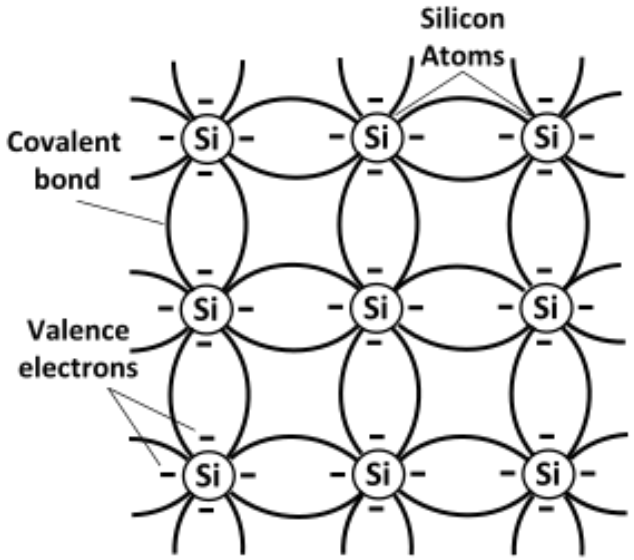
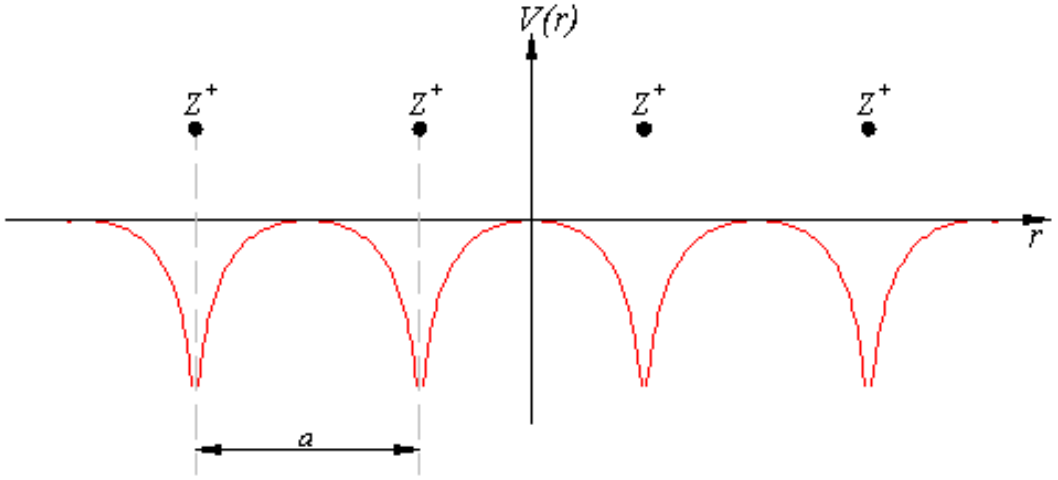
Semiconductors

Individual toms have discrete energy levels but when they bond into a solid the energy levels are dramatically changed. To begin, recall that a crystal consists of a periodic "lattice" . A simple example is shown below. This orderly arrangement makes the potential for any given electron look like a nearly infinite periodic array of potential wells spaced some distance *a* apart where *a* is typically a few Angstroms. At the center of each well is the nucleus with charge + Ze where Z is the number of protons and e is the elementary unit of charge. (A good introduction can be found in *Principles of Semiconductors Devices* (B. Van Zeghbroeck, http://ecee.colorado.edu)



https://en.wikipedia.org/wiki/File:Covalent_bonding_in_silicon.svg

What are the quantum states of this of this periodic potential? If the atoms were very far apart you would expect the energies and wavefunctions to be those of a single atom: 1s, 2s, 2p… orbitals with their corresponding energies.

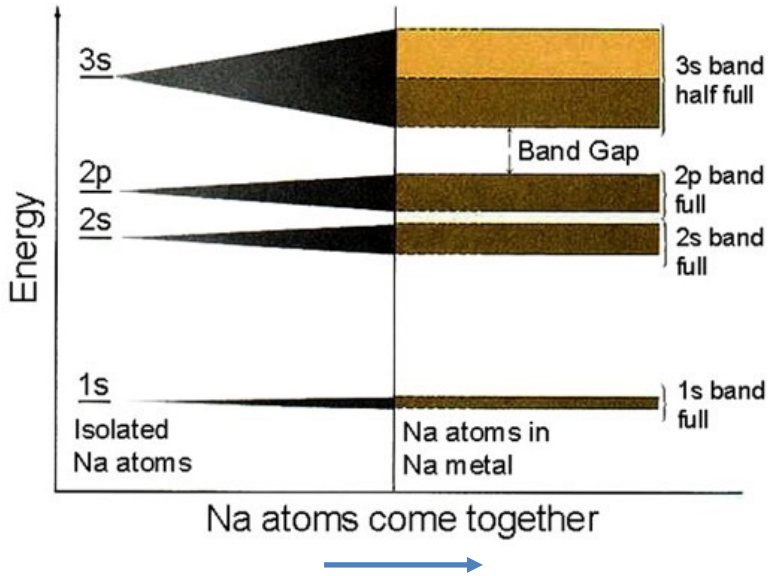


If there are N isolated atoms, there are N identical 1s orbitals, N identical 2s orbitals and so on. By orbital I mean the spatially-dependent part of the wavefunction. For example, the 1s orbital in hydrogen is given by,

$$\psi_{1s} = \frac{1}{{a_0}^{3/2}\sqrt{\pi}}\, e^{-r/a_0} \qquad a_0 = \ Bohr\ radius\ \approx 0.0529\ nm$$

Each orbital can accommodate 2 electrons (spin up and spin down) so there are 2N distinct quantum *states* with energy E_{1s} .

Now imagine bringing the atoms closer together until they are separated by the distance they would have in an actual crystal, a = 0.2 - 0.5 nm. As they come closer together the individual potential wells overlap forming a new potential energy landscape. The N original 1s orbitals develop into N distinct new orbitals, each with a slightly different energy, i.e., a *band* of energies. Something similar happens for the 2s, 2p, etc. other atomic orbitals, each atomic energy level developing into a band of energies. If there are N atoms in the crystal then each band contains N distinct energy levels. And as usual, 2 electrons (spin up and spin down) can occupy each orbital, so the energy band can accommodate 2N electrons. The crystal will have many bands, some of which may overlap. The figure below shows the process for sodium (Na) which is a simple metal.



universe-review.ca

Now think of building the solid just the way you'd build an atom, by filling the available quantum states with electrons. Sodium has 11 electrons. For a solid with N sodium atoms, we must put 11 N electrons into distinct quantum states. The 1s band can take 2N electrons as can the 2s band. Each of the three 2p bands can also take 2N electrons. The remaining N electrons go into the 3s band which will be only half-full as shown. As we will see, the fact that the topmost energy level is within a partially-filled band is the reason sodium is a metal.

Working out this "band structure" for a real material is a messy problem in solid state physics. Suffice to say that energy bands in solids are typically 0.1 eV to to a few eV wide and the band gaps are of similar size. Each band has N distinct energy levels but the spacing between these individual levels within a band is *extremely* small. Consider a band 1 eV wide in a solid with N = 10^{22} atoms. It has N orbitals and therefore N = 10^{22} energy levels. The spacing between the levels within the band is therefore about 10^{-22} eV, which is *very* small. Therefore, the energy levels in a band are essentially continuous.

Incidentally, what are the wavefunctions now that the electrons are in this crystal? It was proven by Felix Bloch back in the 1930's that electronics travelling in an infinite periodic potential have orbitals of the form,

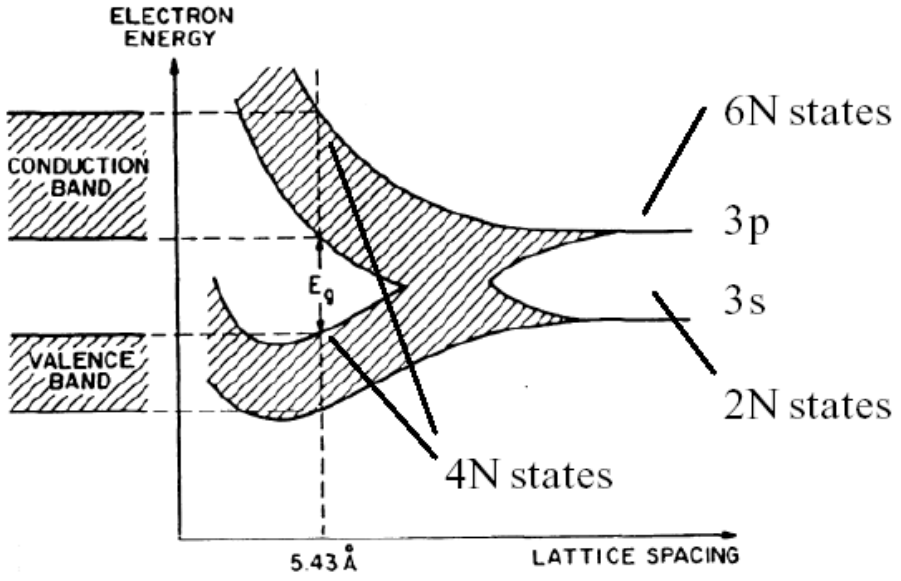
$$\psi_{\vec{k}}(\vec{r}) = e^{-i\,\vec{k}\cdot\vec{r}}\, u_{\vec{k}}(\vec{r})$$

Here $u_{\vec{k}}(\vec{r})$ is a periodic function of $\vec{r}$ with the same periodicity as the potential. $\vec{k}$ is a quantum number somewhat like that for a free electron but with some constraints.

Energy bands in silicon

The electronic world revolves around silicon. Atomic silicon has 14 electrons so a crystal with N silicon atoms must find distinct quantum states for 14N electrons. The first 10N fill up lower energy bands (1s, 2s, 2p) leaving 4N outer shell or *valence* electrons. (Recall that atomic silicon has a $3s^2 3p^2$ configuration which means 4 valence electrons.) Band structure calculations reveal that as the interatomic spacing decreases, the 3s and 3p bands first hybridize into one big band. Then, as the spacing decreases further down to the value it has in a real crystal (labelled a_0 in the figure) this single band splits into two bands, each of which can accommodate 4N electrons. These bands are separated by a gap $E_g = 1.1$ eV.

The diagram below shows the process just described with $a_0 = 0.545$ nm, the spacing in a real silicon crystalline lattice. The lower energy band is called the valence band and the upper band is called the conduction band. Each can accommodate 4N electrons. At temperature $T = 0$ the solid would assume its lowest energy configuration in which all of the lower energy bands (not shown) would be full and the remaining 4N electrons would fill up the valence band, leaving the conduction band empty.



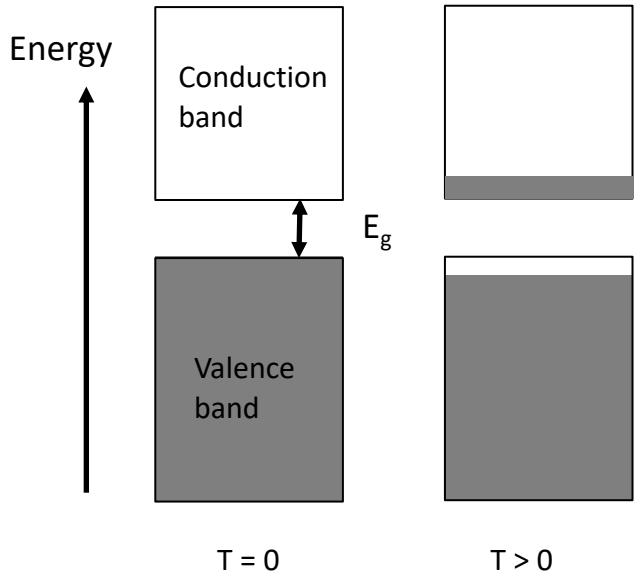
As we'll see, bands which are completely full cannot contribute to the conduction of electricity. For electronics they are inert and therefore irrelevant. For conduction to occur *some electrons must exist in a partially-filled band*. Sodium has its 3s band partially filled so it conducts easily – it's metal.

Since silicon at T = 0 has a filled valence band and an empty conduction band, it won't conduct electricity and would be useless for devices. However, two things can change this. First, we live closer to T = 300 K so there is a finite probability for an electron in the valence band to jump into the conduction band where it can roam around. Even more importantly, by adding suitable impurities, it's possible to selectively add or subtract charge carriers from bands and dramatically increase the conduction of current.

https://www.albany.edu/~soktyabr/NNSE618/NNSE618-L4-so-bands-kp.pdf

Electrons and Holes

Focus now on just the valence and conduction bands of silicon. All the lower energy bands will be full. At T = 0 all the states in the valence band are full (indicated by shaded regions) and those in the conduction band are unoccupied.

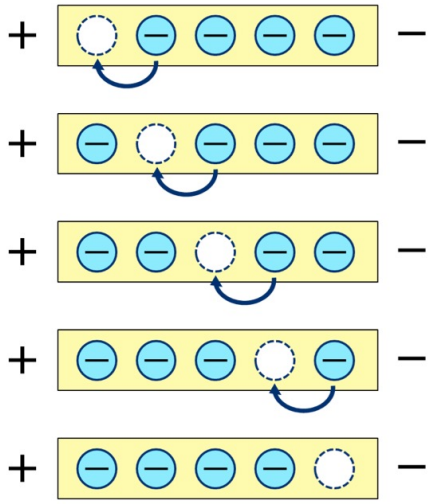


As the temperature is raised there is a non-zero probability for an electron in the valence band to jump across to the conduction band. That gives rise to a non-zero density *n* of electrons in the conduction band for T > 0. It also leaves behind an equal number of empty states in the valence band. We'll call these *holes*. The density of holes will be called *p* and in this case $n = p = n_i$. A result from statistical mechanics known as the *law of mass action* says that the product *np* depends only on temperature and material parameters,

$$n\,p = n_i^2 = 4\left(\frac{2\pi k_B T}{h^2}\right)^3 \left(m_{elec} m_{hole}\right)^{\frac{3}{2}} e^{-E_g/k_B T}$$

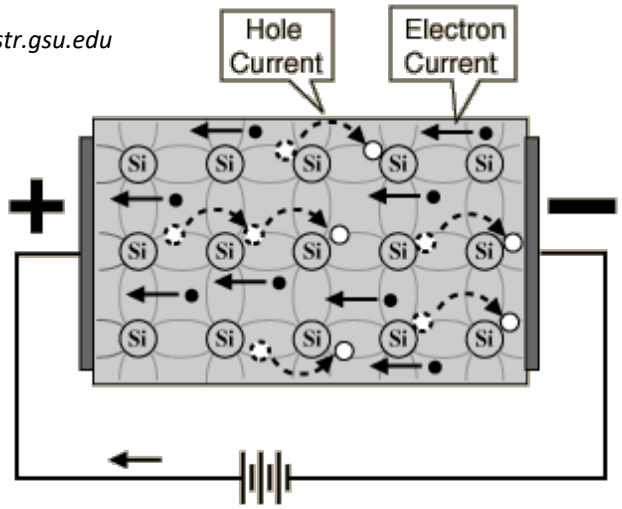
n_i is called the *intrinsic carrier density*. For Si at room temperature, E_g = 1.1 eV so $n_i \approx 1 \times 10^{10}/\text{cm}^3$ at T = 300 K. It's small but not negligible. By comparison, the density of electrons in copper is about $10^{22}/\text{cm}^3$.

m_{elec} and m_{hole} refer to the *effective mass* of the respective charge carriers. For example, an electron in the conduction band acts somewhat like an electron in a vacuum but its interaction with the periodic potential leads to an effective mass that is somewhat larger than the free electron mass. What about the holes? The picture on the right shows what happens with an empty state in the valence band. Imagine we apply a small electric field by making the left side more positive than the right side. Then the closest electron jumps into the hole and moves left. This continues down the line with the net result being a net positive charge moving left to right. That's the hole. It acts like a positively-charged particle even though only electrons are really moving. We can describe the flow of charge in the valence band by the motion of these positively charged holes. Due to the periodic potential, holes will have their own effective mass m_{hole} . Both kinds of charge carriers contribute to the current – electrons in the conduction band and holes moving in the valence band. Both carriers contribute to a current in the same direction.



The picture below illustrates the conduction of electricity in a piece of intrinsic silicon. At any temperature T > 0, for every electron in the valence band there is a hole left behind in the conduction band ($n = p = n_i$). Under the influence of an electric field, electrons move to the left toward the (+) potential. But an electron moving left constitutes a *positive current to the right*. Holes in the valence band act as positive charges and move toward the (-) terminal, also contributing a clockwise current.

hyperphysics.phy-astr.gsu.edu



The conductivity of a material is proportional to the density of charge carriers. At room temperature diamond has very few, silicon has more and metals have far more carriers. Therefore the resistivity (the inverse of the conductivity) at room temperature can vary over 20 orders of magnitude:

Resistivities at T = 293 K

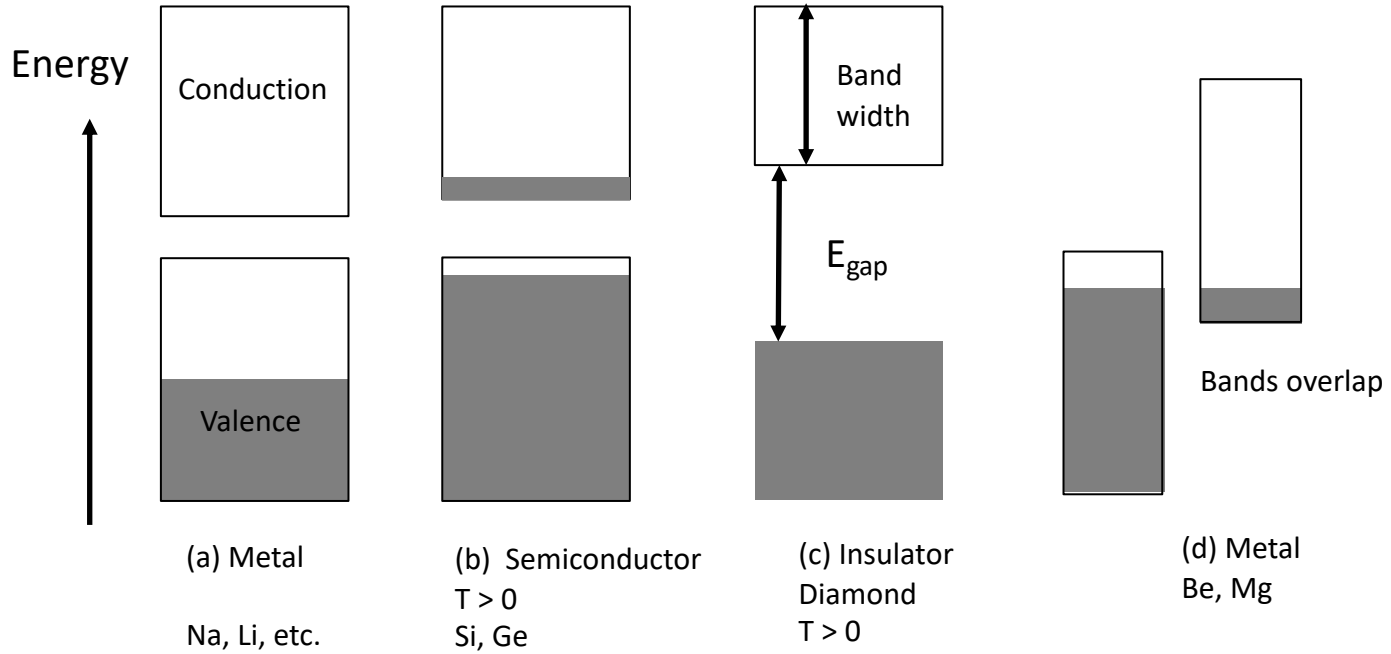
Copper	1.68×10^{-8}	Ω-m
Silicon	6.4×10^{2}	Ω-m
Diamond	10^{12}	Ω-m

A filled energy band conducts no current

We've just seen that the valence and conduction bands in silicon both contribute to an electrical current. But there are lower energy bands. Should we worry about them? To conduct electricity we need a current, i.e., more electrons going one way than the other. The states in an energy band each have a quantum number $\vec{k}$ for which which $\hbar\vec{k}$ is somewhat like momentum. For each state with quantum number $\vec{k}$ there is a corresponding state with $-\vec{k}$. Therefore if all the $\vec{k}$ states in a band are occupied with electrons there is net momentum and *no net current*. Applying an electric field cannot change the occupation of $\vec{k}$ states because there are no empty ones. Therefore the filled band is electrically inert. A filled energy band conducts no current. As we've just seen, things are different with a partially filled band. Now there are empty states available and the field can upset the balance between $+\vec{k}$ and $-\vec{k}$ electrons, leading to a net electrical current.

Metals, insulator and semiconductors

Whether or not a material conducts easily (metal), somewhat (semiconductor) or essentially not at all (insulator) depends on the number of electrons sitting in partially filled bands. This is shown below for different kinds of solids. Since the lower energy bands are already full and therefore irrelevant, we show just the top 2 bands. The darkened regions indicate electron states that are filled.
The band widths and band gaps depend upon the specific material.



(a) The left figure is a metal. There is a partially filled band even at T = 0 so it always conducts electricity.

(b) The next figure is a semiconductor at T > 0 so it conducts. However, the number of charge carriers is far smaller than in a metal so we call it a semiconductor.

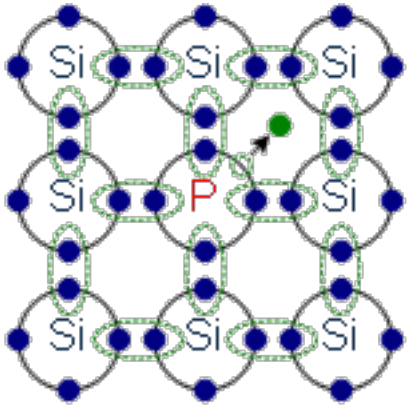
(c) This is an insulator. The energy gap is much larger than in a semiconductor (E_{gap} = 5.4 eV for diamond versus 1.1 eV for silicon) so even at room temperature there are very few charge carriers and essentially no conductivity.

(d) In some materials (e.g., Be and Mg) some of the bands overlap. The lower band could be completely filled with electrons, but the total energy is lowered by partially filling states in the upper band. Now both bands are partially-filled and the material is a metal.

Adding dopants – N-type silicon

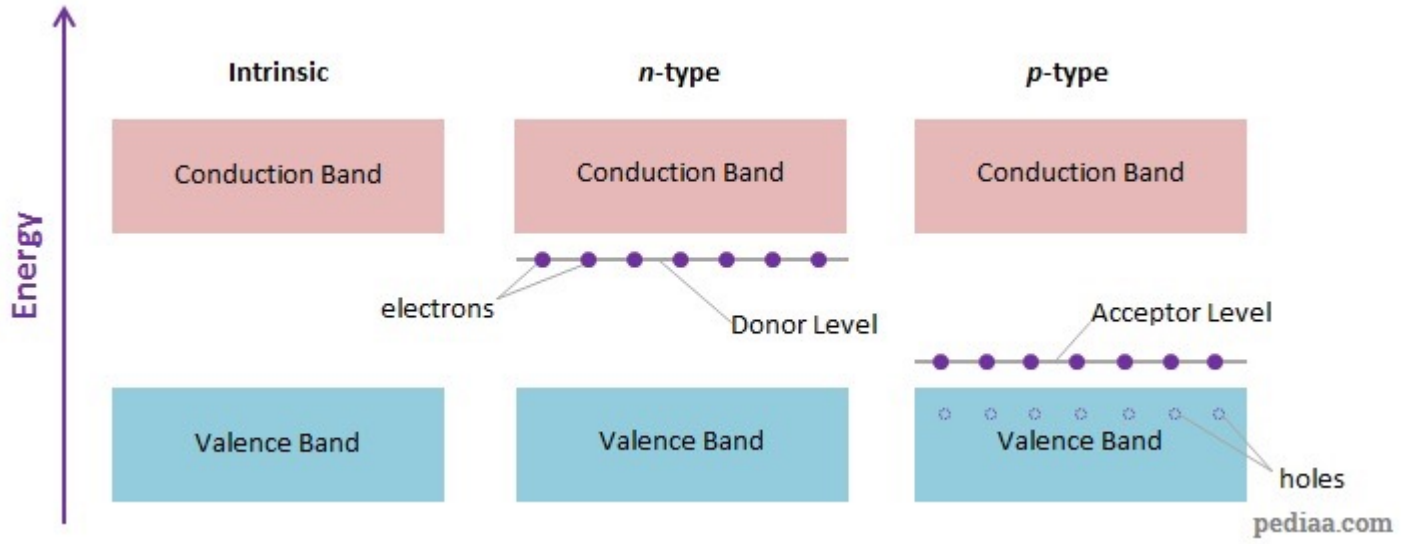
With just pure silicon, semiconductor physics would not be terribly interesting. Dramatic things happen when impurities ("dopants") are added. Recall that silicon has 4 electrons in its outer shell. Now consider replacing a silicon atom by phosphorous which has 5 outer shell electrons ($3s^2 3p^3$). The chemical bonding diagram shown below is useful. All the covalent bonds between atoms are filled but there is now an extra electron that can wander around the lattice.

N-doping : The phosphorous atom *donates* an electron that goes into the conduction band.



At T = 0, this electron orbits around the phosphorous atom it has left behind and which, as a result, has a net charge of +e. It's like a hydrogen atom but inside a solid so the potential energy is much weaker due to the dielectric constant of the silicon, which is about 11. The result is a hydrogen-like bound state, but with a binding energy of about 0.05 eV. Each phosphorous atom, called a *donor*, produces one such bound state, called a *donor level*. The donor levels all sit about 0.05 eV below the conduction band, as shown below.

Since the binding energy in a donor level is only 0.05 eV, electrons in these states are easily excited into the conduction band at room temperature, where they become mobile. Each ionized donor atom contributes one mobile electron to the conduction band. Therefore if the density of donor atoms is N_D, the density of electrons in the conduction band is essentially $n = N_D$. The result is called an **N-type semiconductor**.

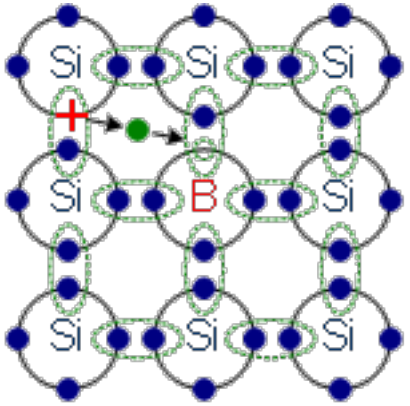


P-type silicon

Now consider adding an atom with 3 outer shell electrons, such as boron ($2s^2 2p^1$). The boron atom has one *less* electron than silicon. In order to fill the covalent bonds around the boron atom an electron is pulled from somewhere else, leaving behind a *hole.* The hole orbital energy sits a small amount above the top of the valence band. The state is called an acceptor level and boron is called an *acceptor* dopant.

At T = 0 the acceptor states would be empty – all electrons would be in the valence band. But at room temperature, electrons can easily jump from the valence band up to an acceptor level, leaving behind a hole in the valence band. These holes now can carry a current in the valence band. At room temperature most acceptor atoms have grabbed an electron from the valence band, leaving behind a hole. Therefore $p = N_A$. This arrangement is called a **P-type semiconductor**.

P-doping : The boron atom grabs an electron and leaves an unfilled bond between two of the silicon atoms. This absence of an electron acts like a mobile, positively charged particle – a hole.



For intrinsic silicon we're stuck with the intrinsic density of electrons and holes, about $10^{10}/cm^3$ at room temperature. With doping, we can control the n and p densities and therefore the conductivity of the silicon. Doping levels typically range from $10^{13}/cm^3$ – $10^{18}/cm^3$. Since the density of Si atoms is close to $10^{22}/cm^3$ the dopants are still very dilute. (They are diffused into the silicon at high temperatures for controlled periods of time.) Now recall the law of mass action:

$$n\,p = n_i^2(T)$$

Suppose the semiconductor is p-type with an acceptor density of N_A. Then at room temperature then $p = N_A$ and the electron density adjusts itself to keep the product constant:

$$n = n_i^2(T)/N_A$$

If $N_A = 10^{16}$ and $n_i = 10^{10}$ then $n = 10^4/cm^3$. The ratio $p/n = 10^{12}$ so there are *far, far* more holes than electrons. Through the law of mass action, adding dopants to a semiconductor allows us to change the density of carriers by many orders of magnitude, something that cannot be done with metals.

Drift and Diffusion

In semiconductors we need to consider two distinct ways in which electricity is conducted. The first contribution is familiar. The current density J is proportional to the electric field *E* :

$$J_{drift} = nqv = nq\mu E = \sigma E$$

In semiconductor physics this contribution is called the *drift* current. q is the charge, n is the density of charge carriers, v is their average velocity, σ is the conductivity and μ is called the mobility. This relationship gives rise to Ohm's law. Electrons in the conduction band and holes in the valence band will each have their own separate drift current.

Useful semiconductor devices involve sandwiches of p and n type materials with abrupt changes in the density of electrons and holes. This means that even without an *E* field, there is a current due to the diffusion of carriers from regions of high density to low density:

$$J_{diff} = -qD\frac{\partial n}{\partial x}$$

D is called the diffusion constant. It depends on how the carrier motion is interrupted by scattering, typically from impurities and lattice vibrations. Why the minus sign? Consider hole diffusion, for which *q* = +e. If the hole density decreases as x increases then the derivative is negative and holes will diffuse to the right. This constitutes a positive current so we need the minus sign. Both electrons and holes can have a diffusion current. Therefore the full current density for holes (p) and electrons (n) takes the form,

$$J_p = \sigma_p E - eD_p\frac{\partial p}{\partial x} \qquad J_n = \sigma_n E + eD_n\frac{\partial n}{\partial x}$$

Remember that if *E* > 0, holes move to the right and give a positive drift current. Electrons move left but this *also* constitutes a drift current to the right. The electron *diffusion* current has a + sign because *q* = - e.

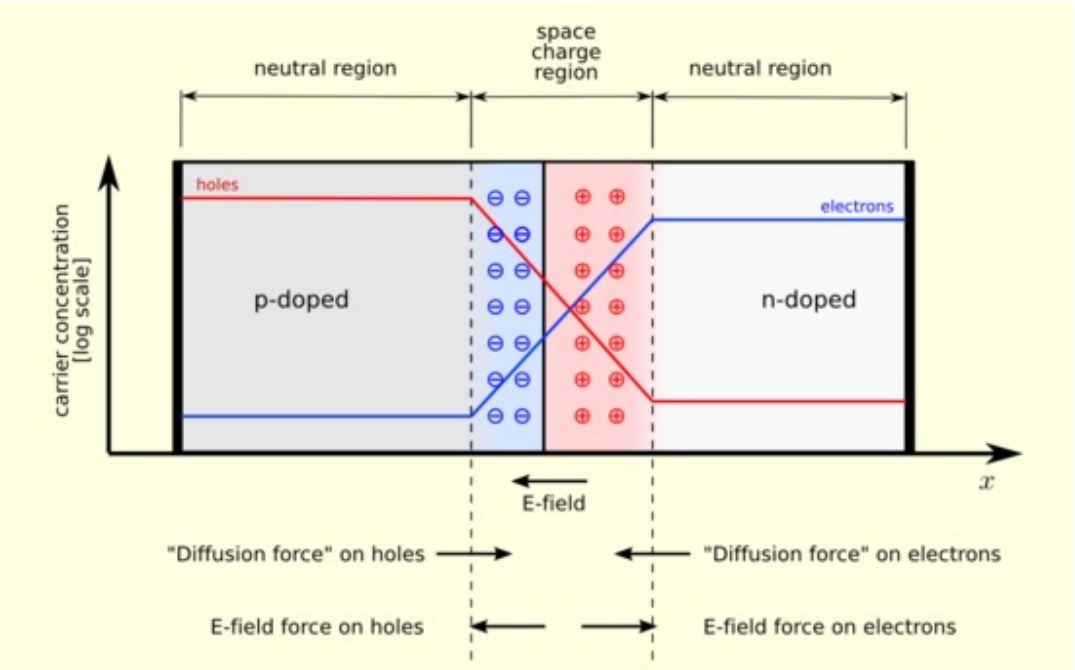
The pn junction

Consider a silicon crystal, the left half of which is p-doped and the right half n-doped, as shown below. In the p-doped region $p = N_A$ (density of acceptor atoms) and in the n-doped region, $n = N_D$ (density of donor atoms.) In general $N_A \neq N_D$. Except near the p-n boundary the material is neutral so there is no electric field. In equilibrium the carrier densities must *everywhere* obey the law of mass action where n_i is the intrinsic density of silicon,

$$n\,p = n_i^2$$

This result implies that *p* must be much lower on the n side than on the *p* side so it drops precipitously as we cross the p-n boundary. Due to this gradient in density, holes from the p-side will diffuse to the n-side and electrons from the n-side will diffuse into the p-side. However, as this happens, holes diffusing right will leave behind a net (-) charge from the fixed acceptor atoms and electrons moving left will leave behind a net (+) charge on the fixed donor atoms. In this *space charge* region near the boundary the free carrier densities both drop, leaving fixed acceptor and donor atoms with their (-) and (+) charges as shown. The free carrier densities are *depleted* so this boundary region is called the **depletion region**.

The arrangement of fixed charges creates a large *E* field in a direction to pull each carrier back to where it's a majority carrier. First focus first on the holes. In equilibrium the net hole current must be zero so the drift and diffusion currents must cancel at each point:



$$J_p = 0 = e\,p\,\mu_p E - eD_p \frac{\partial p}{\partial x}$$

We now borrow a result from statistical mechanics called the *Einstein relation,* that connects the mobility to the diffusion constant for any kind of charge carrier:

$$D = \mu k_B T / e$$

Putting the last 2 equations together and rearranging we have,

$$\frac{dp}{p} = \frac{eE}{k_B T}\,dx = -\frac{e}{k_B T}\,dV$$

$$\frac{dp}{p} = \frac{eE\,dx}{k_B T} = -\frac{e}{k_B T}\frac{dV}{dx}$$

Now integrate this equation across the depletion region from point 1 to point 2:

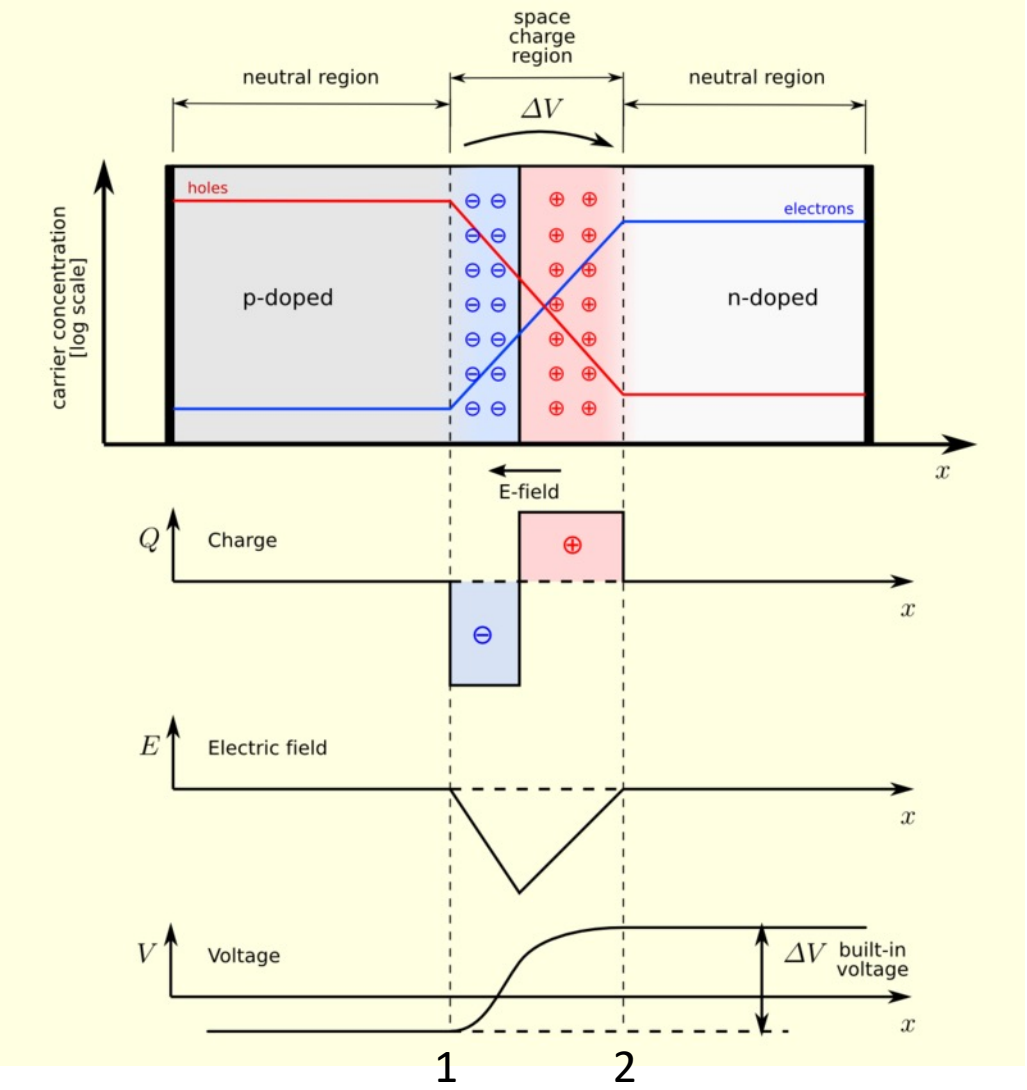
$$\ln\frac{p_2}{p_1} = -\frac{e}{k_B T}\left(V_2 - V_1\right)$$

$$\Rightarrow \quad \frac{p_2}{p_1} = e^{-\frac{e(V_2-V_1)}{k_B T}} = e^{-\frac{\Delta V}{k_B T/e}}$$

This shows that the *E* field in the depletion region leads to a *built-in* potential difference ΔV between the n and p side. Not surprisingly, the ratio of the hole densities at 1 and 2 is just the Boltzmann factor since the holes see a potential energy step eΔV. We could do the same calculation for electrons, which have the same potential energy step as they go from 2 to 1.

How big is this built-in potential ΔV ? Using mass action and the fact than in the neutral regions, $p = N_A$ ($x < x_1$) and $n = N_D$ ($x > x_1$) we have,

$$n_2\,p_2 = N_D\,p_2 = n_i^2 \quad \Rightarrow \quad \frac{p_2}{p_1} = \frac{n_i^2/N_D}{N_A} = e^{-\frac{\Delta V}{k_B T/e}}$$



Wikipedia, pn-junction

Solving for ΔV and inserting typical values for N_A and N_D ($10^{13} - 10^{18}$ / cm^3) and $n_i = 1.5 \times 10^{10}$ /cm^3, we find,

$$\Delta V = \frac{k_B T}{e}\ln\frac{N_A N_D}{n_i^2} \approx 0.3 - 0.9\ V$$

A couple of things are worth noting.

(1) $k_B T/e \approx 26$ mV at $T = 300$ K. It's a number that will come up often in electronics. It's often designated by just V_T .

(2) Let's estimate density of holes in the middle of the depletion region, between points 1 and 2. Take typical value for the built-in potential of ΔV of 0.6 V. In the middle of the depletion region the potential will be about ΔV/2 = 0.3 V. Using the Boltzmann relation we have,

$$\frac{p(middle\,of\ depletion\,region)}{p_1 = p(neutral\ p\ side)} \approx e^{-\frac{\Delta V/2}{k_B T/e}} = e^{-\frac{0.3}{0.026}} \approx 10^{-5}$$

The hole density has fallen by 5 orders of magnitude from its value on the p-side of the junction, where p = N_A. The same argument can be made for the electrons. Therefore most of the charge density in the depletion region comes from the fixed donor and acceptor ions. Of course there must be *some* mobile holes and mobiles electrons in the depletion region in order to carry the current. But since the current = charge density x velocity, the holes and electrons can carry the same total current as in the rest of the bar by moving much faster.

(3) The built-in potential is not a battery! It's just a redistribution of charge in the semiconductor. You can't hook up a voltmeter and measure it since this redistribution of charge builds up changes in potential at the surfaces of the crystal.

Width of the Depletion region

How wide is the depletion region? Referring to the figure, we assume the charge density in the depletion region on the p-side is $-eN_A$ and on the n-side it is $+eN_D$ and we know that $V_2 - V_1 = \Delta V$. Outside the depletion region V = constant. We need to solve Poisson's equation to get the distance between 1 and 2.

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\varepsilon} \qquad \rho_p = -eN_A \quad \rho_n = +eN_D$$

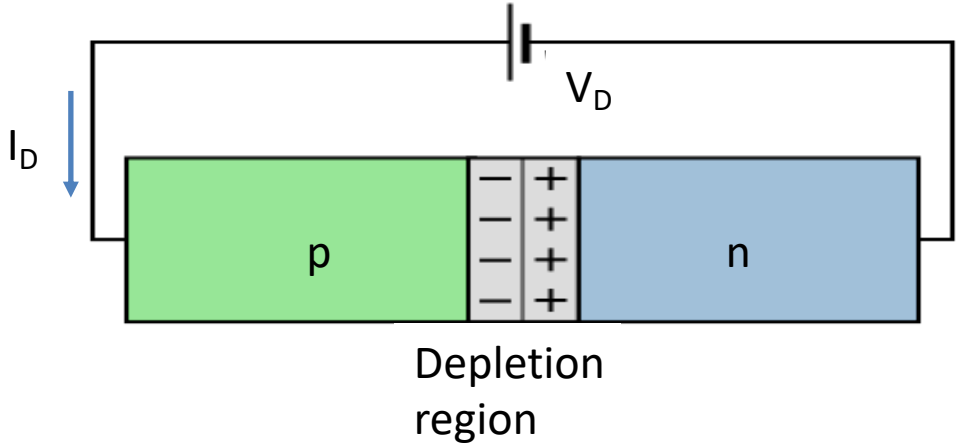
After some algebra the depletion layer width is given by,

$$d_{depletion} \approx \sqrt{\frac{2\varepsilon_{Si}}{e}\left(\frac{N_A + N_D}{N_A N_D}\right)\Delta V} \qquad \varepsilon_{Si} = 11.8\,\varepsilon_0$$

Typical values for the width might be d = 100 – 1000 nm. The depletion region will play a central role in the behavior of semiconductor electronic devices.

Current-voltage characteristic of the diode

Now attach a battery $V_D = V_P - V_N$ across a pn junction. This arrangement will henceforth be called a **diode**. When $V_D > 0$ we say the diode is *forward-biased.* With this polarity, the built-in potential ΔV across the depletion layer is reduced from ΔV to $\Delta V - V_D$. This *lowers* the potential barrier for holes moving left to right and for electrons moving right to left and allows a net current to flow. (The discussion follows C. Kittel, *Introduction to Solid State Physics*.) The convention for diodes is always that $V_D = V_P - V_N$ and the diode current I_D is defined to be positive if it flows *into* the P-side and out the N-side.



Focus first on the holes. In the depletion region there is a right-going diffusion current J_p^{diff} and a left-going drift current J_p^{drift}. When $V_D = 0$ no current flows so these must be equal and opposite:

$$J_p^{diff}(V_D = 0) = -J_p^{drift}(V_D = 0)$$

This diffusion current is also called a *recombination* current since it involves holes that diffuse *from* p to n, become minority carriers and eventually recombine with electrons in the n region. Now turn on V_D. This *lowers* the barrier that holes must surmount to diffuse from the p to the n side. The diffusion current is now enhanced by the Boltzmann factor,

$$J_p^{diff}(V_D) = J_p^{diff}(V_D = 0)\, e^{\frac{eV_D}{k_B T}} \equiv J_p^{diff}(0)\, e^{\frac{V_D}{V_T}} \qquad V_T = k_B T / e \approx 26\,mV$$

By comparison, the hole *drift* current is <u>not</u> changed by V_D . That's because it comes from holes that are continuously generated in the n region to maintain thermal equilibrium. Some of these holes wander to the left, fall into the depletion region and are quickly swept back to the p region by the huge E field in the depletion region. The rate at which this happens, and therefore the current, is limited by how many holes are generated per second, which is not affected by V_D. This contribution is sometimes called the hole *generation* current. The total hole current density is given by,

$$J_p(V_D) = J_p^{diff}(V_D) + J_p^{drift}(V_D) = J_p^{diff}(0)\, e^{\frac{V_D}{V_T}} + J_p^{drift}(0)$$

$$\Rightarrow \quad J_p(V_D) = J_p^{diff}(0)\left(e^{\frac{V_D}{V_T}} - 1 \right)$$

We can make just the same argument for electrons diffusing from the n-side to the p-side. But remember that electrons moving left constitute a positive current moving to the right. The electron and hole currents therefore *add*, giving a total diode current density,

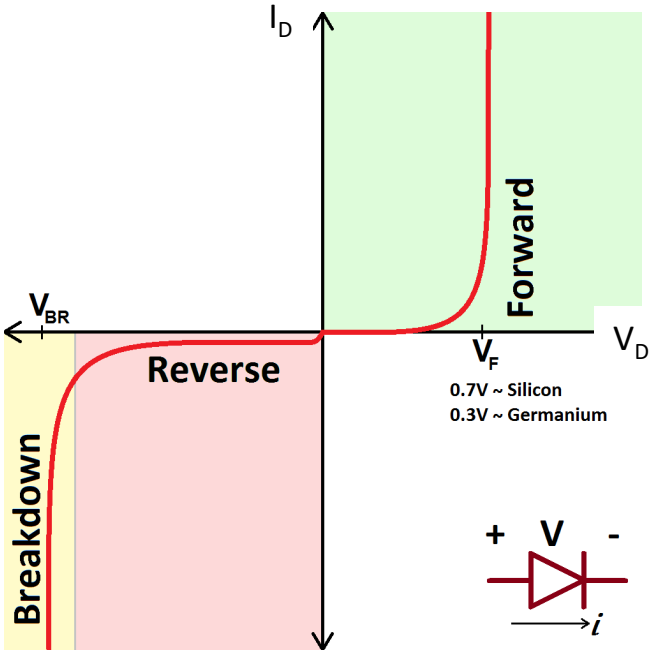
$$J_{total}(V_D) = J_p(V_D) + J_n(V_D) = \left(J_p^{diff}(0) + J_n^{diff}(0) \right)\left(e^{\frac{V_D}{V_T}} - 1 \right)$$

Multiply the current density *J* by the cross-sectional area *A* to get the current *I* in Amps. This current-voltage characteristic often termed the "Law of the Diode",

$$I_D = I_S(T)\left(e^{V_D/V_T} - 1\right)$$

In the forward bias region ($V_D > 0$) real diodes follow this equation moderately well if we multiply V_T by a fudge factor $\eta \approx 1 - 3$ depending on the value of V_D. For reverse bias ($V_D < 0$) the reverse current saturates at $-I_S$ which might be 25 nA for a small-current diode like an IN914. For much larger reverse voltages (V_{BR} = -75 V for an IN914) the diode enters the *reverse breakdown* region in which it conducts much more current. This is not contained in the simple derivation we just did.

Due to the very rapidly increasing exponential, the diode looks like it turns on at at $V_F \approx 0.7$ V in Silicon diodes. The IV curve is actually exponential all the way down to $V_D = 0$ but in many circuits, 0.7 V is an approximate value at which the diode begins to conduct significant current.

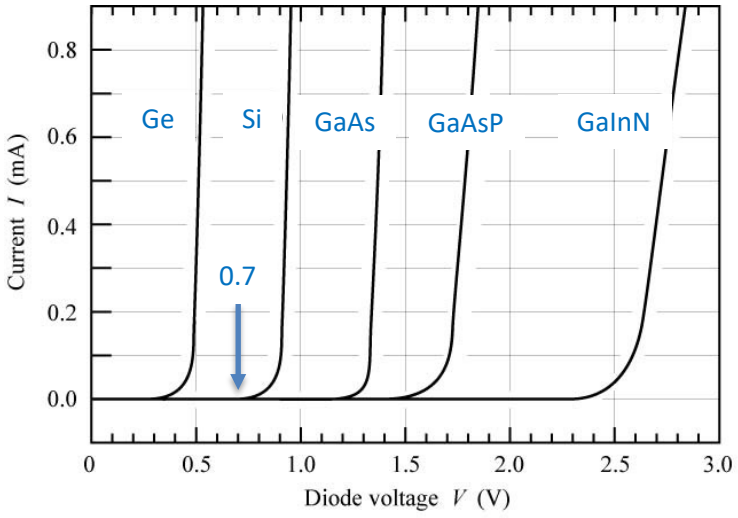


Diode threshold voltage

The prefactor I_S in the diode current-voltage characteristic is proportional to n_i^2 , the square of the intrinsic carrier density. This leads to a dependence of I_S on both temperature and the energy gap E_G of the semiconductor:

$$I_D = I_S(T)\left(e^{V_D/V_T} - 1\right) = A\,T^3 e^{-E_G/k_B T}\left(e^{V_D/V_T} - 1\right) \qquad V_T \equiv \frac{k_B T}{e}$$

Here A is a temperature-independent constant that depends on the diode geometry. All else the same, a diode made from silicon ((E_G = 1.1 eV) requires a smaller V_D to achieve a certain current (say 1 microamp) than the same diode made from GaAs (E_G = 1.4 eV). When you plot the IV curve for a diode it appears to conduct no current until V_D reaches some some threshold voltage. For silicon the threshold is around V_D = 0.7 Volt, shown by the arrow. The diode *does* actually conduct all the way down to V_D = 0, but since the IV characteristic is exponential, the current falls *very* rapidly as V_D goes down. The diode threshold voltage is evident for IV curves made from the different semiconductors as shown below. In LEDs the threshold depends strongly on the color, with blue having a higher threshold than red.



$T = 295$ K

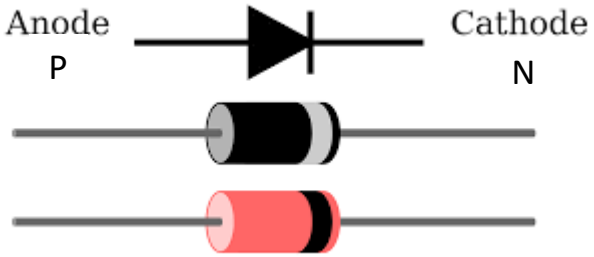
(a)	Ge	$E_g \approx 0.7$ eV
(b)	Si	$E_g \approx 1.1$ eV
(c)	GaAs	$E_g \approx 1.4$ eV
(d)	GaAsP	$E_g \approx 2.0$ eV
(e)	GaInN	$E_g \approx 2.9$ eV

Fig. 4.2. Room-temperature current–voltage characteristics of p-n junctions made from different semiconductors.

Our model for diode conduction is oversimplified. In real diodes, the V_D dependence is more like :

$$I_D = I_S(T)\left(e^{V_D/\eta V_T} - 1\right)$$

The factor $1 < \eta < 3$ varies depending on the size of V_D. However, it turns out that the ideal $\eta = 1$ behavior *is* fairly well-obeyed in transistors.

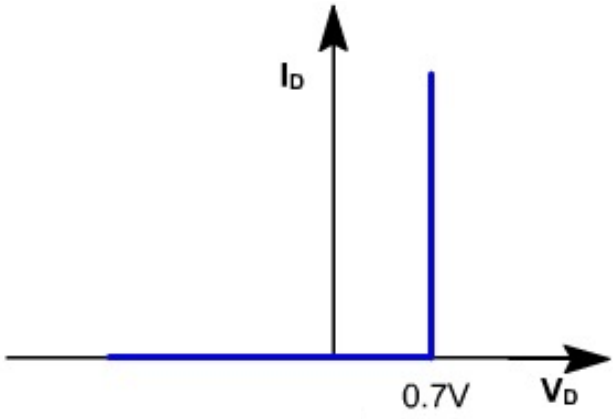


The schematic symbol for a diode is shown along with drawings of actual diodes. Anode and cathode are old terms from the days of vacuum tube electronics. In most diodes the N side (cathode) is marked with a circular band. Diodes come in sizes that can carry maximum currents anywhere from about 300 mA (IN914) to 100 A or more.

The exponential function makes the diode IV characteristic highly asymmetrical. To emphasize this point, just use the law of the diode to estimate how much V_D needs to change to increase the current by a factor of 100:

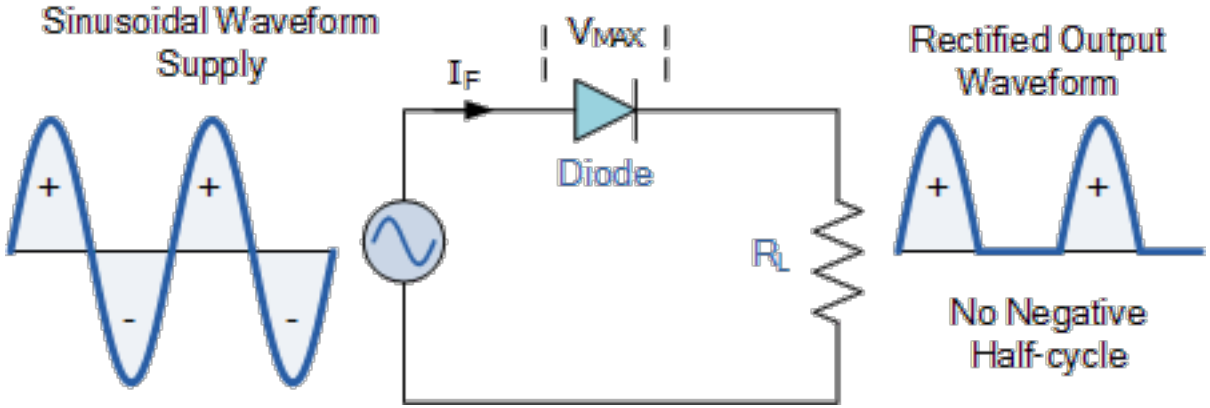
$$\Delta V_D \approx \frac{k_B T}{e} \ln 100 \approx 0.12 V$$

This small change in V_D for a very large change in I_D suggests that to a first approximation, the diode IV curve can be approximated by the piecewise linear function shown in the next figure. The 0.7 V turn-on voltage is, again, just an approximation. Current does flow below 0.7 V but it's typically down in the micro to nanoamp range. The basic idea is that the diode is a one-way valve for current. Lots of current when forward-biased and very little current when reverse-biased. That's the feature used to turn an AC (alternating current) signal like the one coming out of your power wall socket, into a waveform that has a non-zero time average – a DC (direct current).

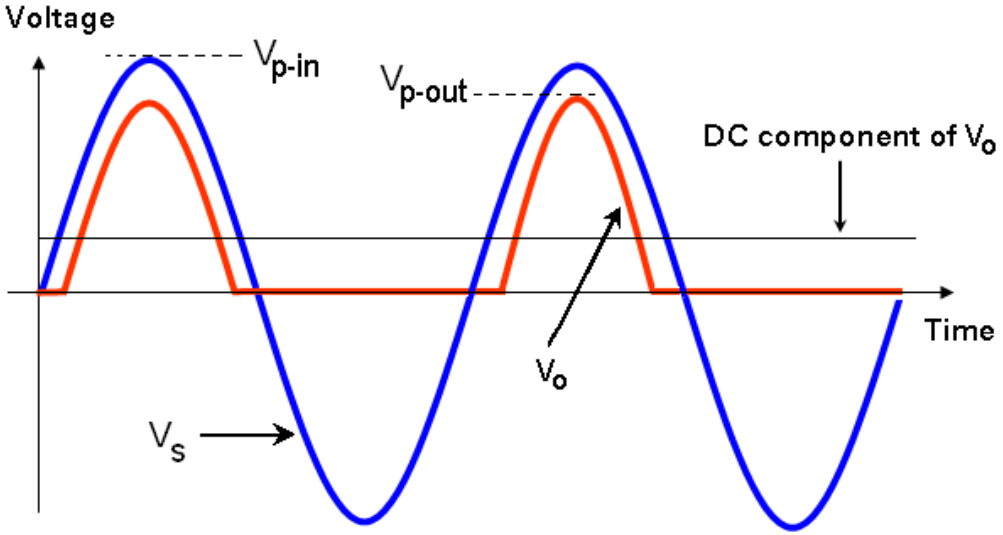


Rectifier

Diodes are widely used for *rectification*: turning an AC signal into a waveform with a DC component. The simplest rectifier circuit is shown below. The diode conducts only on positive-going half-cycles of the input. It's therefore called a *half-wave* rectifier. The upper figure ignores the 0.7 V threshold voltage of the diode. The lower figure shows the effect of the threshold voltage on the output. (*www.electronics-tutorials.ws*)



www.electronics-tutorials.ws

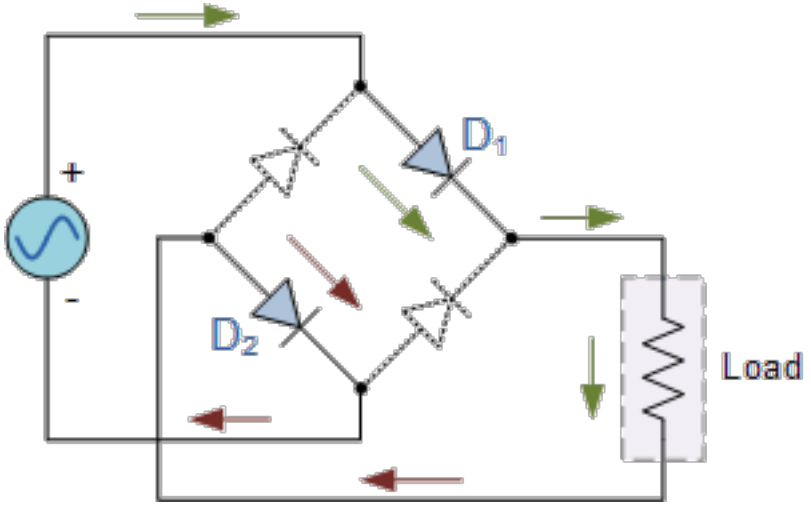


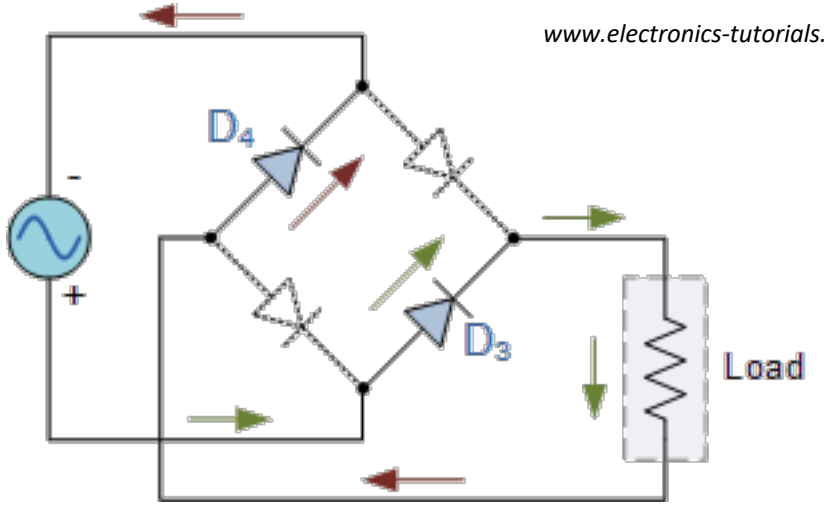
Taking the 0.7 V threshold into account, the output (red) is zero until the input exceeds about 0.7 V. The diode then conducts and the output remains 0.7 V below the input. Once the input falls below 0.7 V the output again falls to zero and remains there for half a cycle. The DC component is given by,

$$V_{DC} = \frac{1}{T}\int_0^T V_{out}(t)\,dt$$

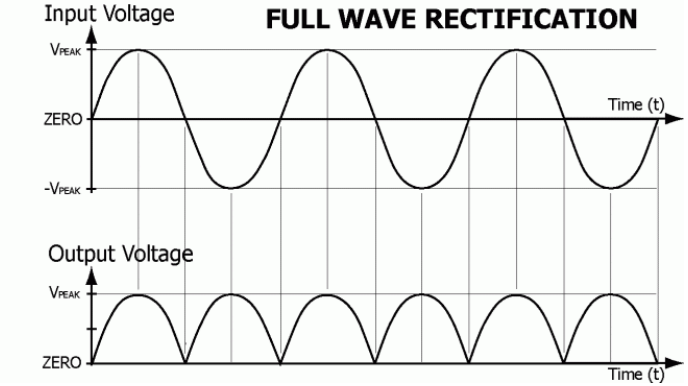
Full wave rectifier – diode bridge

Half-wave rectification isn't efficient since the output is zero half the time. You can do better by using 4 diodes, as shown below. During one half cycle, diodes D_1 and D_2 conduct while on the other half cycle, D_3 and D_4 conduct, producing the waveform shown below. Again, we're ignoring the 0.7 V diode drop which is okay if the input sinewave amplitude >> 0.7 V. The 4 diode arrangement is called a *diode bridge*. These come packaged as one unit with current-capability generally > 1 amp. They are in virtually every power supply you will find. It will require more electronics to convert the rectified output into a nice, steady, purely constant voltage.



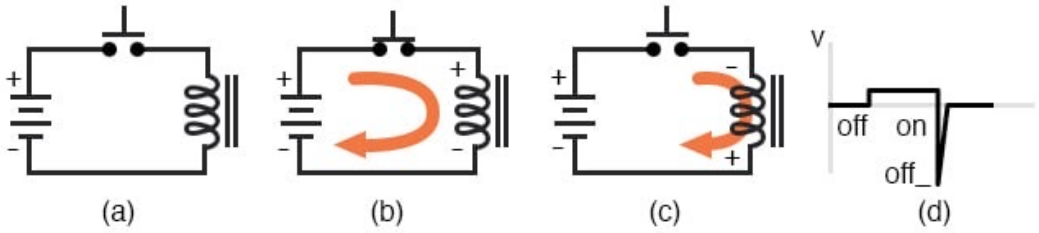


www.electronics-tutorials.ws



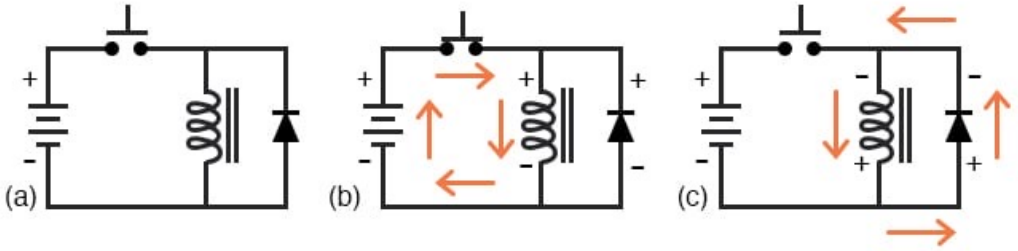
Diodes in protection circuits

In the circuit below, suppose you close the switch. Current will flow through the inductor in the direction shown. Now open the switch. Faraday's law says that we cannot the inductor will generate a back-EMF that tries to keep the current going in the same direction. This can create a potentially damaging voltage spike, as shown. This is often called "inductive kickback".



https://www.allaboutcircuits.com/textbook/semiconductors/chpt-3/inductor-commutating-circuits/

To protect against this voltage spike, add a diode in parallel with the inductor. When the switch is close the diode is reverse-biased no current flow through it and it's irrelevant. But when the switch is opened, the back-EMF now *forward biases* the diode and current flows in the loop shown in part c. As we've seen, the diode can accommodate a very wide range of current and maintain a voltage of about 0.7 V or less. So the original voltage spike across the inductor is now limited to about 0.7 V. Diodes acting in this way, as voltage limiters, are found throughout modern electronics.



https://www.allaboutcircuits.com/textbook/semiconductors/chpt-3/inductor-commutating-circuits/

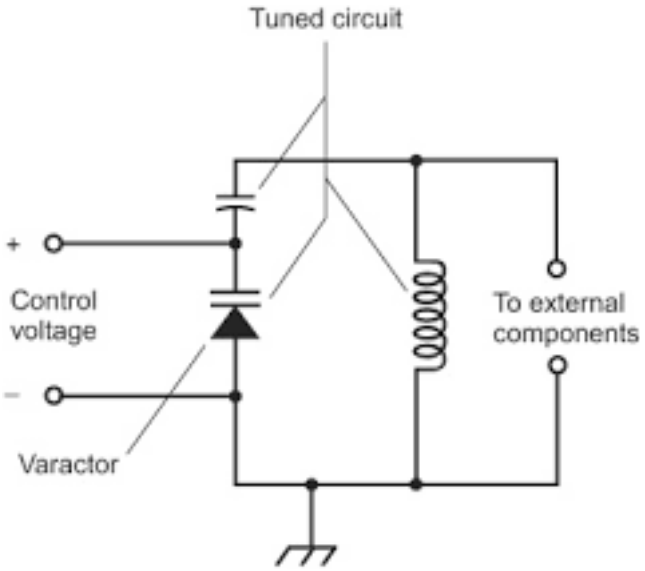
Varactor diodes

Recall that the width of the depletion layer is given by, $$d_{depletion} = \sqrt{\frac{2\varepsilon_{Si}}{e}\left(\frac{N_A + N_D}{N_A N_D}\right)\Delta V}$$

Biasing the diode by V_D implies that we replace ΔV by $\Delta V - V_D$. Forward biasing ($V_D > 0$) reduces W and reverse biasing ($V_D < 0$) increases W. In the reverse-biased state, the depletion region acts like a capacitor with plates of area *A* and spacing *W*. (This is plausible but takes some proof.) It's called the *transition capacitance* and is given by,

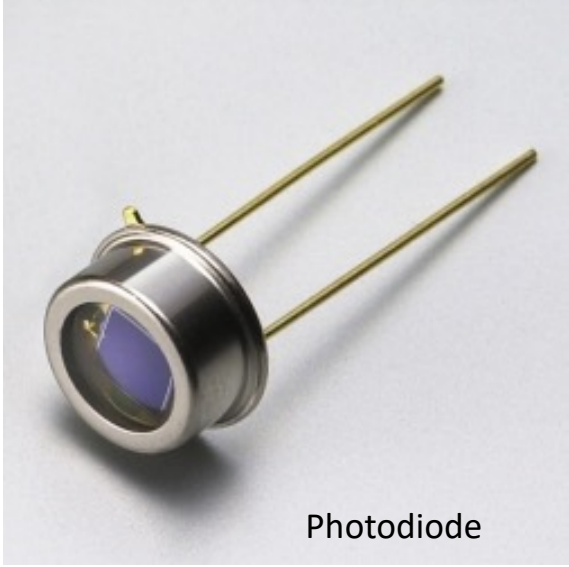
$$C_{transition} = \frac{A}{\varepsilon d_{depletion}} \propto \frac{1}{\sqrt{(\Delta V - V_D)}}$$

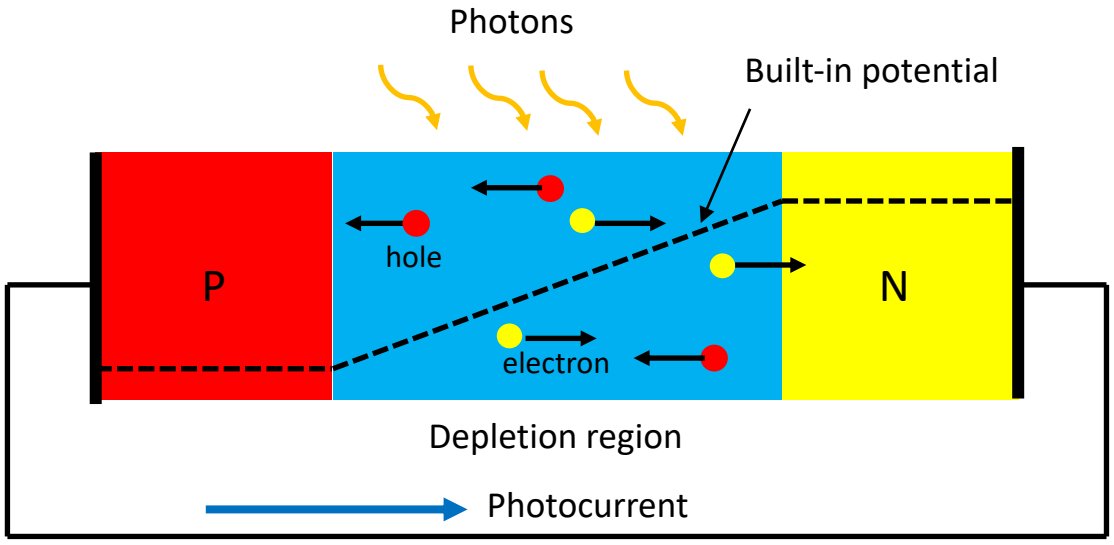
A reverse-biased diode can therefore be used as a voltage-variable capacitor. Diodes optimized for this application are called *varicaps*, *varactor diodes* or just *varactors*. The range of capacitance variation is typically 5 – 100 pF which is very useful for applications in high frequency tuned circuits and oscillators. The circuit on the right shows an LC tuned circuit where part of the capacitance is a varactor. The resonant frequency of the circuit $\omega_0 = (LC)^{-1/2}$ can be tuned by changing the control voltage that reverse-biases the varactor and changes the total capacitance. The diode symbol with the two capacitor bars indicates a varactor diode.



Photodiodes, solar cells

Imagine shining light on a PN junction. Photons with energy larger than E_{gap} = 1.1 eV can excite electrons from the valence band into the conduction band, creating an electron and a hole. In the depletion region the built-in potential will force newly created holes to move toward the P-region and newly created electrons to move toward the N-end of the junction. If we connect the two ends with a wire a current will flow *out* of the P side and *into* the N side, as shown. In effect, the PN junction, together with incoming photons, now constitutes a source of current. This photo-induced current can be used to detect light, in which case we call it a photodiode, or it can be used to power a washing machine, in which case we call it a solar cell. Obviously the demands for each are very different so the details of how the PN junction is designed and the circuit to which it's attached will be very different. But the basic idea is the one shown. (An excellent reference can be found at https://www.pveducation.org/pvcdrom/welcome-to-pvcdrom/instructions .


Photodiode




Solar panels

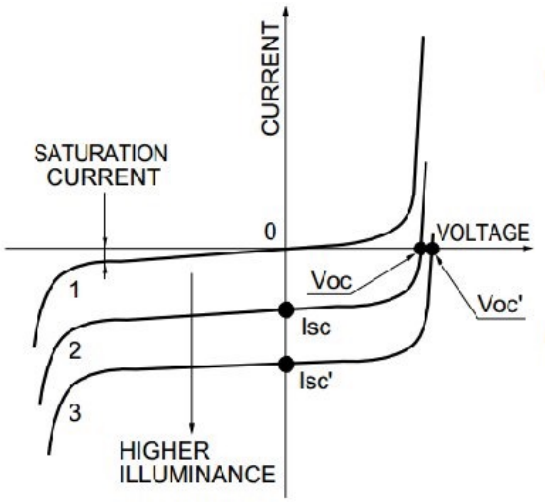
In electronics, a voltage source provides the same *voltage* regardless of what is hooked across its terminals. A battery is one example, though it's less than ideal. A current source, though less familiar, provides the same *current* regardless of what is attached across its terminals. A photodiode (plus the incoming flux of photons) is an example of a current source (again, less than ideal.)

Photodiodes and solar cells work on the same principles. The total current consists of a piece that's just like an ordinary diode and an additional contribution from the electron-hole pairs generated by the incoming photons:

$$I_D = I_S \left(e^{V_D/V_T} - 1\right) - I_{Photo}$$

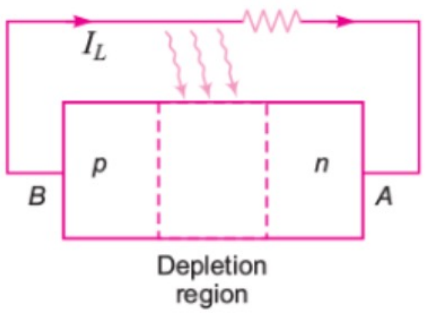
The behavior is better illustrated by the IV (current-voltage) characteristic shown. Looking at the previous page, if you short circuit the terminals then $V_D = 0$. A current I_{SC} will flow *out* of the P-terminal and *into* the N-terminal. That's the photocurrent and it is negative. It's magnitude increases with the number of photons illuminating the diode. On the other hand, if you open-circuit the diode then the total current equals zero. A "photovoltaic" voltage V_{OC} builds up across the diode terminals in order to satisfy the equation,

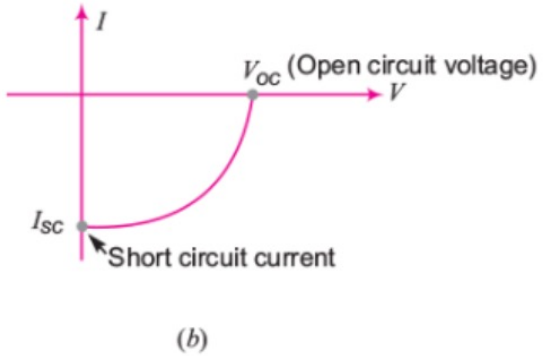
$$0 = I_0\left(e^{V_{OC}/V_T} - 1\right) - I_{Photo}$$



https://www.toppr.com/ask/question/describe-the-briefly-using-the-necessary-circuit-diagram-the-three-basic-processes-which-take-place/

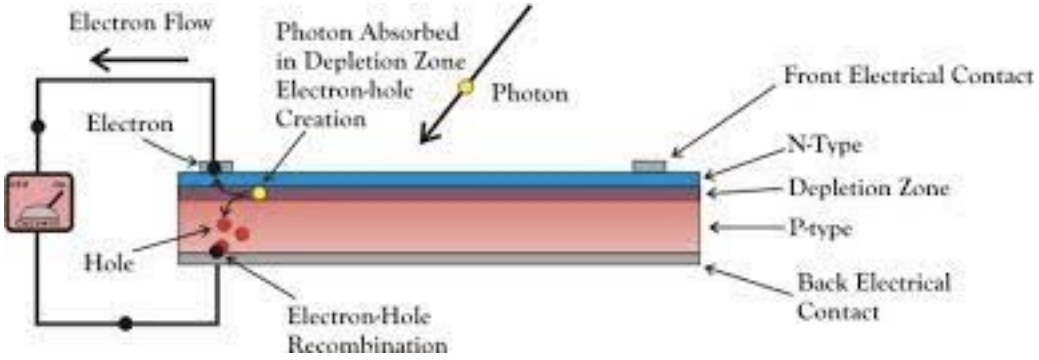
Neither short circuit nor open circuit permits us to dissipate any power in the external circuit. For a solar cell we certainly want to use the sun's energy to provide power so attach a resistor between the photodiode terminals as shown. The flow of current through the resistor generates, by Ohm's law, a positive diode voltage ($V_p > V_n$) so the circuit would operate somewhere on the IV curve between short circuit and open circuit points.



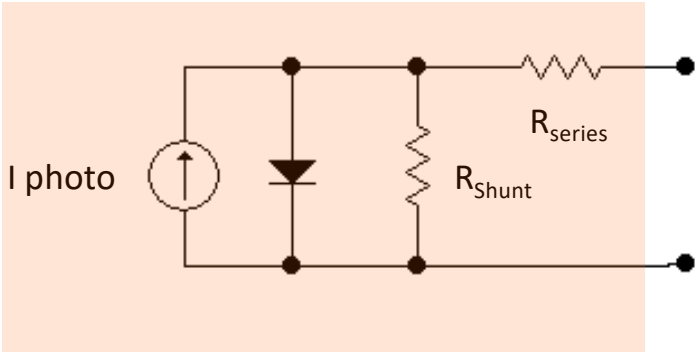


Solar Cells

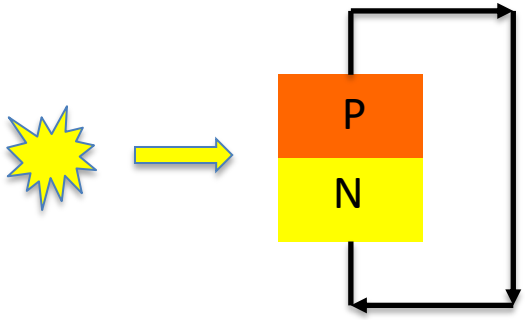
The geometry of a solar cell is shown below. The effective circuit looks like that of a photodiode detector, except that we don't care about the dark current or capacitance. The shunt resistance comes from manufacturing defects while the series resistance is intrinsic. An excellent reference on solar cells can be found at *https://www.pveducation.org/pvcdrom/welcome-to-pvcdrom/instructions.*

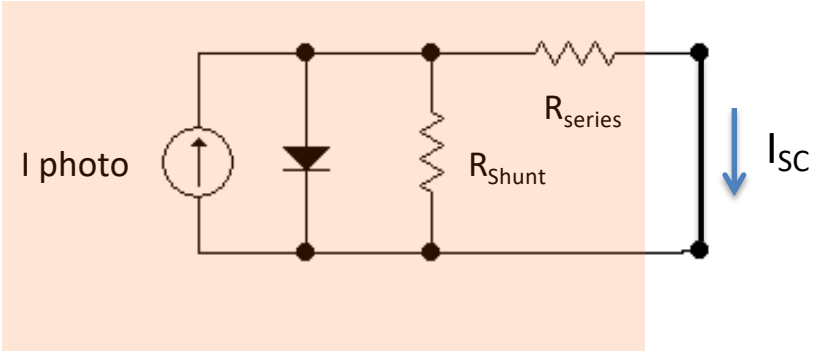


Imagesco.com



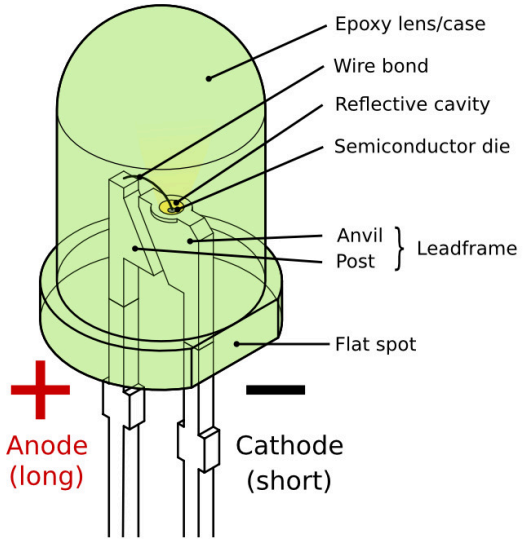
If you short circuit the terminals, current will flow *from* the P terminal *to* the N terminal which, again, is a diode *reverse* current.



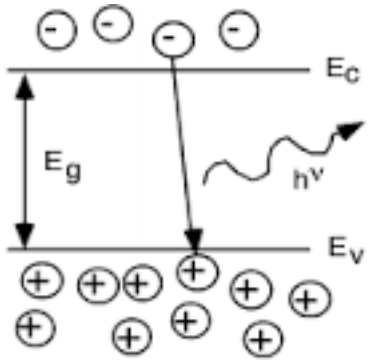


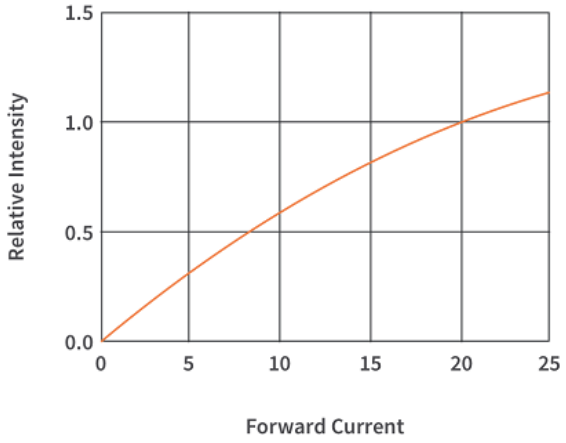
LEDs

If we forward bias the diode, diffusion currents carry holes through the depletion region and into the neutral n region. Similarly, electrons will diffuse into the neutral p region. Eventually these excess minority carriers will recombine with the majority carriers and establish the equilibrium densities inside the neutral regions. Recombination amounts to an electron leaving the conduction band and filling a hole in the valence band as shown below. The energy difference is usually given off as heat (lattice vibrations) but for some semiconductors the process can emit a photon, in the same way an atom emits light by jumping from one energy level to another. Then the diode becomes an LED. Common semiconductors for LEDs are GaAs, AlInGaP and InGaN. While many people contributed it's fair to say that the primary development of the LED came from the late Prof. Nick Holonyak in the ECE department of the University of Illinois.

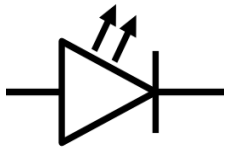


https://soldered.com/learn/led-light-emitting-diode-explained/





The schematic symbol for an LED is shown below. As a rule of thumb, you can light up a typical LED with 5-10 mA of current. The IV characteristic for an LED is very similar to a conventional diode, but the turn-on voltage is higher, typically 2-3 V depending on the color.



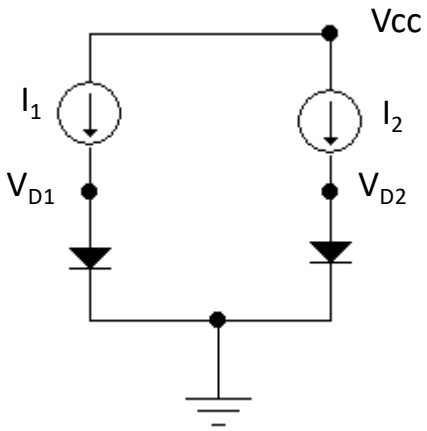
Diode thermometry

At fixed current I_D the diode voltage V_D depends on temperature T: $V_D = V_T \, ln\left(\frac{I_D}{I_S}\right)$. In principle V_D could be used as a thermometer. As it stands, however, the T-dependence comes from both I_S(T) and the thermal voltage V_T, making things too complicated. The dependence on I_S(T) can be eliminated by using two identical diodes driven by constant current sources I_1 and I_2. (See *Microelectronic Circuit Design*, p. 88 by R.C. Jaeger and T.N. Blalock, , McGraw Hill 2011).

By measuring the *difference* between the two diode voltages we eliminate the dependence on I_S. The result is a voltage directly proportional to the absolute temperature in Kelvins:

$$V_{D1} - V_{D2} = V_{PTAT} = V_T \, ln\left(\frac{I_{D1}}{I_S}\right) - V_T \, ln\left(\frac{I_{D2}}{I_S}\right) = V_T \, ln\left(\frac{I_{D1}}{I_{D2}}\right)$$

$$V_{D1} - V_{D2} = V_{PTAT} = \frac{k_B T}{e} \, ln\left(\frac{I_1}{I_2}\right)$$



I_1 and I_2 is a fixed ratio and doesn't depend on temperature. V_{PTAT} is referred to as a *PTAT* (proportional to absolute temperature) voltage. This concept is widely used in digital thermometers of the type you might buy at a drugstore. They're typically accurate to about 0.1 K. Later, we'll see how transistors, op amps and feedback can improve on this diode circuit. The main takeaway is the use of a *differential circuit* to eliminate the dependence on certain device parameters.

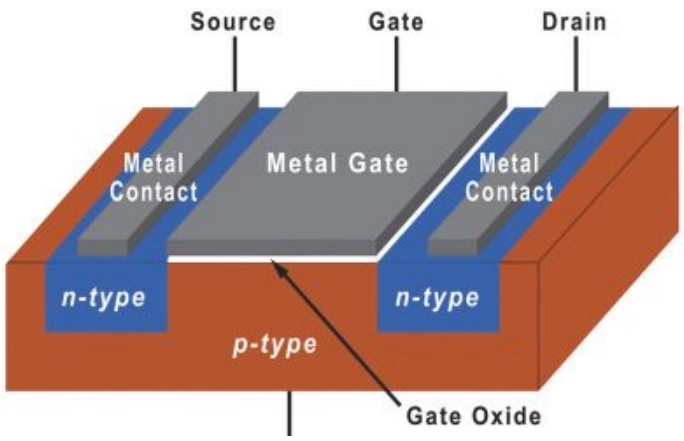
MOSFETs

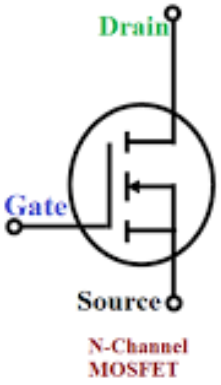
The transistor is a three terminal device that makes the modern electronic world possible. There are two basic kinds: FETs (field effect transistors) and BJTs (bipolar junction transistors). Both kinds are used for analog electronics but the digital world is ruled by FETs. In fact, the FET is the most widely manufactured object in human history. You probably own a few billion yourself. We'll focus on the MOSFET (*metal oxide insulated* field effect transistor) which is by far the most common variety and the type used in your phone and computer.

We'll begin with the n-channel MOSFET, shown in perspective as well as its schematic symbol. It consists of two n-type regions, called source and drain, embedded in a p-type substrate, together with a metallic gate isolated from everything else by a thin layer of insulating oxide material. The main thing to understand about this type of MOSFET is that current flows between source and drain through a *channel* that is induced by applying a voltage between the gate and the substrate. The gate/oxide/substrate sandwich acts like a capacitor. Applying a positive voltage to the gate induces a sheet of electrons right at the edge of this p-type substrate. This sheet is called an inversion layer because it's a region of electrons sitting in an otherwise p-type material. But insofar as the MOSFET behavior is concerned, we can treat it like a capacitor. There is one hitch. In order to induce this channel, the voltage between gate and substrate must exceed a threshold voltage V_{Th} , which might be a volt or more. A FET $V_{Th} > 0$ is called an *enhancement-mode* FET and this is indicated in the schematic by the 3-segment line.

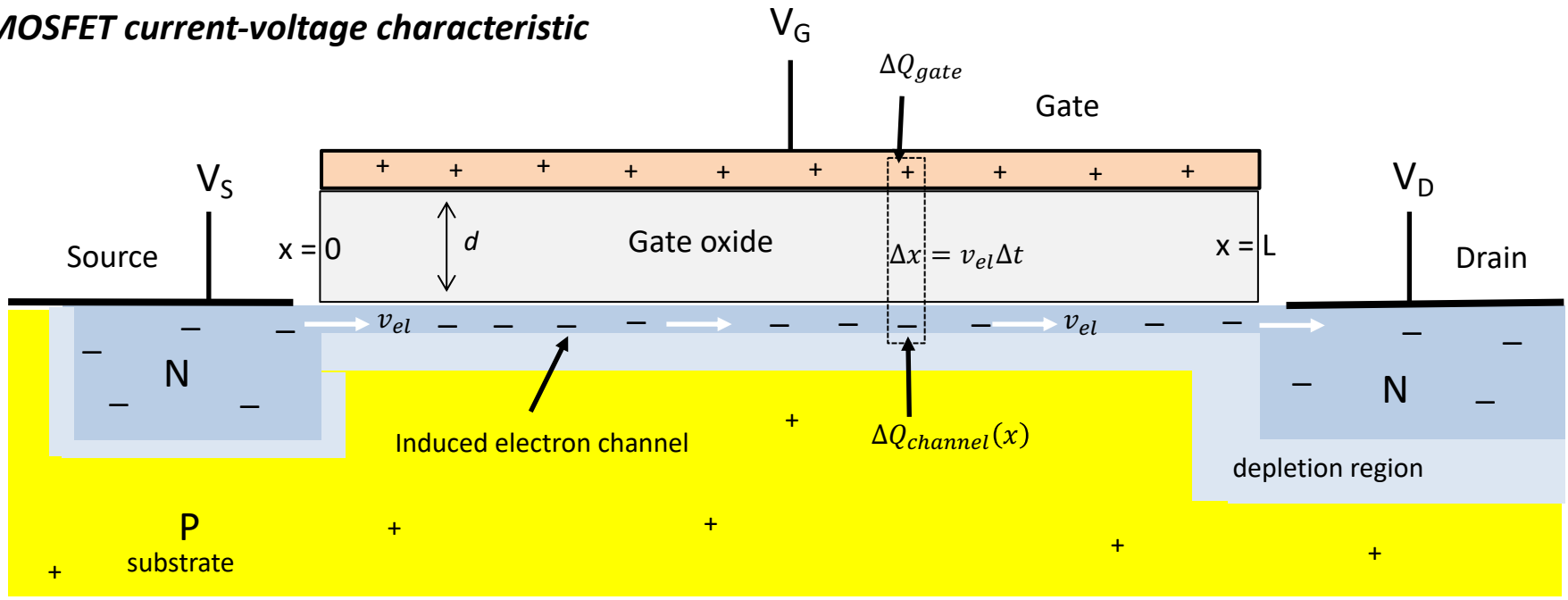
Usually the substrate is internally connected to the source, as indicated by the schematic symbol on the left and in our derivation, all voltages will be measured relative to the source voltage. Both the source/substrate and drain/substrate form pn junctions. In general those junctions should be reverse-biased so no current should *ever* flows through the substrate. For now, think of the substrate as a source of charge carriers.

https://www.mks.com/n/mosfet-physics





MOSFET current-voltage characteristic



Triode region

The figure shows the cross section of an N-channel MOSFET. Imagine that it extends into the page to a distance W. Assume that we've connected the substrate electrically to the source and grounded both. Now apply a voltage between gate and source. If $V_G - V_S = V_{GS} > V_{Th}$ then negative charge will be induced in the region between the gate oxide and the substrate, just like a capacitor. V_{Th} acts like a little battery in opposition to the applied gate-source voltage V_{GS}. The charges induced between the substrate and gate oxide now form a conducting channel for electrons. Consider a tiny region Δx wide, enclosed by the dashed lines. This is a little capacitor so the charge induced in the channel (i.e., the lower capacitor plate) will be,

$$\Delta Q_{channel}(x) = -\Delta Q_{gate}(x) = -C\big(V_{GS} - V_{Th} - V_{channel}(x)\big)$$

The capacitance of this little section will be $C = \frac{\epsilon \Delta x\, W}{d}$ where ϵ is the dielectric constant of the oxide and *d* its thickness. $V_{channel}(x)$ is the voltage in the channel. Now connect a power supply between the drain and source so $V_D > V_S$. $V_{channel}(x)$ will now increase from V_S at x = 0 to V_D at x = *L* causing electrons to flow from source to drain with an average velocity v_{el}. That constitutes a positive current I_{DS} *from* drain to source. To find I_{DS} consider a tiny interval Δt during which time all the electrons in region Δx will move to the right by $\Delta x = v_{el}\Delta t$. The current is given by the charge passing through the dotted line per unit time,

$$I_{DS} = -\frac{\Delta Q_{channel}}{\Delta t} = \frac{v_{el}\,\epsilon W}{d}\left(V_{GS} - V_{Th} - V_{channel}(x)\right)$$

The (-) sign is there because a flow of electrons to the right is equivalent to a positive current flowing to the left. What about v_{el} ? That depends on the electronic mobility μ_{el} ,

$$v_{el} = \mu_{el} E_{channel} = -\,\mu_{el}\,\frac{dV_{channel}}{dx}$$

Putting it all together gives,

$$I_{DS} = \frac{\mu_{el}\,\epsilon W}{d}\left(V_{GS} - V_{Th} - V_{channel}(x)\right)\frac{dV_{channel}}{dx}$$

Now integrate both sides from x = 0 to x = *L*. The current is the same throughout the channel so it comes outside the integral,

$$\int_0^L I_{DS}\,dx = I_{DS}L = \int_{V_S}^{V_D} \frac{\mu_{el}\epsilon W}{d}\left(V_{GS} - V_{Th} - V_{channel}(x)\right)dV_{channel}$$

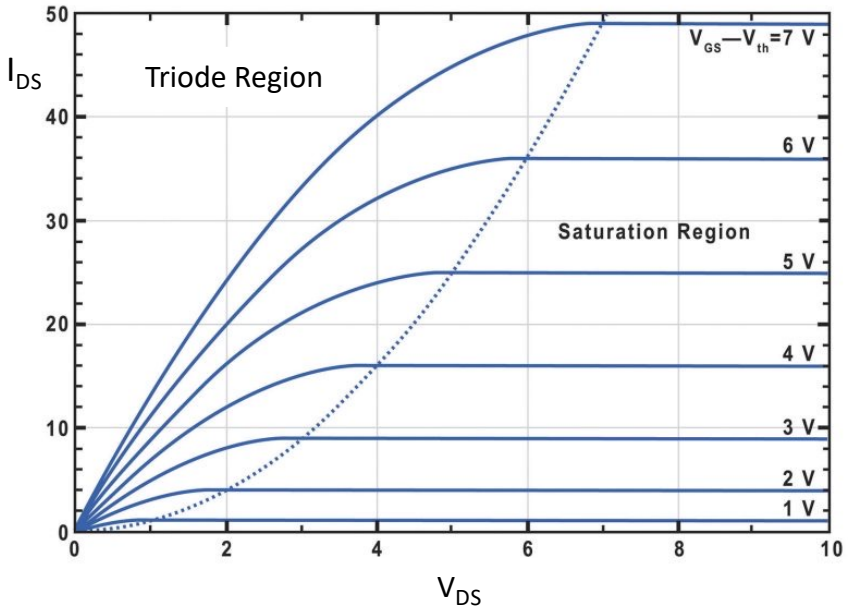
Doing the right hand integral we obtain the IV characteristic of the MOSFET,

$$I_{DS} = \frac{\mu_{el}\,\epsilon W}{L\,d}\left((V_{GS} - V_{Th})V_{DS} - \frac{1}{2}V_{DS}^2\right) = 2\kappa\left((V_{GS} - V_{Th})V_{DS} - \frac{1}{2}V_{DS}^2\right) \qquad \kappa \equiv \frac{\mu_{el}\,\epsilon W}{2\,L\,d} \qquad \textbf{\textit{Triode region}}$$

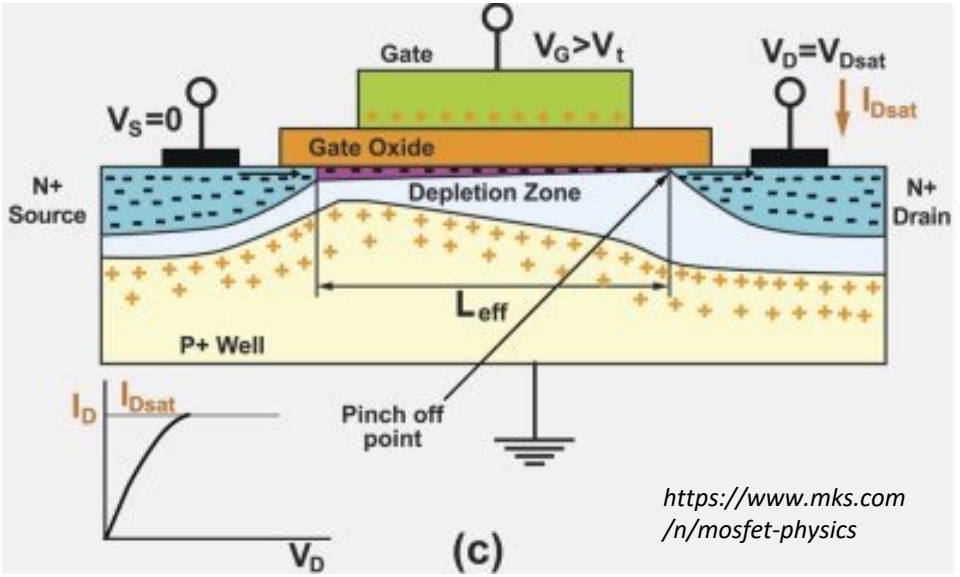
This equation holds true *so long* as there is induced charge everywhere in the channel from x = 0 to x = *L*. The transistor is a 3-terminal device so V_{GS} and V_{DS} can be controlled independently. The full IV characteristics are shown on the next page. The equation we just derived corresponds to all curves on the left of the dotted line, the so-called *triode* region of the MOSFET. Each curve corresponds to a different value of $V_{GS} - V_{Th}$.

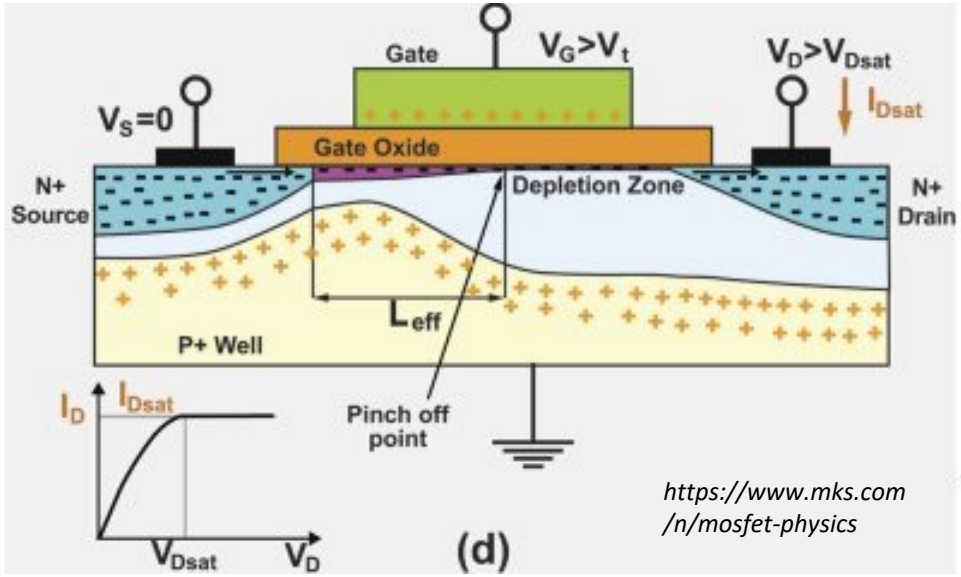
Pinch off and saturation

The figure shows that beyond the triode region I_{DS} stops increasing with V_{DS} and flattens off. This region is called *saturation*. Think of the channel like a capacitor plate. In order to induce charge in the channel the MOSFET requires that $V_{GS} - V_{Th} > V_{channel}(x)$. The channel voltage is largest at x = *L* where $V_{channel}(x) = V_{DS}$. Once $V_{DS} > V_{GS} - V_{Th}$ then $V_{GS} - V_{Th} - V_{channel}(x = L) \leq 0$ and no charge can be induced at x = *L*. The channel starts to *pinchoff* as shown in the next cross sectional view.



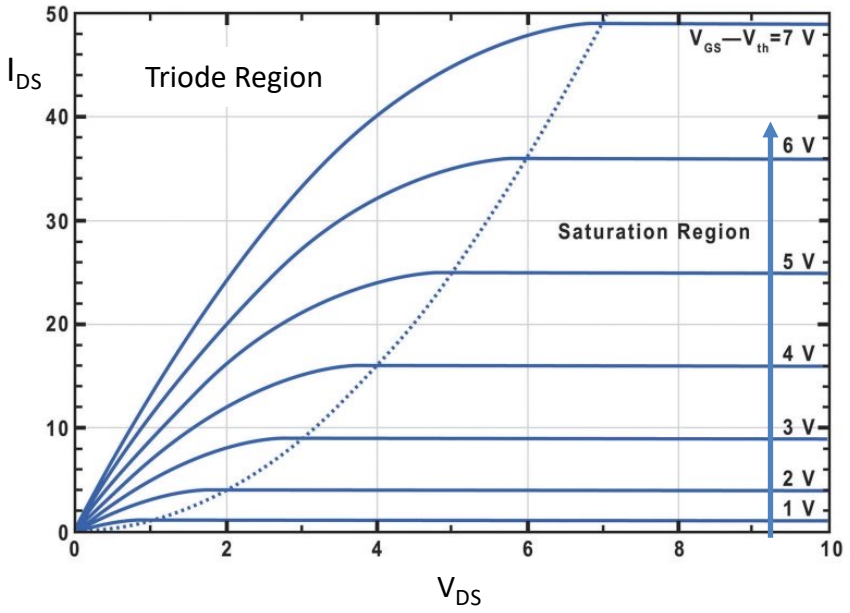
https://www.mks.com/n/mosfet-physics





Now increase V_D still further. There will be a longer stretch in the region under the gate oxide where V_{DS} is too high to induce a channel. The induced channel length will decrease from L to L_{eff} and the region between L_{eff} and L will have no free charge. Electrons *are* able to travel through this region but their velocity no longer depends on the electric field so I_{DS} no longer depends on V_{DS} . The *saturation region* is where the MOSFET can be used in analog circuits as an amplifier. It also plays a role in digital electronics as circuits switch from one logical state to another.

Notice also that the depletion zone between the p-type substrate and the n-type drain and channel becomes wider as V_{DS} increases. That's because the substrate/channel interface acts like reverse-biased pn junction. The larger is V_{DS}, the larger the reverse-bias and the larger the width of the depletion region.

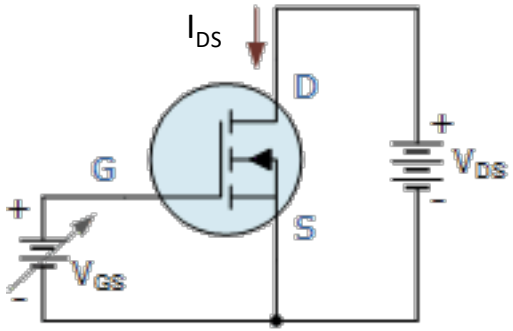


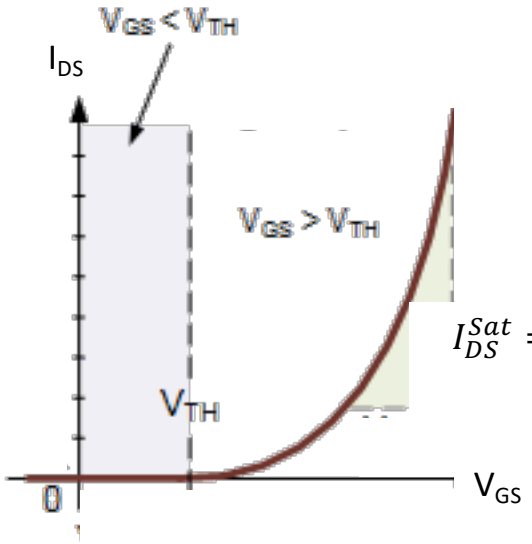
The current in the saturation region is just the value attained once the channel starts to pinch off and that occurs when $V_{DS} = V_{GS} - V_{Th}$. Substituting that value into the equation for the triode region we have,

$$I_{DS}^{Sat} = \kappa\,(V_{GS} - V_{Th})^2 \qquad \textbf{\textit{Saturation region}}$$

In the figure on the left focus on the blue arrow. Hold V_{DS} fixed and increase V_{GS} . No current flows until V_{GS} - V_{Th} > 0 after which I_{DS} increases quadratically.

The simple circuit shown below summarizes the behavior in the saturation region. The red curve shows I_{DS} versus V_{GS} with V_{DS} held fixed by the power supply. So long as you stay in the saturation region, you'll get the same curve regardless of the value of V_{DS} . In real FETs the curve is not exactly flat but has a slight positive slope due to details which our simple model has ignored.



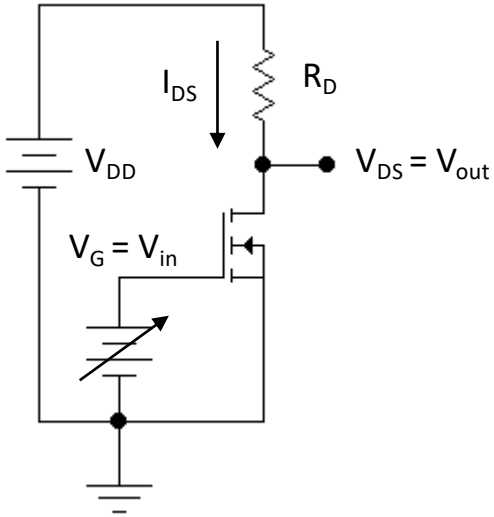


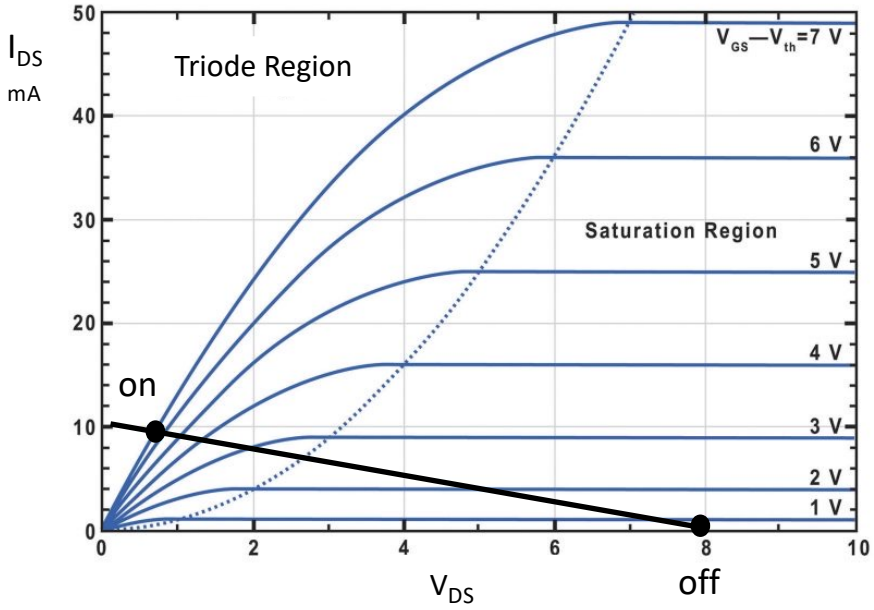
$$I_{DS}^{Sat} = \kappa\,(V_{GS} - V_{Th})^2$$

A simple logic circuit.

To see how the MOSFET can be used in digital electronics, consider the circuit below. We've simply added a resistor R_D between the drain and the power supply voltage V_{DD} which we'll assume is 8 Volts for illustration. It's not obvious whether the MOSFET will be in the triode or the saturation region. To see it graphically, use a basic result from circuit theory – that the sum of the voltage changes around any loop must equal zero, otherwise known as Kirchoff's voltage law. Going clockwise around the outer loop we have,

$$V_{DD} - I_{DS}R_D - V_{DS} = 0 \quad \rightarrow \quad I_{DS} = \frac{V_{DD}}{R_D} - \frac{V_{DS}}{R_D}$$



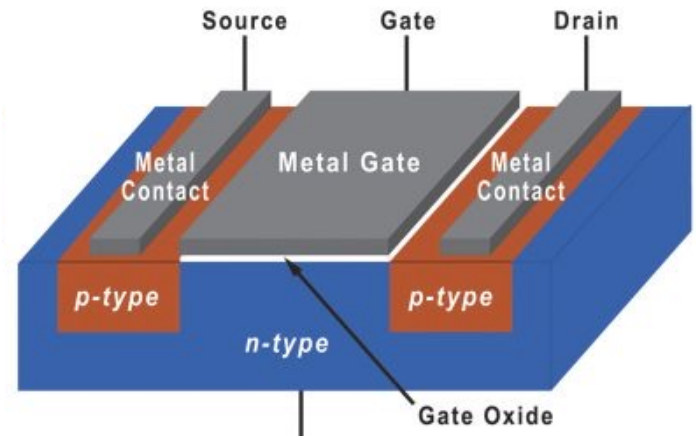


This linear relation between I_{DS} and V_{DS} is called the *load line*, shown as a black line on the IV characteristics. That's one equation in two unknowns. The other equation is the IV characteristic of the MOSFET. The solution is the intersection the IV curve and the load line. Which IV curve? That's determined by the gate voltage, or more precisely, $V_{GS} - V_{Th}$. Suppose $V_{Th} = 1\,V$. If we set $V_{GS} = 0$ the FET will turn off and no current will flow, as marked **off.** If no current flows then $V_{DS} = V_{DD} = 8$ Volts. Alternatively, set $V_G = 8$. Then $V_{GS} - V_{Th} = 7$. The solution is the point marked **on**. In that case $V_{DS} \approx 0.6\,V$.

This is a logic circuit in which the input is the gate voltage and the output is the drain voltage. If we call logical 1 any voltage above 7 Volts and logical 0 any voltage below 1 volt then you can see that this circuit performs a logical NOT operation, $V_{out} = \bar{V}_{in}$.

P-channel MOSFET

It turns out that resistors are difficult to fabricate with the enormous density required in modern digital logic chips. In addition, they dissipate power. In the previous circuit, there is a non-zero current flowing when the MOSFET is on. That dissipates power $P = I_{DS}^2 R_D$ and when you have hundreds of millions of circuits in a very small place the heating would be intolerable. The solution is to replace the resistor by another MOSFET, but this time the *complement* of the n-channel MOSFET we just discussed. That is, make the substrate n-type silicon and make source and drain both p-type. The geometry is the same but now, when we apply a *negative* voltage between gate and source, that induces a conducting channel of *holes* between source and drain. It's called a p-channel MOSFET.



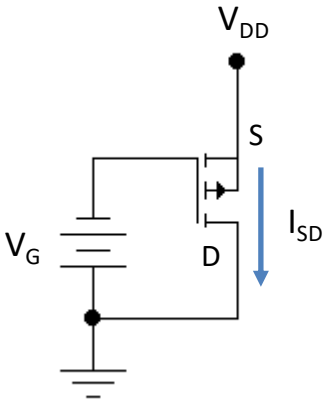
https://www.mks.com/n/mosfet-physics

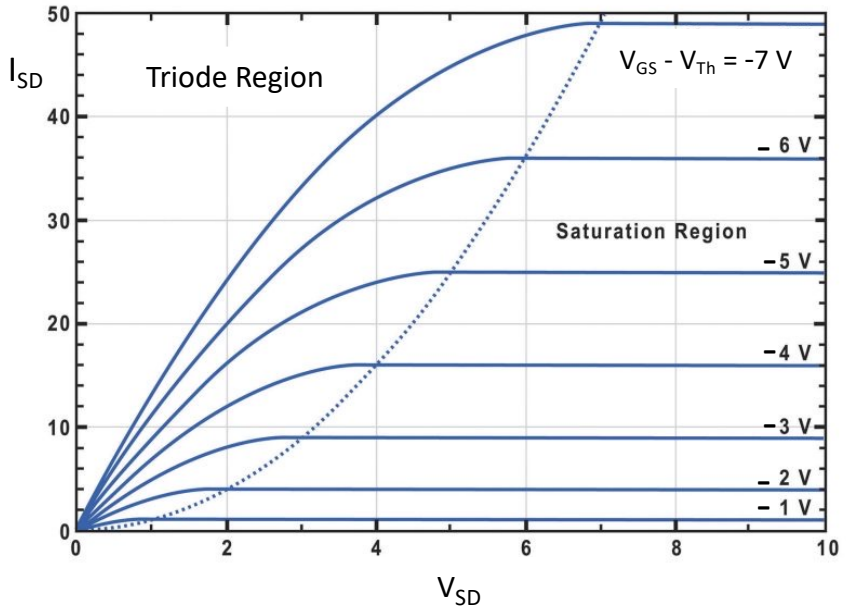
p-channel MOSFETs work just like n-channel MOSFETs except that a sufficiently *negative* gate-source voltage is required to induce a conducting channel of holes. That is, to have current flow we must have,

$$V_G - V_S = V_{GS} < V_{Th} < 0$$

If $V_{Th} < 0$ it's called a p-channel *enhancement* MOSFET. The IV characteristics look qualitatively like the ones for n-channel FETs but with the signs of all the voltages and currents reversed. (The actual numbers on the plot will differ from transistor to transistor.)

The schematic symbol is shown in the circuit to the right. Once $V_{GS} < V_{Th} < 0$ holes flow through the induced channel from source to drain. Therefore $V_{SD} = V_S - V_D$ and I_{SD} are both positive. You can think of the little arrow on the schematic symbol like a pn junction, pointing *from* the p-type channel *to* the n-type substrate.





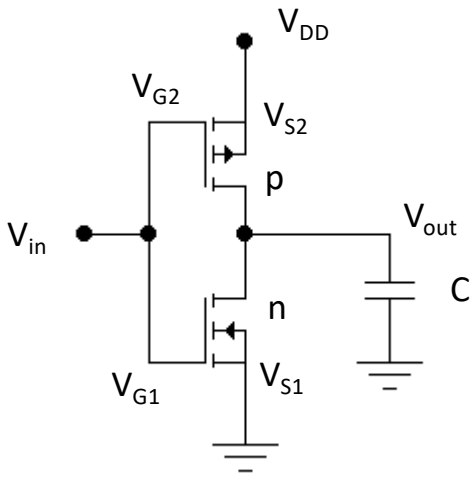
CMOS

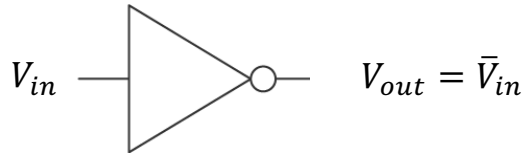
We're now ready to discuss what might be the most important circuit in modern electronics. It consists of an n-channel and a p-channel MOSFET as shown below. This setup is called CMOS since it relies on complementary MOSFETs. I've attached a capacitor to the output. That would typically be the capacitance to ground for one or more similar circuits to which the output is connected. There are two states of interest.

State 1: $V_{in} = 0$. Then $V_{G1} - V_{S1} = V_{GS1} = 0$ which is less than the threshold voltage V_{Th} for the n-channel MOSFET so it is turned **off.** $V_{in} = V_{G2} = 0$ so $V_{G2} - V_{S2} = V_{GS2} = -V_{DD}$. That's more negative than the threshold voltage for the p-channel MOSFET, so it's turned **on**. The p-channel FET becomes a low resistance channel between V_{DD} and the output so $V_{out} = V_{DD}$.

State 2: $V_{in} = V_{DD}$ so $V_{GS1} > V_{Th}$ and the n-channel FET is turned **on**. The p-channel FET now has $V_{G2S} = 0$. This is not negative enough to turn it on so it becomes an open circuit**.** The lower FET now provides a low resistance channel to ground and $V_{out} = 0$.

If we assign $V = V_{DD}$ = logic 1 and $V = 0$ = logic 0 then this circuit is a logical inverter or NOT gate.





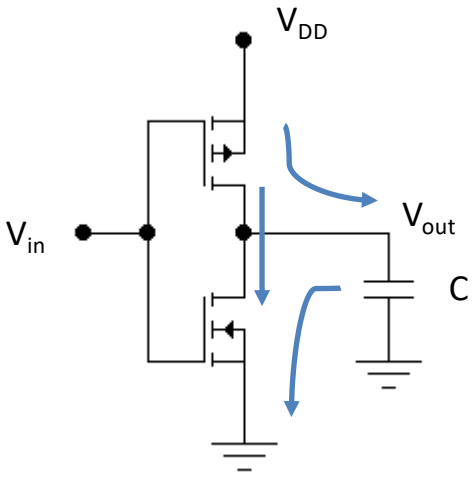
This circuit is a big improvement over our previous logic circuit that employed a resistor. Now, when the n-FET is on, the p-FET is off and vice versa. In real circuits the inputs and outputs change states. Suppose we're in state 1 so $V_{out} = V_{DD}$. To change states, the n-FET must turn on and the p-FET must turn off. During that time current flows though the n-FET to discharge the capacitance. After that, no current flows so there is no power dissipation. To change the output back to V_{DD} the n-FET must turn off and the the p-FET must turn back on. During that time current flows through the p-FET to charge the capacitor. Once that stops $V_{out} = V_{DD}$ and no more current flows so there is no dissipation. In both states final states the dissipation is zero. It's only *during* transitions that power is dissipated. That is the key advantage of CMOS.

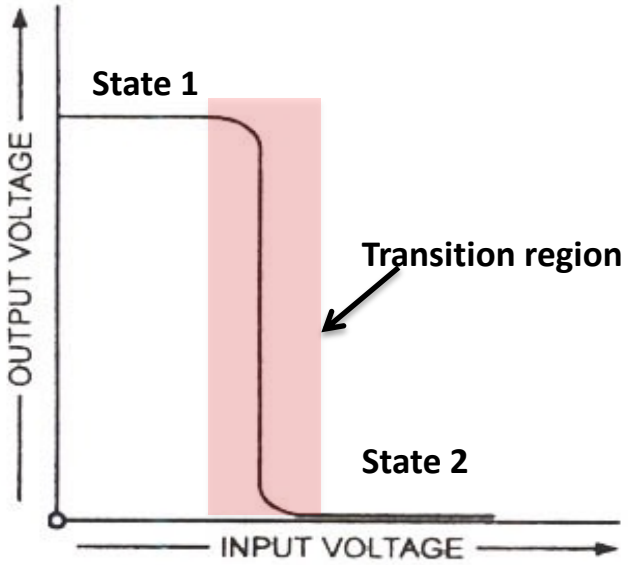
Power Dissipation

Ideally, in CMOS, current flows *only* during the transition from one state to the other so that's the only time power is dissipated. But as we run computers faster and faster these transitions come more often. Suppose you want to charge the load capacitance C up to voltage V_{DD}. A current must flow through the FET to do that and the instantaneous power dissipated is V_{DS}*I_{DS} . The total power dissipated in the transistor will be just $C\,V_{DD}^2/2$. Switching on and off f times per second leads to a "dynamic" power dissipation,

$$Power = V_{DD}^2 * f * C$$

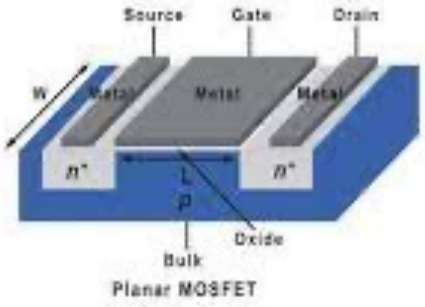
Reducing the switching speed *f* isn't an option in the modern world so the industry is moving to lower values of V_{DD}. That, however, has its limits because noise becomes a bigger fraction of the signal size and more errors accumulate. In addition, as transistors get down to the nanometer scale, it becomes harder to turn completely turn off the MOSFET. There is leakage current that flows directly through the two transistors regardless of the capacitance and this leads to additional power dissipation.



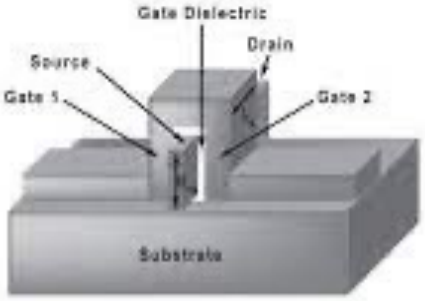


Modern Implementation: FinFETs

MOSFETs have have historically been fabricated in a *planar* geometry, as shown below for an N-channel device. However, as the size of the transistor shrinks, this approach leads to problems. As the channel gets shorter, it becomes more difficult to turn the FET fully off. This results in leakage current through the channel, more power dissipation and shorter battery life for mobile devices. In the semiconductor industry the direction has been toward a more 3-dimensional structure called a *FinFET.* In this device, the channel between source and drain is a "fin" that sticks out above the substrate. The gate then wraps around the channel on 3 sides.

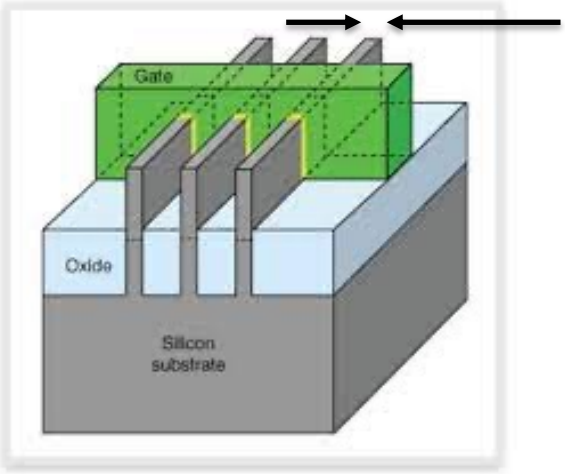


Planar MOSFET



FinFET

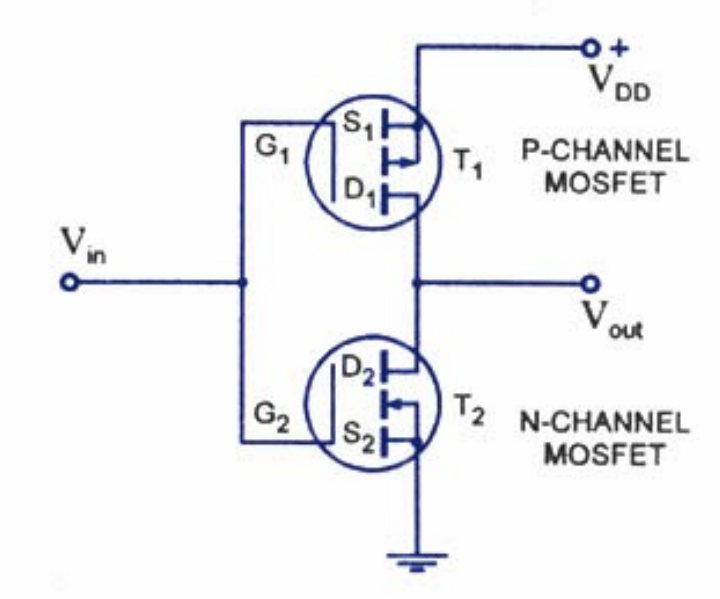
From mksinstr.com

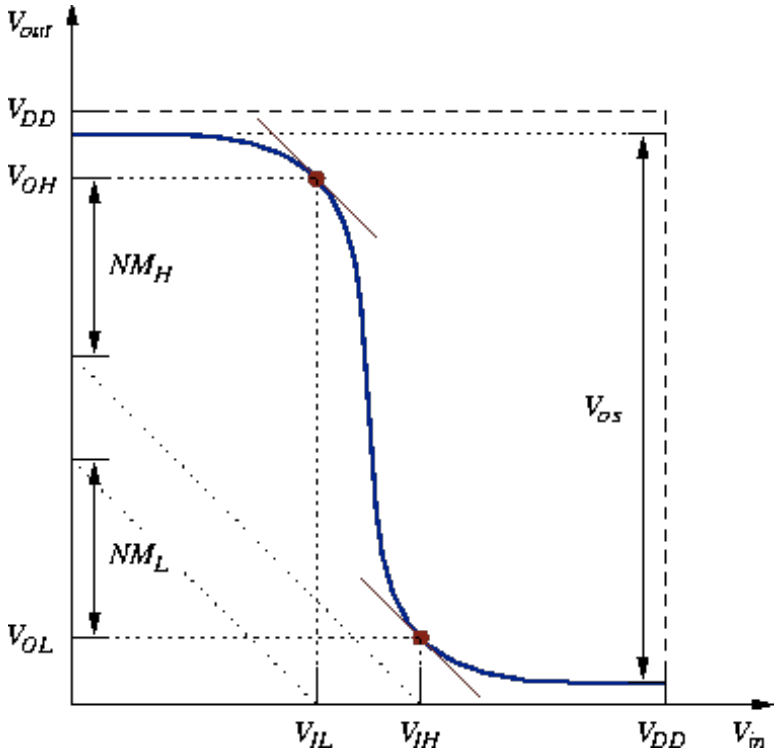


From synopsys.com

Even better performance is obtained by using many fins in parallel, as shown on the left. Multi-fin FETs provide lower source-drain leakage, lower power dissipation, lower threshold voltages and better speed than conventional planar MOSFETs. The logic gates are still made with complementary N and P type transistors, but the transistors themselves are made in this FinFET geometry. It's also possible to electrically isolate the gates on either side of the channel. With two gates on acting on each fin, it's possible to accurately control the threshold voltage. Such transistors would be called IG FinFETs meaning *independent-gate* FinFETs. When people talk about 22 or 15 or 7 nm technology, this refers to the minimum width of the channel. The current limit is around 3 nm.

Logic voltage limits





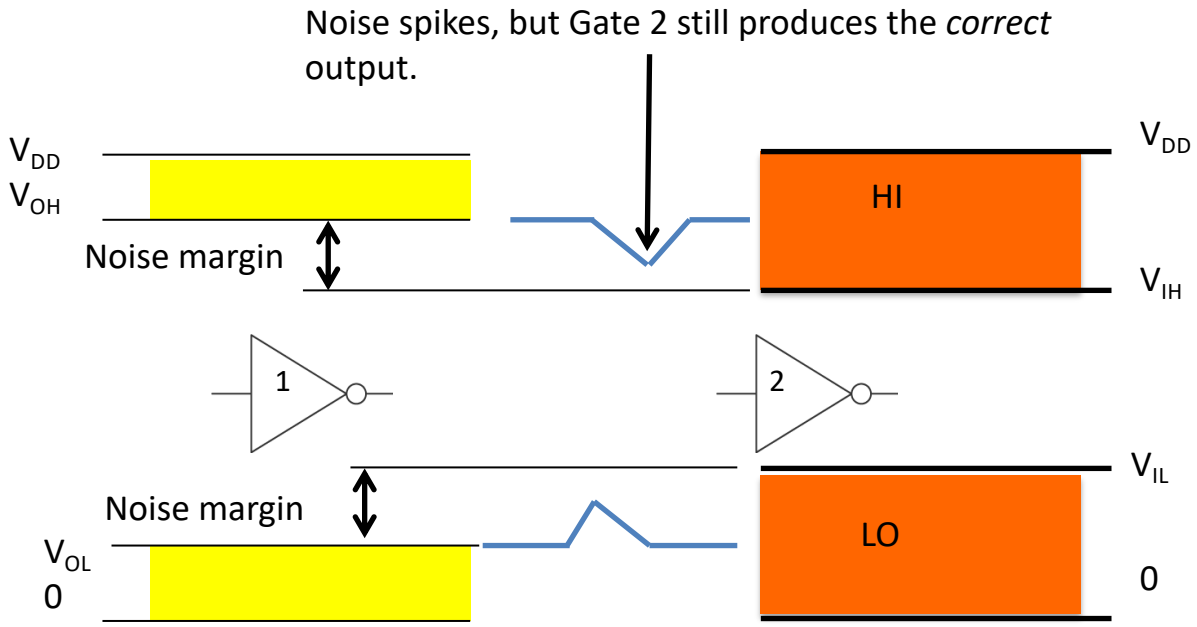
https://www.iue.tuwien.ac.at/htmlpapers/schrom-may98/node7.html

It's worth discussing the operation of the CMOS inverter in more detail to see how logic HI and LO limits are determined. The graph shows the output as the input is varied continuously from 0 up to V_{DD}. For $0 < V_{in} < V_{IL}$ the output stays between V_{DD} and V_{OH}. V_{OH} is the *minimum* value for which the output is considered logic HI. As V_{in} increases further the output drops sharply until $V_{out} = V_{OL}$ when $V_{in} = V_{IH}$. Outputs in this steep region between V_{OH} and V_{OL} are considered *indeterminate* – neither HI nor LO. As V_{in} increases beyond V_{IH} the output remains below V_{OL}. V_{OL} is the *maximum* the output can be and still be considered logical LO.

To reiterate, any input between 0 and V_{IL} will produce an output that is recognized as logic HI. Any input between V_{IH} and V_{DD} will produce an output considered logic LO. This is best seen on the next diagram.

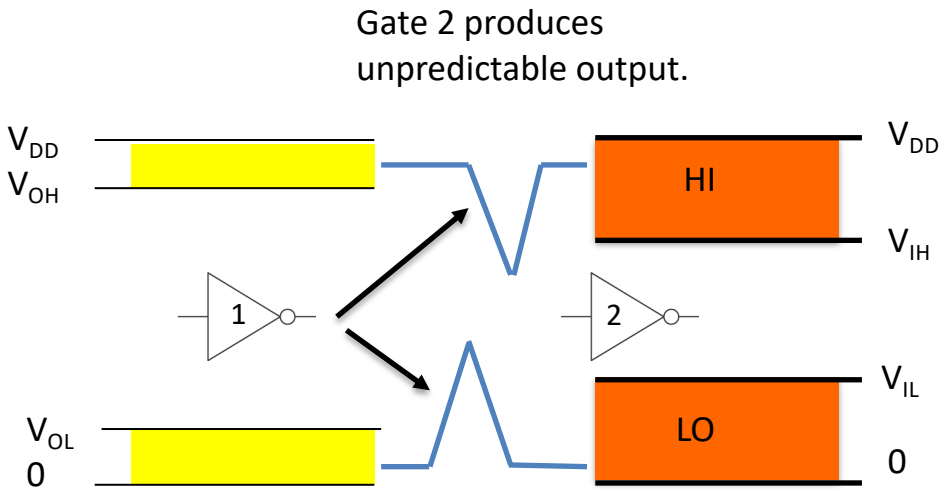
Noise Margins

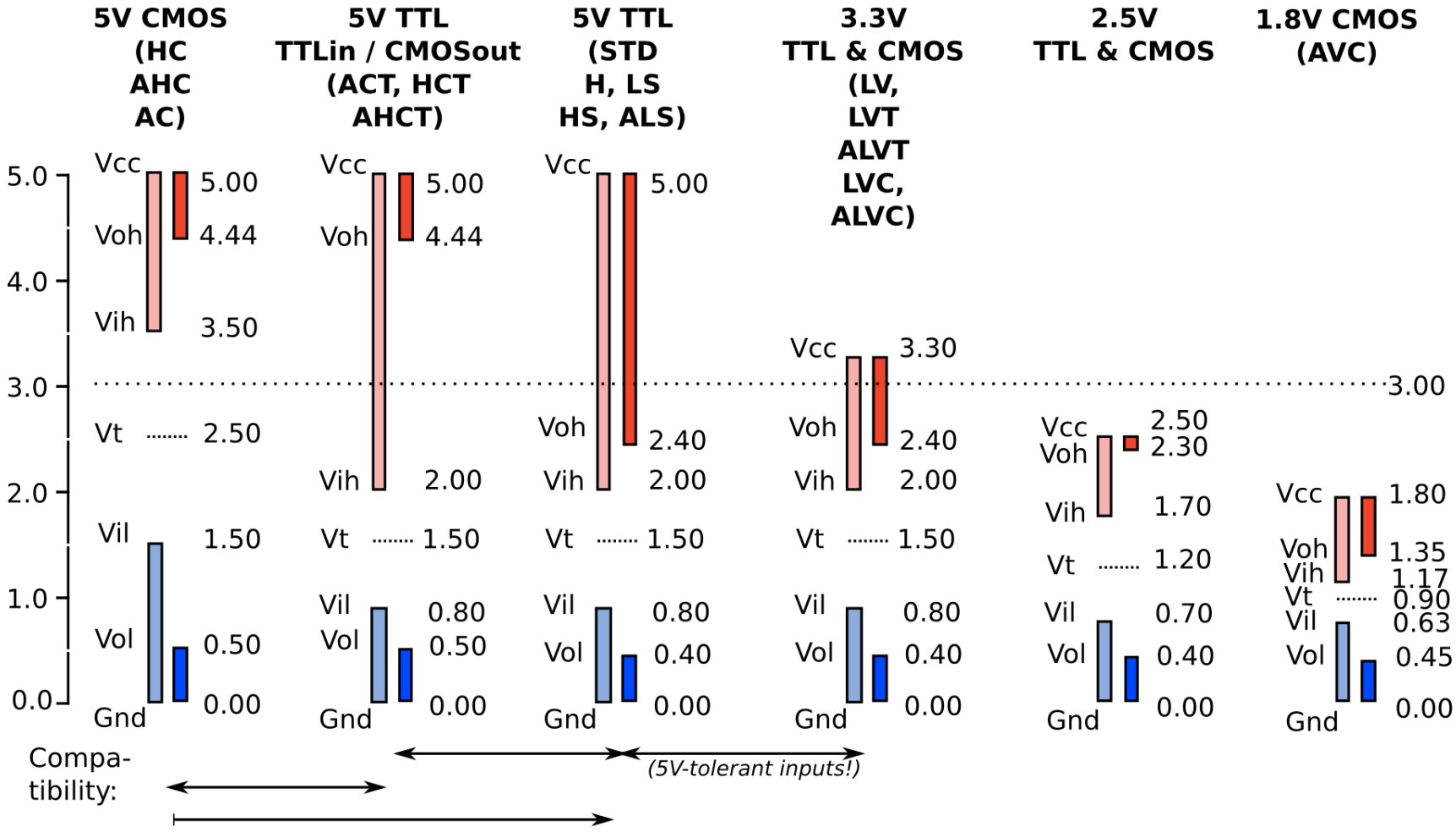
The yellow bands indicate the guaranteed range of output voltages for logic LO and logic HI. The orange bands show the range of input levels that will produce an acceptable logic LO or logic HI output. The yellow bands are narrower than the orange bands by an amount called the *noise margin*.



Suppose that noise on gate 1 results in an output spike. So long as Gate 2 receives an input signal larger than V_{IH}, it will correctly interpret it as a logic HI and produce the correct output. Or, suppose Gate 1 sends a spiky signal that is just below V_{IL}. Then Gate 2 will still interpret the signal as a valid input LO and produce the correct output. In each case, a noisy input to a digital logic gate still results in a correct logical output. This is known as *noise immunity*. In effect, digital logic allows us to regenerate a clean signal from a noisy signal. The figure shows that **larger noise margins give greater noise immunity.**

There are limits, of course. Suppose Gate 1 sends out a logic HI with a big noise spike. Gate 2 receives a signal that that falls into the range between V_{IN} and V_{IL}. We've exceeded the noise margin and the output from Gate 2 will be unpredictable. Or, maybe Gate 1 was sending out a logic LO and a noise spike drove the voltage above V_{IL}. Again, the output of Gate 2 will be unpredictable.



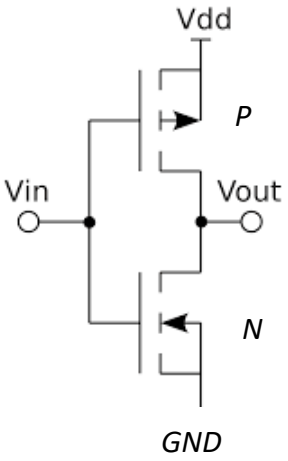


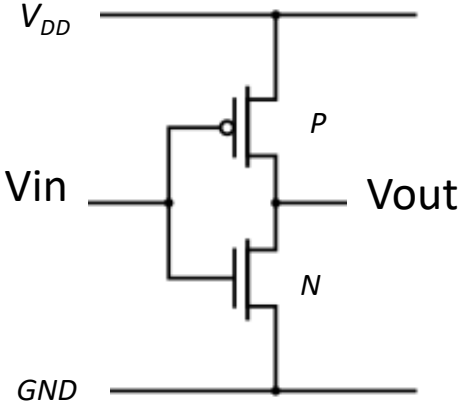
Data source: EETimes, A brief recap of popular logic standards (Mark Pearson, Maxim).

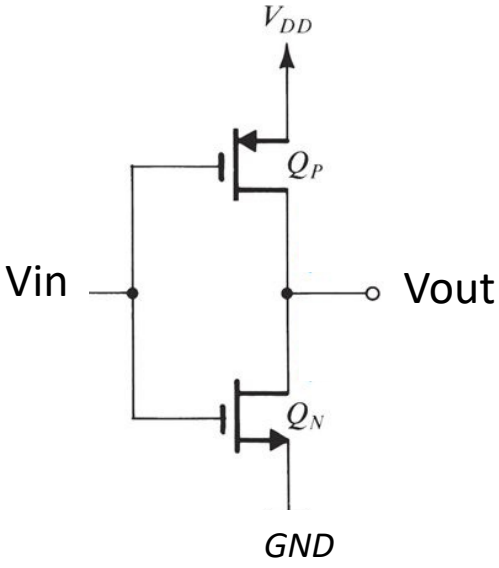
The chart shows input and output levels for several logic families. For 5 Volt CMOS the noise margins are $V_{OH} - V_{IH} \approx V_{IL} - V_{OL} \approx 1$ Volt. 3.3 V CMOS is now very common for sensors. It has noise margins of 0.4 V. Older transistor-transistor logic (5V TTL) also has a noise margin of 0.4 V so it has much lower noise immunity than 5V CMOS.

Other schematic symbols

You'll see several different symbols for MOSFETs. The CMOS NOT gate is shown on the left. Those schematic symbols tells you the most about the transistor type (p-channel or n-channel, enhancement mode, source connected to substrate) but you'll often see the other two versions. In each case the lower transistor is an n-channel enhancement MOSFET and the upper one is a p-channel enhancement MOSFET. In the middle figure the small circle on the gate indicates a p-channel MOSFET. Since the circle is the schematic symbol for logical negation, an input voltage of $V_{IN} = V_{dd}$ (logical 1) would turn the top transistor *off* and a LO input voltage ($V_{IN} = 0$) would turn it *on*. The right-hand circuit uses a notation that indicates the direction of current through the channel and distinguishes the source from the drain.

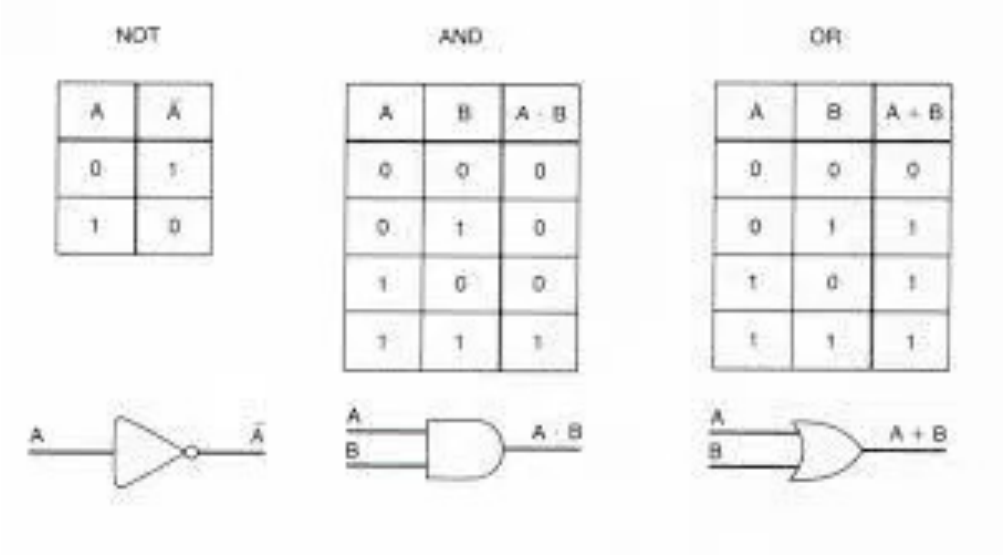






Logic Gates

In digital electronics we deal with signal levels that represent logical True and False or alternatively 1 and 0. A +5 V signal might represent 1 and 0 V represents 0. When the high voltage corresponds to logical 1 we call it *positive logic.* When the high voltage corresponds to logical 0 we call it *negative* logic. Both are perfectly acceptable ways of doing things. Truth tables for the three basic Boolean logic operations (NOT, AND and OR) are shown below along with the corresponding logical operator symbol. We'll use the symbol $\bar{A}$ for NOT A , A+B for (A OR B) and AB for (A AND B).



https://gyandakids.wordpress.com/2016/01/14/q-basic-logical-operators/

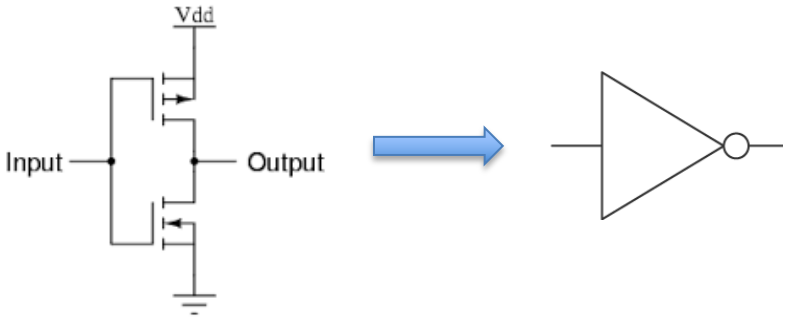
Summary of Boolean logic operations

A + 0 = A A0 = 0 A + B = B + A AB = BA A + 1 = 1 A1 = A A + A = A AA = A A + $\bar{A}$ = 1

Distributive property A(B+C) = AB + AC Associative Property (AB)C = A(BC)

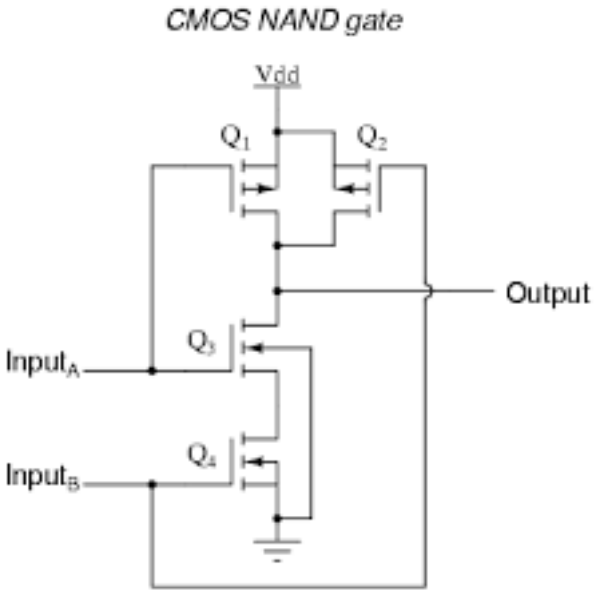
DeMorgan's Laws: $\overline{ABC\ldots} = \bar{A} + \bar{B} + \bar{C} + \ldots$ $\overline{A + B + C + \cdots} = \bar{A}\,\bar{B}\,\bar{C}\ldots$

Implementation of logical operations

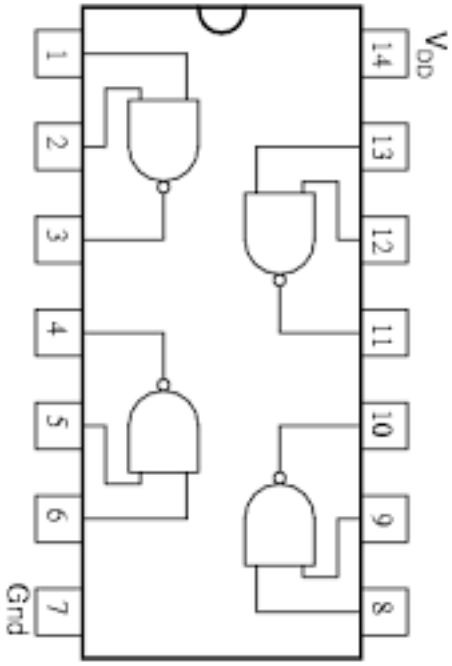


In digital electronics, logic gates perform the logical operations on incoming signals. We've already seen the simplest logic gate - the CMOS inverter. In the figure, logical 0 = logic LO = 0 V and logical 1 = logic HI = V_{DD} . (CMOS is extremely versatile and in fact the voltage levels could in principle be +/- 5 V or anywhere in a range that doesn't damage the transistors.)

One very common gate is the NAND = $\overline{AB}$. Using DeMorgan's laws, you can build any logical operation with enough NAND gates. Its CMOS implementation is shown below. To how it works, suppose that A and B are at logical 1, i.e., voltage = V_{DD} . Then Q_1 and Q_2 are off while Q_3 and Q_4 are on. The output has a low resistance path to ground through Q_3 and Q_4 so the output is at logical 0. In this way it's easy to show that the circuit has the truth table of a NAND. For very simple digital electronics you can buy chips with 4 NAND gates inside, all powered by an external voltage V_{DD}.



Input		Output
A	B	Y= $\overline{A.B}$
0	0	1
0	1	1
1	0	1
1	1	0

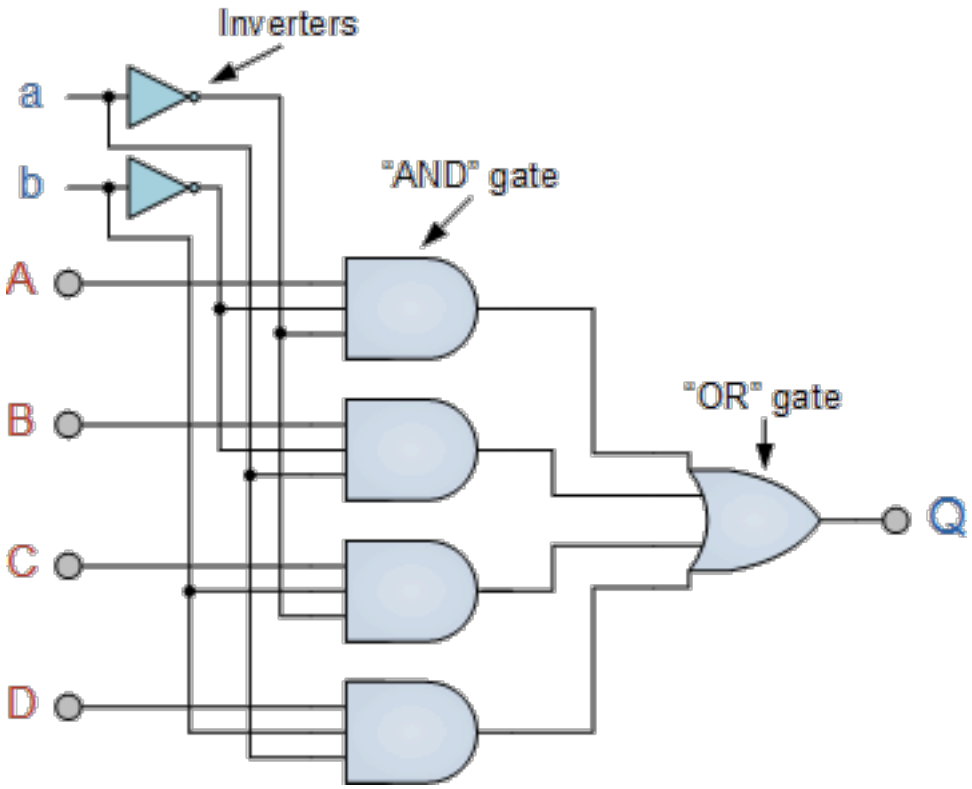


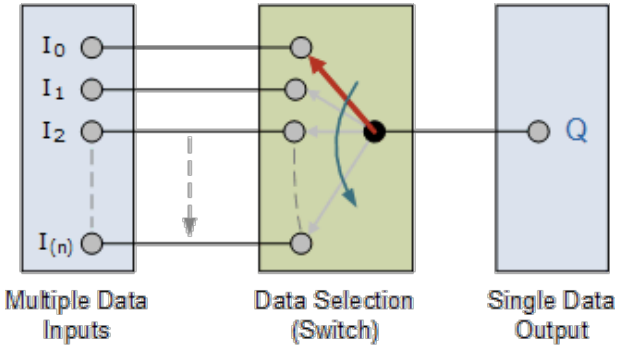
Combinatorial logic

Logic circuits in which the output is a unique function of the input(s) are called *combinatorial* logic. There are many kinds but once you see a few you'll get the idea.

Multiplexer

Suppose you want to send digital data from 1 of 4 different places A, B, C or D to an output Q. This device shown below does that. It's called a 4:1 *multiplexer*. It's like a mechanical switch that can move between 4 different positions.





Figures from http://www.electronics-tutorials.ws/combination/comb_2.html

Decoder

A 2-bit decoder is shown below. It has 4 possible outputs lines: D_0, D_1, D_2 or D_3. Only one of them goes to logic 1 depending on the two address bits It takes two *address* bits A_0 and A_1 . The truth table is shown along with *miniterms* which are Boolean algebraic expressions for the outputs. With 2^N address lines we can light up one of 2^N possible outputs.

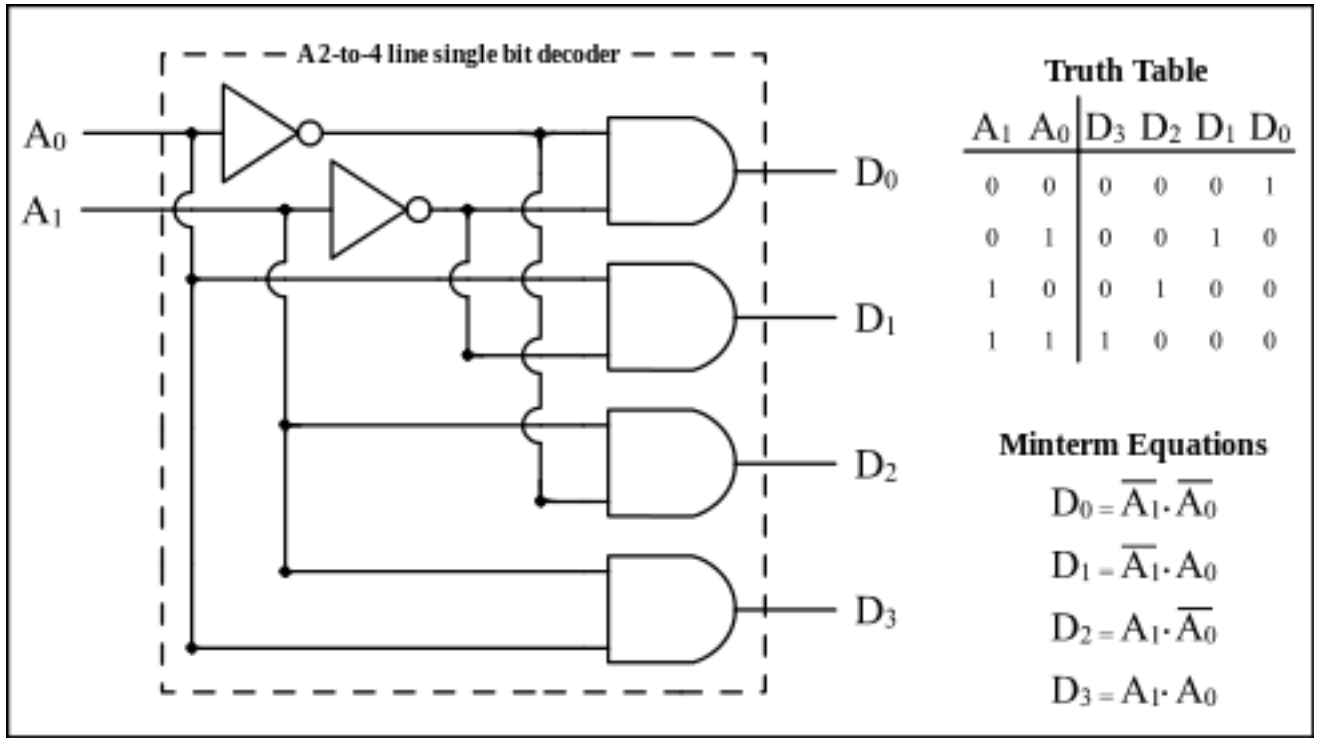
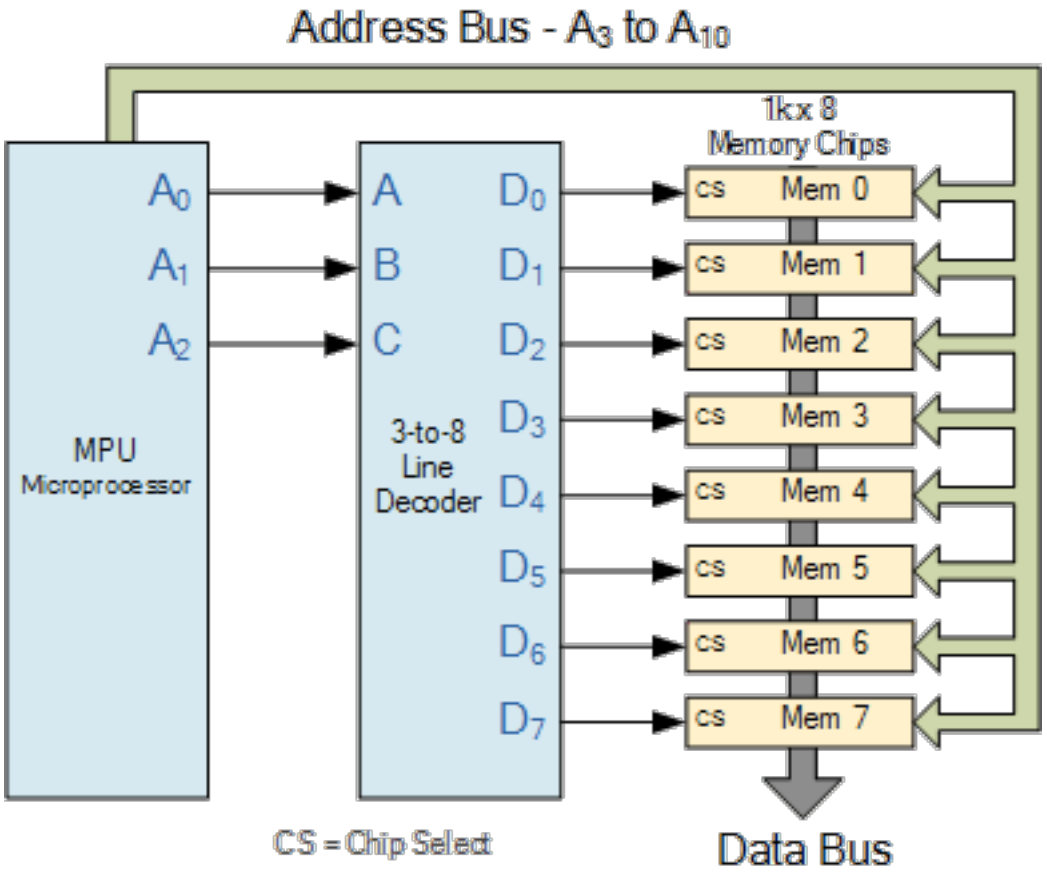


Figure from https://en.wikipedia.org/wiki/Binary_decoder

The figure below illustrates how decoders are used. Suppose we want to store some information in a memory location. Assume for now that a memory location is a transistor circuit that can store a string of 1's and 0's. Each memory location has a unique *address*. The microprocessor (MPU) might want to store the information 10110101 in address 11110110 on Memory chip 3. Think of each memory chip as a Zip code and each memory location on the chip as someone's address within that Zip code.



The MPU has address outputs $A_0 - A_{10}$. These are connected to a set of wires called the Address Bus.

1. The MPU sets $A_2A_1A_0$ to 011.

2. The decoder "decodes" 011 and sets D_3 to 1, which in turn activates the Mem 3 chip. It is ready to accept data.

3. The MPU also sets $A_{10} - A_3$ to 11110110, the location on Mem 3 where the byte of information is to be stored.

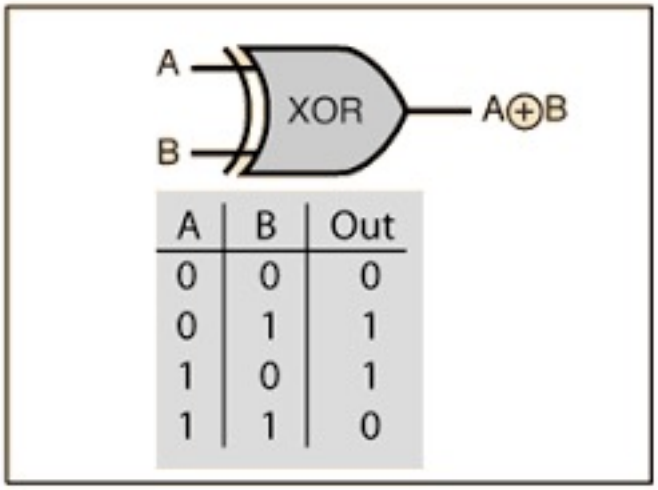
4. The actual data, 10110101, is now sent on the Data Bus which is another set of parallel wires connected to logic outputs on the MPU.

http://www.electronics-tutorials.ws/combination/comb_5.html

XOR

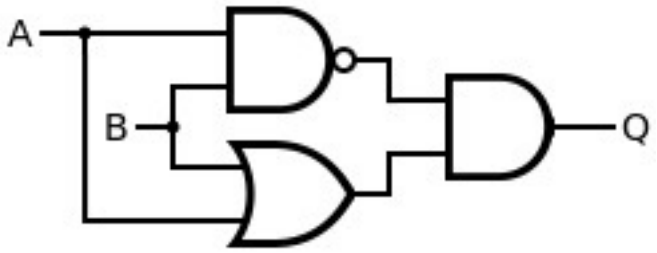
The XOR is a useful combinatorial function whose schematic symbol and truth table are shown. Its output is logic 1 if and only if one or the other but *not* both inputs are 1. The circuit shown below is one way to make an XOR. The Boolean expression for the circuit is,

$$A \oplus B = \overline{AB}(A + B)$$



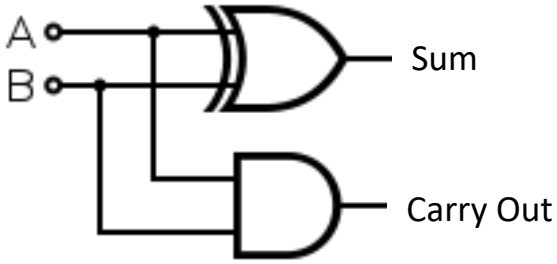
A	B	Out
0	0	0
0	1	1
1	0	1
1	1	0

This set of logic gates will do the trick,



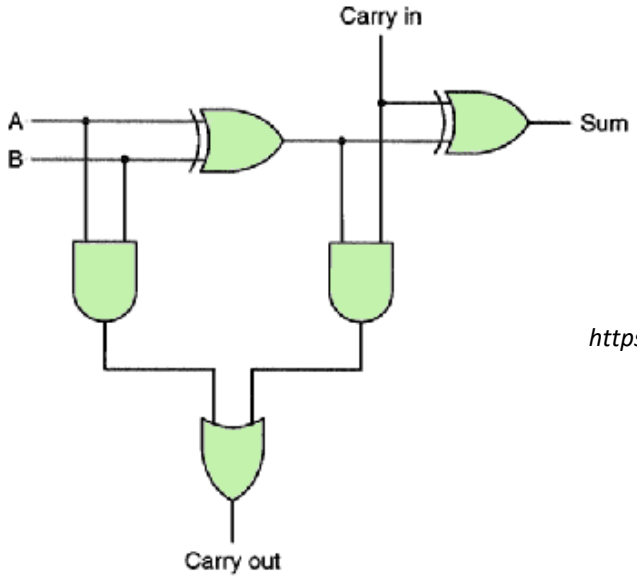
http://hyperphysics.phy-astr.gsu.edu/hbase/Electronic/xor.html

The XOR is an essential component of a *half-adder*, shown below. The circuit takes two bits, A and B, adds them and produces a sum (S) and carry (C) bit. It's not a *full* adder because that requires the circuit to also add the carry bit from a previous addition. To do binary addition the least significant bits (LSB) can be summed with a half adder. Higher order bits in red each require a full adder, shown next.




```
  10101
+ 01101
_________
 100010
```

A	B	Carry In	Sum	Carry out
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1



Here is a *full adder* which gives you some appreciation for how many transistors are required to do a simple task like adding two bits.

https://www.quora.com/What-are-the-uses-of-a-full-adder

XOR Encoding

Since the internet came along we've all become worried about keeping our digital information private. XOR is one way to encode data. Consider the byte *B* = 10101 that we want to encode. This is called the *Plaintext*. The idea is to XOR *B*, bit by bit, with a *key*, which is a string of bits that is known only to those allowed to access to the original information. The encrypted version of *A* is called the *Ciphertext*. To decrypt the Ciphertext, just XOR it with the key again.

10101 A
01101 key

encrypted A 11000 $A \oplus key$

11000 encrypted A
01101 key

10101 $(A \oplus key) \oplus key = A$

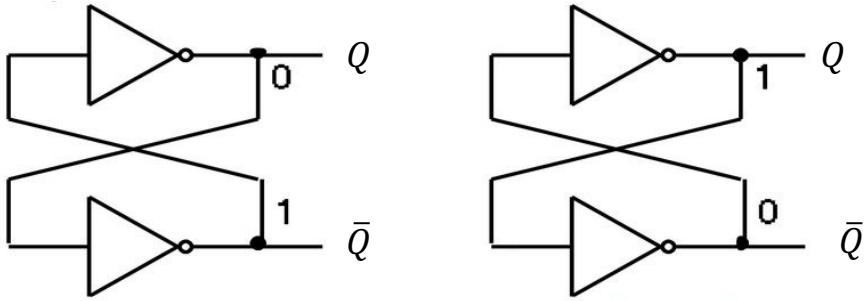
Okay, but what's special about XOR? Why not use AND or NOR? Compare the truth tables when we use XOR as opposed to AND. Suppose the key for a given bit is *randomly* generated; 0 or 1 with equal probability. Suppose the original information bit (the *Plaintext*) = 0. Then the encrypted bit (the *Ciphertext*) has equal probability of being 0 or 1. The same holds true if the Plaintext = 1. Therefore, a probability analysis of the Ciphertext can't tell you anything about the Plaintext. On the other hand, if you chose to encrypt with the AND operation, then a Plaintext = 0 is *always* encrypted into a 0 and a probability analysis of the encrypted message can eventually break the key. If you generate the key with a random number generator then XOR encryption is apparently unbreakable. You can read more about this in the Wikipedia article called *"XOR Cipher"*. Of course, if someone happens to have copy of both the Plaintext and the Cyphertext then the key is immediately obtained by the operation: $Key = Plaintext \oplus Cyphertext$.

Plaintext	Key	Plaintext $\oplus$ Key = Ciphertext
0	0	0
0	1	1
1	0	1
1	1	0

Plaintext	Key	Plaintext $\cdot$ Key = Ciphertext
0	0	0
0	1	0
1	0	0
1	1	1

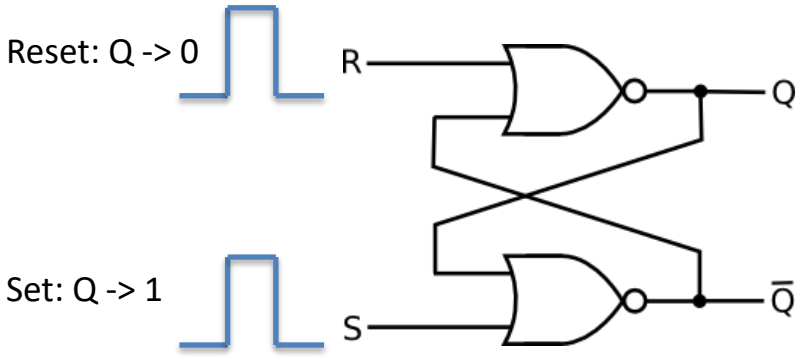
Memory Circuits

We've presented just a few example of combinatorial logic circuits. These are a significant part of digital electronics but not the whole story. Since the outputs are unique functions of the inputs, there is no memory! To see how to achieve memory consider the simple cross-coupled pair of inverters,



This circuit has <u>two</u> stable output states: Q = 1 and Q = 0. It we put the system into either state it will stay that way indefinitely, so long as the power is on. This is a memory circuit. As it stands there is no obvious way to change the state since there is no input. We can remedy that by using a pair of NOR gates instead. As its truth table shows, with one input of a NOR gate held at 0, it becomes an inverter with respect to the other input. This new circuit shown below is called a *flip-flop*. The inputs are traditionally called S (Set) and R (Reset.) With S and R both 0 it's just like the circuit above. But now we can change the state.

Input		Output
A	B	$\overline{A+B}$
0	0	1
0	1	0
1	0	0
1	1	0

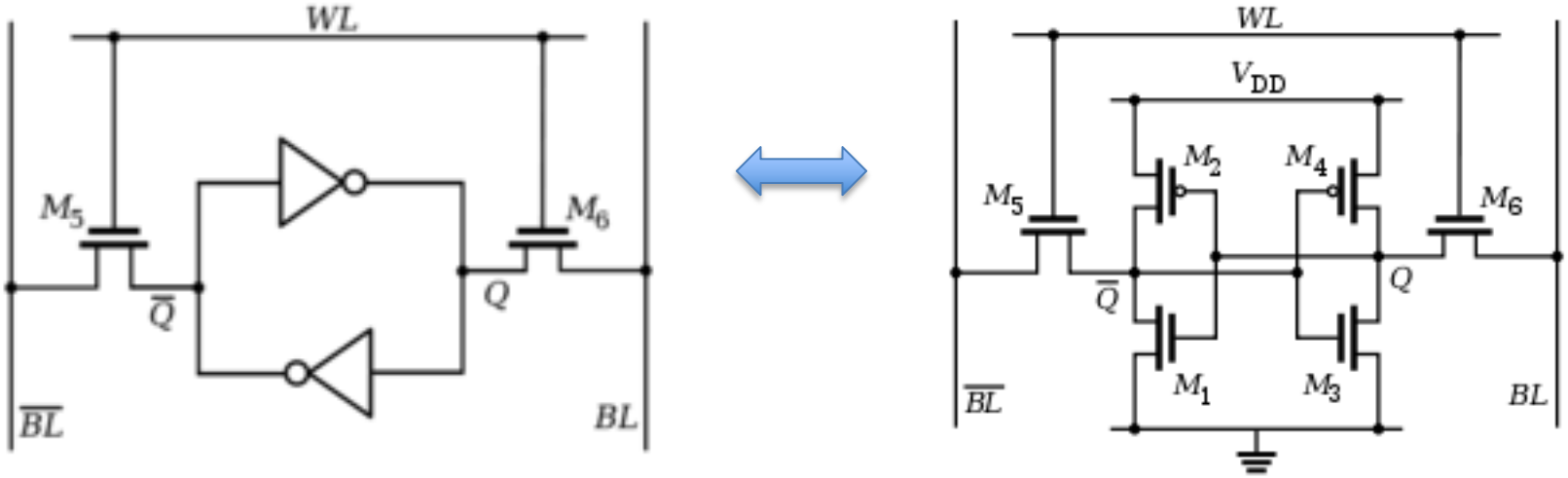


S	R	Q	State
0	0	Previous State	No change
0	1	0	Reset
1	0	1	Set
1	1	?	Forbidden

To make Q = 1, make S = 1. Once S goes to 1 and Q goes to 1 then you can bring S back to 0 and Q remains at 1. In electronics we say you've "latched" a 1 into the flip flop and for that reason circuit is also called an SR latch. To make Q = 0, then send the R input to 1. Again, you can return it to 0 and Q will remain at 0. The "truth table" for the latch shows that the state with *both* S and R = 1 is to be avoided because the final output depends on which one returns to zero first and that can be unpredictable. There are many different kinds of flip-flops but all have one element in common – feedback from the output to the input.

Static Ram

The cross-coupled inverter is a basic storage element in what's called Static Random Access Memory (SRAM) in which we need Gigabytes of memory cells that can be written to and read from in a few nanoseconds. The circuit shown is a single cell of SRAM. In addition to the cross-coupled inverters there are two MOSFETS, M5 and M6 that provide access to the cell. When the Word Line (WL) is at 0 these FETs are turned off and the inverters are isolated. When the word line is driven to logical 1 they are turned on and provide a conducting path into the memory cell. The actual data (1 or 0) that's either written into the cell or read from it is carried on the bit line (*BL*) and its complement $\overline{BL}$. In SRAM we need to both store data but also read it out without changing the state of the memory cell.

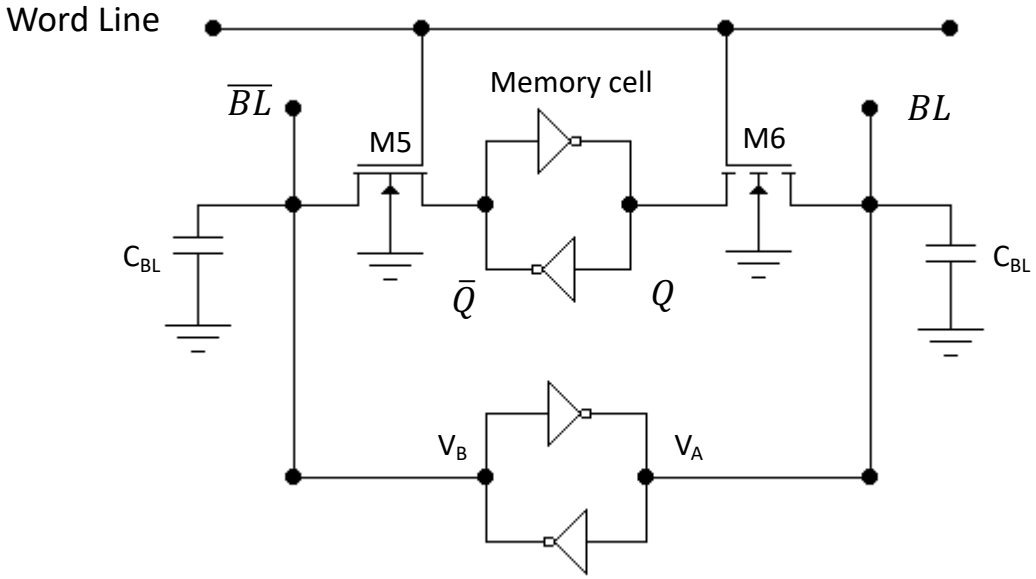


https://en.wikipedia.org/wiki/Static_random-access_memory

Writing data: To store Q = 1 set *BL* = 1 and $\overline{BL}$ = 0. Then drive WL to 1 which turns on M_5 and M_6 . There is now a conducting channel from the bit lines to the memory cell. A logical 1 on *BL* forces Q = 1 and $\overline{Q}$ = 0. This is a stable state so once the *WL* line goes to 0, the memory cell is isolated and Q = 1 is stored. To store Q = 0 drive *BL* = 0 and $\overline{BL}$ = 1. In principle you would need only the *BL* line and M6 for writing data. However, reading it out is more problematic as explained next.

Reading Data

In an ideal world, reading the data out of each SRAM memory cell would amount to driving the Word Line to 1 which would connect the memory cell to BL and then reading the voltage on BL. Then drive the WL line back to 0 to isolate the cell. Unfortunately the BL line has capacitance C_{BL} which, together with resistance in circuit, slows down the process. And for other types of memory cells, like dynamic RAM, this would lead to a smaller voltage on the BL line and bit errors. What is needed is a way to (1) read out the data unambiguously and (2) restore the full voltage that was in the cell before we read it out. It's a little like opening the refrigerator, stealing someone's beer and replacing it with an identical can.



Sense amplifier

Suppose Q = 1 corresponding to voltage V_{DD}. Some external circuitry now sets both BL and $\overline{BL}$ to $V_{DD}/2$. With $V_A = V_B = V_{DD}/2$ the sense amplifier is in a metastable state. A tiny imbalance between V_A and V_B will drive it to one of its stable states, either $V_A = V_{DD}$ and $V_B = 0$ or vice versa. Now open gates M5 and M6. The voltage on BL will increase to $V_{DD}/2 + \Delta V$ and the voltage on $\overline{BL}$ will decrease to approximately $V_{DD}/2 - \Delta V$. The sense amplifier sees this imbalance and jumps to $V_A = V_{DD}$ and $V_B = 0$. That, in turn, drives BL and Q back to V_{DD}. We can latch this information into a flip-flop connected to BL and at the same time we have restored Q back to its original state V_{DD} so all is well. We can now close the refrigerator, i.e., turn off M5 and M6 and the data in the memory cell has been restored.

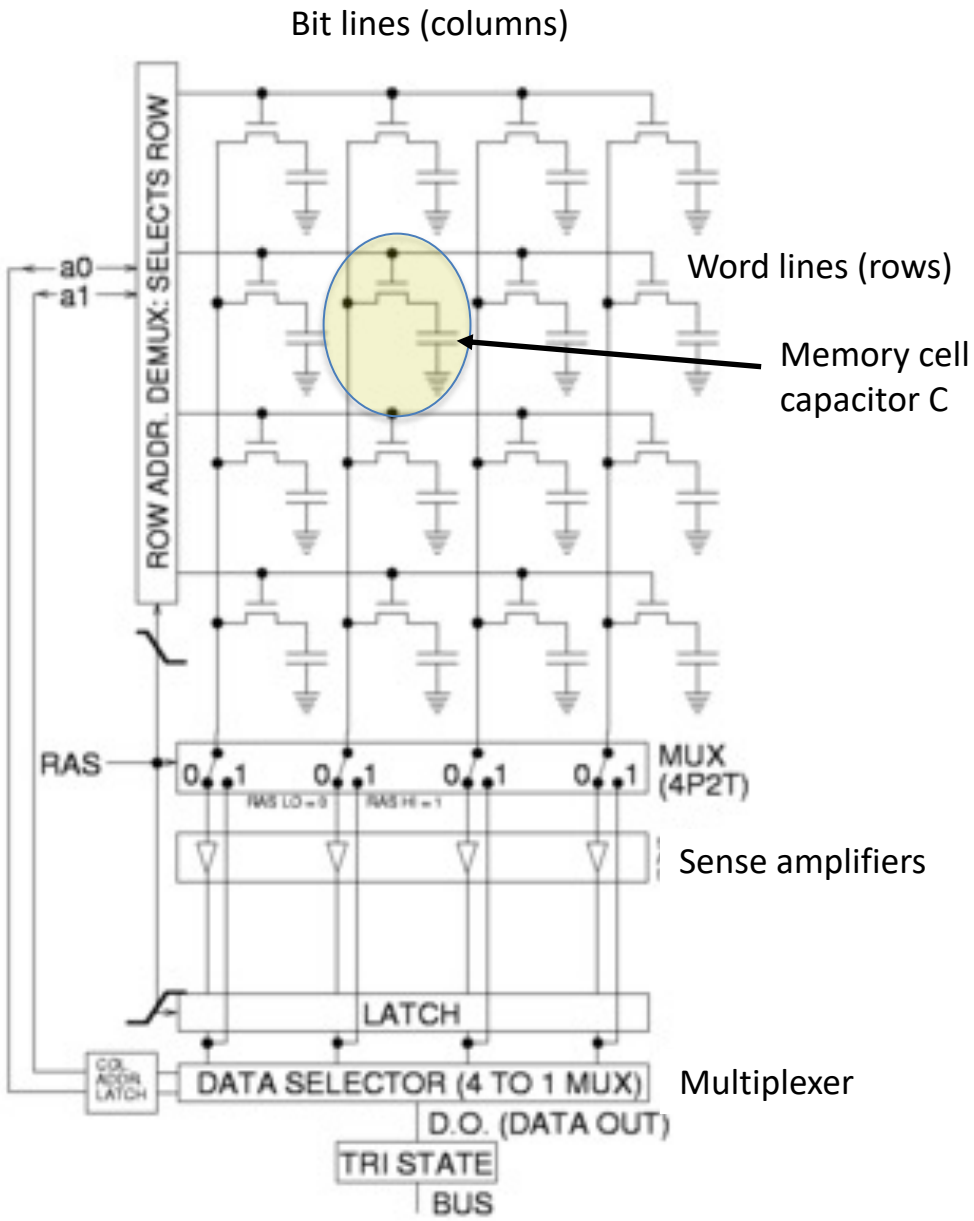
The sense amplifier speeds up the whole process and essentially regenerates the original information that was stored in the memory cell. One sense amplifier is used for all the memory cells connected to BL and $\overline{BL}$ but each cell still requires 6 transistors so there is an incentive to do it all with fewer transistors, which leads us to DRAM.

Dynamic RAM (DRAM)

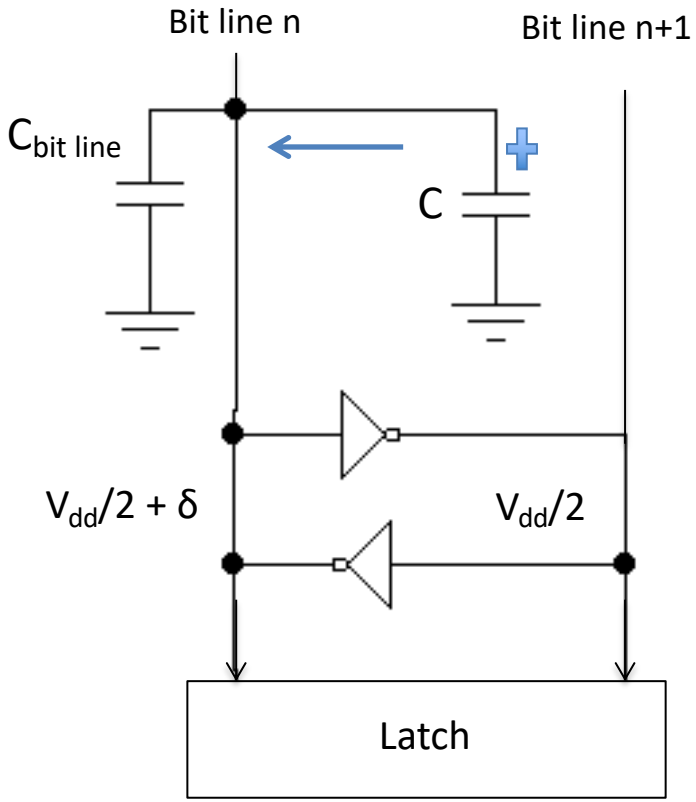
To make random access memory with fewer transistors DRAM (dynamic random access memory) was invented at IBM in 1966. DRAM is the scratch pad memory on which a computer temporarily stores data. Each bit of information is stored on a capacitor whose charge can be accessed through the channel of a MOSFET, which acts like a switch. Since each bit requires just one transistor, as compared to 6 for 6T SRAM, the density of DRAM is much higher. Each capacitor is small to keep the density high.

The packing density in DRAM is higher than SRAM but the data is more fragile. The small capacitor holding each data bit discharges through the leakage resistance of the MOSFET with an RC time constant of order 1 -10 sec. To avoid losing data, the charge on the capacitor must be constantly *refreshed*, typically every 64 msec, making the circuitry more complicated than for SRAM.

With DRAM, again, the extra capacitance of the bit line presents a problem. Suppose the memory cell capacitor C is charged up to V_{DD}. To read it, drive the word line to 1. That turns on the MOSFET and connects the capacitor to the bit line. But now the charge originally on C is *shared* between the C and the capacitance of the bit line. This lowers the voltage from V_{DD} making the reading more unreliable. And when you're done reading and turn off the MOSFET switch, C is left with less charge than before. To remedy this, a similar sense amplifier scheme as described for SRAM is used to read out and regenerate the data stored in C. The difference is that it's done every 64 msec or so to prevent the charge on C from leaking away.



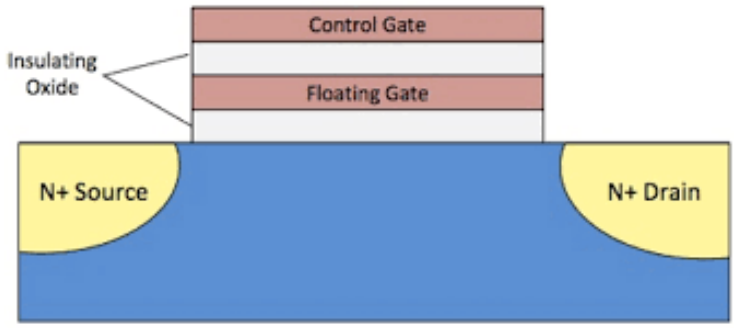
https://en.wikipedia.org/wiki/Dynamic_random-access_memory



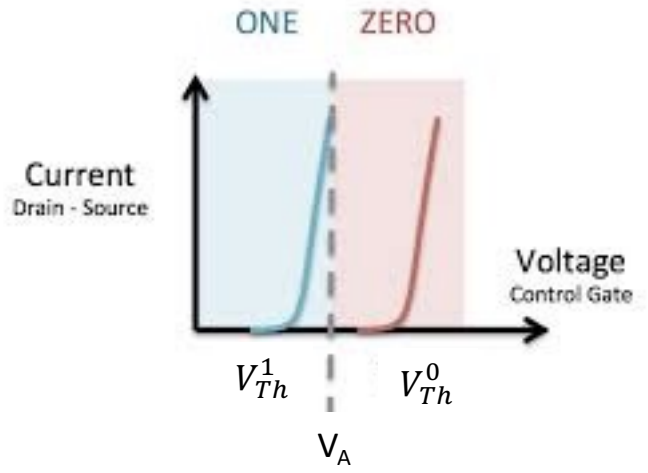
Read out and restore memory capacitor voltage: Suppose C is charged up to V_{dd}. Prior to turning on the MOSFET, Bit lines n and n+1 are driven to $V_{dd}/2$. Now the Word Line turns on the FET and C is connected to $C_{bitline}$. Charge flows as shown driving bit line n to a voltage slightly *above* $V_{dd}/2$. This drives the upper NOT gate to its stable state with an output of 0, which sends Bit line n+1 to 0. That in turn drives the lower NOT gate and Bit line n to V_{dd}. This *recharges* C to voltage V_{DD}. The data can be latched if needed and we've regenerated the original data on C. The word line can now be driven to 0 leaving C isolated.

Flash Memory

SRAM and DRAM lose data once the power goes off so they are *volatile* memory. Flash memory is nonvolatile. The information stays there even with the power off. The key element is a floating gate MOSFET, shown below. The floating gate is a piece of conductor sandwiched between the channel and the regular control gate. It is insulated from everything else and can store charge more or less indefinitely.



https://www.tegile.com/flash-storage-introduction/

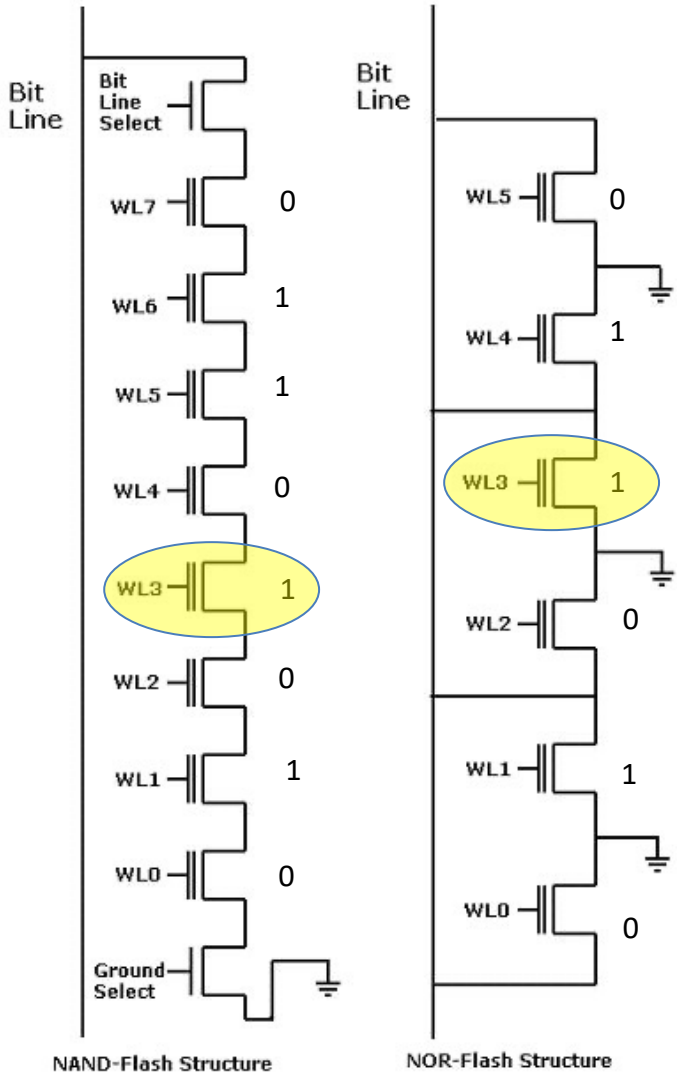


https://flashdba.com/2015/01/09/understanding-flash-floating-gates-and-wear/

To load electrons onto the floating gate a large voltage (> 10 V) is applied to the control gate. This voltage is much larger than the logic HI level which might be 3 V or so. Electrons tunnel from the channel to the floating gate. To get them off, a large negative voltage is applied to the control gate and the electrons tunnel to the channel.

Suppose there are electrons on the floating gate. They create an additional electric field that competes with the one generated by the control gate. As a result, the threshold voltage for the MOSFET is raised up to V_T^0 as shown in the figure. On the other hand, when the floating gate is uncharged, the threshold voltage = V_T^1. <u>The floating gate charge changes the IV characteristics of the FET.</u>

If we now apply $V_{GS} = V_A$ to the control gate then the MOSFET will conduct current if there is no charge on the floating gate. If there *is* charge on the floating gate then the FET will *not* conduct current and is effectively an open circuit. In effect the data is stored as charge on the floating gate and it's accessed by seeing whether or not the control gate voltage can induce a conducting channel.



https://www.eeherald.com/section/design-guide/esmod16.html

NAND and NOR flash memory

The arrangement shown below has floating gate FETs arranged in what is called NOR flash memory. The transistors are labeled 0 or 1. Imagine that the Bit Line is normally at logic HI voltage.

NOR structure: Each source and drain is individually connected to ground and the Bit Line respectively. Suppose you wish to know if cell 3 is 1 or 0. If it's a logic 1 then the floating gate is uncharged. Therefore, if you drive the gate (Wordline WL3) to V_A, that's *above* the threshold V_T^1 so the FET will conduct. It provides a short circuit between the Bit Line and ground, pulling the Bit line to 0. On the other hand, if the floating gate on cell 3 is charged (logic 0) then driving the gate to V_A *won't* turn on the FET and the Bit line remains at logic 1. Looking at each pair of FETs, the Bit Line will go LO if one or the other or both is HI. That's the truth table of a NOR so it's called NOR memory. It was introduced by Intel in 1988.

NAND structure: In this structure the source of one transistor is connected to the drain of its neighbor. The ordinary FETS labelled Bit Line Select and Ground Select allow us to isolate the string of floating gate FETs during operations on other transistors. To read cell 3, it's necessary to turn all the other FETs on, *regardless* of their state. That means we need to drive all but WL3 to a voltage above V_T^0 . Now apply V_A to WL3. Again, if it has charge on its floating gate then it won't conduct and the Bit Line will remain HI. If it doesn't have charge on its floating gate then it will turn on and provide a short circuit between the Bit Line and ground. The only way to get a conducting channel between the Bit Line and ground is if all the FETs conduct. That's like the truth table of a multiple input NAND gate so it's called a NAND structure. It was introduced by Toshiba in 1989.

There are fewer connections in NAND so it is cheaper, faster and denser than NOR. Your thumb drives use NAND. Each cell can be addressed in NOR so it is used in microcontrollers and to replace older ROM chips.