to be zero. It is not.

What happens physically is that the reverse bias lifts the Fermi level on the n-side relative to the p-side. The built-in field grows, the depletion region widens, and any minority carrier that wanders into that region is immediately swept across to the other side. The question of "how much current" is therefore really the question of "how many minority carriers are produced per second, and where".

In a real n+/p-substrate junction there are two places where minority carriers appear:

1. In the *quasi-neutral* p-substrate, a short diffusion length away from the junction. These thermally generated electrons diffuse to the depletion edge and are then collected.
2. *Inside* the depletion region itself, through generation at Shockley-Read-Hall trap centers in the silicon bandgap.

The first gives the Shockley diffusion saturation current I_S, the second gives the Sah-Noyce-Shockley generation current I_{gen}. They have different temperature dependencies, and which one dominates depends on temperature, doping, and how clean the silicon is.

Diffusion leakage

$$I_S = qAn_i^2 \left(\frac{1}{N_A} \sqrt{\frac{D_n}{\tau_n}} + \frac{1}{N_D} \sqrt{\frac{D_p}{\tau_p}} \right)$$

This is exactly the I_S from the forward-bias derivation, just re-interpreted. Under reverse bias the minority-carrier concentration at the edge of the depletion region is forced to *zero* (every carrier that arrives is swept across). The concentration gradient between the bulk equilibrium $n_p = n_i^2/N_A$ on the p-side and zero at the depletion edge drives a steady diffusion of electrons toward the junction, and similarly of holes from the n-side.

The factor that matters for temperature is n_i^2. From Equation (1), $n_i^2 \propto T^3 \exp(-E_g/kT)$. The T^3 comes from the densities of states $N_c, N_v \propto T^{3/2}$, which is a genuine quantum-mechanical effect: as temperature rises, more states become thermally accessible above E_c and below E_v. The $\exp(-E_g/kT)$ dominates the slope, but the T^3 prefactor contributes a non-trivial correction that is bigger than people usually admit. The log-derivative of n_i^2 is

$$\frac{d \ln n_i^2}{dT} = \frac{E_g}{kT^2} + \frac{3}{T}$$

At $T = 300\,K$ the $3/T$ term is about $7\,\%$ of the exponential term; at $T = 1000\,K$ it is about $25\,\%$. So the density-of-states prefactor visibly steepens the leakage curve at high temperature, and you cannot drop it if you care about anything beyond a back-of-envelope estimate.

The diffusion current doubles roughly every $4\text{-}5\,K$ near room temperature - steeper than the familiar "reverse current doubles every 10 K" rule of thumb, which belongs to the generation-limited regime [4].

[4]This rule of thumb applies to the diffusion-dominated regime. For the generation-dominated regime, which is where most small junctions actually live at room temperature, the doubling interval is closer to 8-10 K because $I_{gen} \propto n_i$ rather than n_i^2.

For an n+/p-well antenna diode, the $1/N_D$ term is negligible because $N_D \gg N_A$, and I_S is set almost entirely by electron injection from the p-well.

Both D and τ depend on temperature too, and that matters once the sweep is several hundred kelvin wide. Phonon-limited mobility goes as $\mu \propto T^{-2.4}$ above $\sim 300\,K$, so via the Einstein relation $D = \mu kT/q \propto T^{-1.4}$. The SRH lifetime goes as $\tau = 1/(\sigma v_{th} N_t)$ with $v_{th} \propto \sqrt{T}$, so $\tau \propto T^{-1/2}$ (taking σ and N_t constant). The combination $\sqrt{D/\tau}$ in I_S therefore drifts as $T^{-0.45}$, so I_S at $1000\,K$ is about $0.6\times$ what a "constant D, τ" model would predict. That is small compared to the many decades of swing in n_i^2, but it is the dominant reason the diffusion curve in Figure 5 starts to flatten at the top end. The script `ex/antenna_diode_leakage.py` includes both T-dependencies, and the interactive version lets you switch them off to see how little difference they make next to the bandgap.

Generation in the depletion region

$$I_{gen} = \frac{qAn_iW}{\tau_g}$$

$$W = \sqrt{\frac{2\varepsilon_{si}(\Phi_0 + V_R)}{q} \cdot \frac{N_A + N_D}{N_A N_D}}$$

Real silicon is not a perfect crystal. Process damage, residual metallic impurities and dangling bonds at interfaces create trap levels somewhere in the bandgap. A trap near mid-gap can capture an electron from the valence band (creating a hole) and then emit it to the conduction band (creating an electron). Inside a reverse-biased depletion region, both carriers are immediately swept apart by the field, the trap empties, and the cycle repeats. The net effect is a steady generation of electron-hole pairs that shows up as reverse current.

The Sah-Noyce-Shockley analysis gives the result above, where W is the depletion width and τ_g is an effective generation lifetime (typically a few τ_n). For a one-sided n+/p junction with $N_A \ll N_D$ the depletion width simplifies to

$$W \approx \sqrt{\frac{2\varepsilon_{si}(\Phi_0 + V_R)}{qN_A}}$$

which sits almost entirely on the lightly doped substrate side.

The temperature scaling is now n_i, not n_i^2. The exponential in n_i has $E_g/(2kT)$ and the density-of-states prefactor is $T^{3/2}$ rather than T^3, so the slope is roughly half that of I_S. I_{gen} doubles every 8-10 K near room temperature. At low and moderate temperatures, where n_i is small, this slower-scaling term still dominates because it has the smaller exponent. As temperature rises and n_i grows by many decades, the steeper $I_S \propto n_i^2$ eventually overtakes it.

The $1/\tau_g$ in I_{gen} also drifts with temperature: with $\tau \propto T^{-1/2}$ the generation term picks up an extra $\sqrt{T}$ factor on top of the n_iW scaling, which is why at the top of the sweep I_{gen} does not flatten out as much as one might expect from n_i alone.

The reverse bias V_R enters through W, so I_{gen} has a weak $\sqrt{V_R}$ dependence. Diffusion current I_S is essentially independent of bias once the junction is reverse-biased by more than a few kT/q. This is the standard way to separate the two experimentally: I_S saturates, I_{gen} grows with $\sqrt{V_R}$.

Plasma charging currents

$$J_{plasma} \sim \text{1-10}\,\mathrm{mA/cm^2}$$

$$I_{ant} = J_{net} \cdot A_{antenna}$$

How much current does the diode actually have to sink? A reactive-ion plasma delivers ion and electron fluxes to the wafer surface, each typically in the $\text{1-10}\,\mathrm{mA/cm^2}$ range [4]. If the two fluxes were perfectly balanced everywhere, no floating conductor would charge up and there would be nothing to drain. They never are, and two mechanisms produce a net DC current that the antenna diode must shunt:

- **Macroscopic non-uniformity.** Ion and electron fluxes vary across the wafer because of source geometry, magnetic-field shaping and edge effects. The local mismatch is typically a few percent of J_{plasma}, giving a net DC charging current density on the order of $10\,\mu\mathrm{A/cm^2}$ to $0.1\,\mathrm{mA/cm^2}$ collected by any conductor connected to that area [4].
- **Electron shading.** In a high-aspect-ratio opening through photoresist, vertical ions reach the bottom of the feature while near-isotropic electrons are blocked by the sidewalls. The bottom charges positively regardless of plasma uniformity. Hashimoto identified this as the dominant damage mechanism for deep-submicron processes [5], and it is what makes high-density-plasma etch and HDP-CVD particularly aggressive.

A metal trace that the routing tool calls the *antenna* collects the net plasma current density over its full area $A_{antenna}$ and funnels it onto the much smaller gate. Foundry rules cap the *antenna ratio* $A_{antenna}/A_{gate}$ at typically a few hundred to a few thousand, but even within those limits a net plasma current of $\sim 10\,\mu\mathrm{A/cm^2}$ over a $10^4\,\mu m^2$ antenna gives $\sim 1\,\mu A$ injected onto a sub-μm^2 gate. With no discharge path, that current would charge the gate node and force Fowler-Nordheim tunneling through the oxide (fields above $\sim 10\,\mathrm{MV/cm}$). Once tunneling starts, defects accumulate and the breakdown lifetime collapses [6].

The antenna diode bypasses this failure by giving the antenna a reverse-biased path to substrate that conducts at far lower voltage than the oxide. As long as the diode reverse leakage at the relevant process temperature exceeds the antenna-collected charging current, the gate voltage stays clamped well below the FN threshold and the oxide is safe. This is what makes Figure 5 a sizing tool: at the coldest plasma step the leakage

per μm^2 is only $\sim 1\,\mathrm{fA}$, so a small antenna ratio or a generous diode area is required to maintain $I_{leak} \gtrsim I_{ant}$.

Antenna ndiode, 200 K to 1000 K

A useful integrated-circuit example is the *antenna diode*. During plasma etch and ion-implant steps in fabrication, long metal traces collect charge. If the trace is connected to a transistor gate, the gate dielectric can break down before the chip ever sees a power supply. Foundry design rules specify maximum *antenna ratios* (metal area to gate area) per layer; if a net exceeds the ratio, the routing tool inserts an *antenna diode* on the net to bleed the plasma charge to substrate. In normal operation that diode sits reverse-biased and must contribute negligible static current.

The diode itself is simply an n+ source/drain implant placed in the NMOS p-well. The relevant doping is:

- p-substrate (handle wafer): $N_{sub} \approx 10^{15}\,cm^{-3}$
- p-well (retrograde, set by short-channel control): $N_A \approx 10^{17} - 10^{18}\,cm^{-3}$
- n+ source/drain: $N_D \approx 10^{20}\,cm^{-3}$

It is the p-well doping that matters at the junction, because the n+ sits inside the well, not directly on substrate. The effective N_A seen by the antenna ndiode is therefore one to two orders of magnitude higher than the wafer doping, which makes the depletion narrower and the reverse leakage *smaller* than a naive "n+/p-substrate" estimate would suggest.

Figure 5 plots the reverse leakage of such a junction: $N_A = 10^{17}\,cm^{-3}$ (p-well), $N_D = 10^{20}\,cm^{-3}$ (n+ S/D), $V_R = 1\,V$, swept from 200 K to 1000 K. The current is plotted as a *leakage density* in $\mathrm{A}/\mu m^2$, so multiplying by the actual junction area in μm^2 gives the absolute current - a $0.2 \times 0.2\,\mu m^2$ antenna ndiode is the curve times 0.04, a $1 \times 1\,\mu m^2$ diode reads off directly. The script is `ex/antenna_diode_leakage.py` and reuses the $n_i(T)$ derivation from `ex/vd.py`. Both exist as interactive pages: antenna diode leakage, where the doping, the reverse bias and the junction area are sliders, and diode vs temperature for the $n_i(T)$ model on its own.

The two shaded bands mark where plasma-induced damage actually happens during fabrication, which is the *only* time an antenna diode does any work. The cooler band is plasma etch (reactive ion etch, with the wafer chuck cooled to roughly 25-100 °C, so $\sim$ 300-375 K); the hotter band is plasma deposition steps such as PECVD, HDP-CVD and sputter ($\sim$ 200-400 °C, so $\sim$ 470-675 K, with 400 °C a hard upper limit set by BEOL metal reliability). Across this span the leakage available to bleed plasma charge ranges from roughly $1\,\mathrm{fA}/\mu m^2$ at room-temperature etch to $\sim 6\,\mathrm{nA}/\mu m^2$ at 400 °C deposition - nearly seven decades of variation depending only on which process step you are in.

Three things to take away from the figure:

- Generation current dominates from cryogenic temperatures up to roughly 575 K, exactly where the steeper n_i^2 slope of the diffusion term catches up. Above that the diode is in the "diffusion-limited" regime. The crossover happens to fall right in the middle of the plasma-deposition band, so antenna diodes during deposition steps see contributions from both mechanisms.
- The leakage per μm^2 swings nearly seven decades over the wafer-fab thermal range. An antenna diode that easily sinks the plasma charge during a 400 °C HDP-CVD step may be orders of magnitude too small during a room-temperature metal etch. The worst case for sizing is therefore the *coldest* plasma step.
- The slope on a log axis is set by the $\exp(-E_g/(2kT))$ in n_i plus the $T^{3/2}$ density-of-states prefactor. Every junction in a given silicon process scales with temperature the same way; only the absolute level changes with area, doping and lifetime.

It is also worth noting that at the upper end of the sweep n_i approaches and eventually exceeds N_A. The diode then loses junction behaviour and the silicon behaves as an intrinsic resistor. This is why high-temperature electronics typically uses wide-bandgap materials such as silicon carbide.

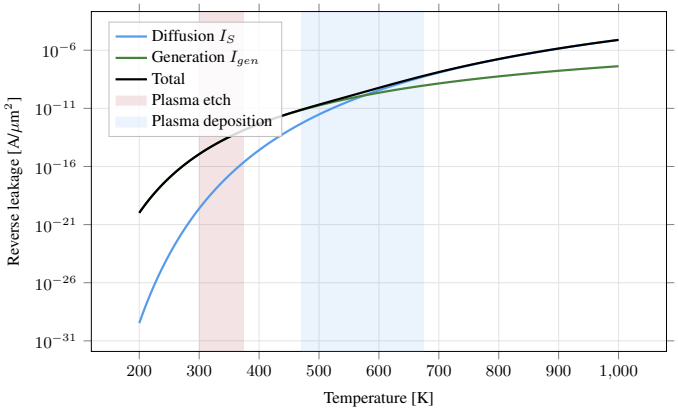

Figure 5: Reverse leakage density (A/um^2) of an n+/p-well antenna ndiode, from 200 K to 1000 K, $V_R = 1$ V. Multiply by your actual diode area in um^2. Shaded bands are the wafer temperature during plasma etch (300-375 K) and plasma deposition (470-675 K), the steps at which plasma-induced gate damage occurs.

Wires below 1 mm

$$I_{ant} = J_{net} \cdot W_{wire} \cdot L_{wire}$$

The antenna current scales linearly with wire length, so for wire lengths below 1 mm - which covers practically everything that stays inside a single block or hierarchical cell - the sizing problem is concrete. Take a minimum-width interconnect, $W_{wire} \approx 0.1\,\mu m$, and the worst-case cold-etch net plasma current density, $J_{net} \approx 10\,\mu\mathrm{A/cm}^2$:

L_{wire}	$A_{antenna}$	I_{ant} at room-T etch
$1\,\mu m$	$0.1\,\mu m^2$	~ 1 fA
$10\,\mu m$	$1\,\mu m^2$	~ 10 fA
$100\,\mu m$	$10\,\mu m^2$	~ 100 fA
$1\,mm$	$100\,\mu m^2$	~ 1 pA

Now compare with Figure 5. At room temperature the antenna ndiode leaks only $\sim 1\,\mathrm{fA}/\mu m^2$. A $1\,\mathrm{mm}$ wire collecting $1\,\mathrm{pA}$ would therefore need a diode active area of $\sim 900\,\mu m^2$ (a $30\times 30\,\mu m$ device) to keep the gate clamped during cold etch, which is clearly impractical. The same $1\,\mathrm{mm}$ wire is comfortably protected by a $0.15\,\mu m^2$ diode at a $200\,^\circ C$ deposition step ($\sim 3\,\mathrm{pA}/\mu m^2$) and by a sub-$0.01\,\mu m^2$ diode at $400\,^\circ C$.

Two consequences:

- For wires shorter than a few μm, the collected charging current is below the room-temperature diode leakage even for the smallest practical antenna ndiode. No protection is needed - which is why foundry antenna rules always have a length-threshold exemption.
- For long wires the cold-etch step dominates the sizing, and a single big diode is rarely the right answer. The routing tool instead *jumps* the long net up to a higher metal layer (patterned only after the lower stack already protects the gate), or distributes many small antenna ndiodes along the wire so each one only has to drain its local section of the antenna current.

Wider metals (upper-stack power and clock) scale antenna area linearly with width, so a $1\,\mu m$-wide, $1\,\mathrm{mm}$-long top-metal trace collects $10\,\mathrm{pA}$ rather than 1, and the same conclusions apply with one decade less margin.

Equations aren't real

Nature does not care about equations. It just is.

We know, at the fundamental level, nature appears to obey the mathematics on quantum mechanics, however, due to the complexity of nature, it's not possible today (which is not the same as impossible), to compute exactly how the current in a diode works. We can get close, by measuring a diode we know well, and hope that the next time we make the same diode, the behavior will be the same.

As such, I want to warn you about the "lies" or "simplifications" we tell you. Take the diode equation above, some parts, like the intrinsic carrier concentration n_i has roots directly from quantum mechanics, with few simplifications, which means it's likely solid truth, at least for a single unit cell.

But there is no reason nature should make all unit cells the same, and infact, we know they are not the same, we put in dopants. As we scale down to a few nano-meter transistors the simplification that "all unit cells of silicon are the same, and extend to infinity" is no longer true, and must be taken into account in how we describe reality.

Other parts, like the exact value of the bandgap E_g, the diffusion constant D_p or diffusion length L_p are macroscopic phenomena, we can't expect them to be 100 % true. The values would be based on measurement, but not always exact, and maybe, if you rotate your diode 90 degrees on the integrated circuit, the values could be different.

You should realize that the consequence of our imperfection is that the equations in electronics should always be taken with a grain of salt.

Nature does not care about your equations. Nature will easily have the superposition of trillions of electrons, and they don't have to agree with your equations.

But most of the time, the behavior is similar.

Summary

The one-page version of this chapter:

- Silicon's bandgap makes it a semiconductor: few free carriers at room temperature, exponentially more when heated
- Doping moves the Fermi level - towards the conduction band with donors, towards the valence band with acceptors
- Join n to p and diffusion fights drift until the depletion field balances: the built-in voltage
- Forward bias lowers the barrier exponentially: $I = I_S(e^{V/V_T} - 1)$
- The diode voltage falls roughly 2 mV/K and extrapolates to the bandgap at 0 K - the seed of every bandgap reference
- Reverse leakage grows exponentially with temperature: an antenna diode sized at etch temperature is a different diode at deposition

Would you like to know more?

[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/

[3] T. Masuhara and R. Muller, "Complementary DMOS process for LSI," *IEEE Journal of Solid-State Circuits*, vol. 11, no. 4, pp. 453–458, 1976, doi: 10.1109/JSSC.1976.1050758.

[4] K. P. Cheung, *Plasma charging damage*. Springer-Verlag London, 2001 [Online]. Available: https://link.springer.com/book/10.1007/978-1-4471-0247-2

[5] K. Hashimoto, "Charge damage caused by electron shading effect," *Japanese Journal of Applied Physics*, vol. 33, pp. 6013–6018, 1994, doi: 10.1143/JJAP.33.6013. [Online]. Available: https://iopscience.iop.org/article/10.1143/JJAP.33.6013

[6] S. Krishnan, A. Amerasekera, S. Rangan, and S. Aur, "Antenna device reliability for ULSI processing," in *International electron devices meeting (IEDM) technical digest*, 1998, pp. 247–250, doi: 10.1109/IEDM.1998.746430.

Carsten Wulff received the M.Sc. and Ph.D. degrees in electrical engineering from the Department of Electronics and Telecommunication, Norwegian University of Science and Technology (NTNU), in 2002 and 2008, respectively. During his Ph.D. work at NTNU, he worked on open-loop sigma-delta modulators and analog-to-digital converters in nanoscale

CMOS technologies. In 2006-2007, he was a Visiting Researcher with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada. Since 2008 he's been with Nordic Semiconductor in various roles, from analog designer, to Wireless Group Manager, to currently Principle IC Scientist. From 2014-2017 he did a part time Post.Doc focusing on compiled, ultra low power, SAR ADCs in nanoscale technologies. He's also an Adjunct Associate Professor at NTNU. His present research interests includes analog and mixed-signal CMOS design, design of high-efficiency analog-to-digital converters and low-power wireless transceivers. He is the developer of Custom IC Compiler, a general purpose integrated circuit compiler, and makes the occational video on analog integrated circuits at https://www.youtube.com/@analogicus. For full CV see https://analogicus.com/markdown-cv/.