Introduction to MOSFET Operation

ECE/CS 5720/6720

Analog Integrated Circuit Design

Prof. Reid R. Harrison

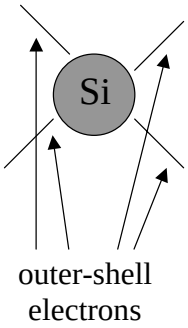
Department of Electrical and Computer Engineering

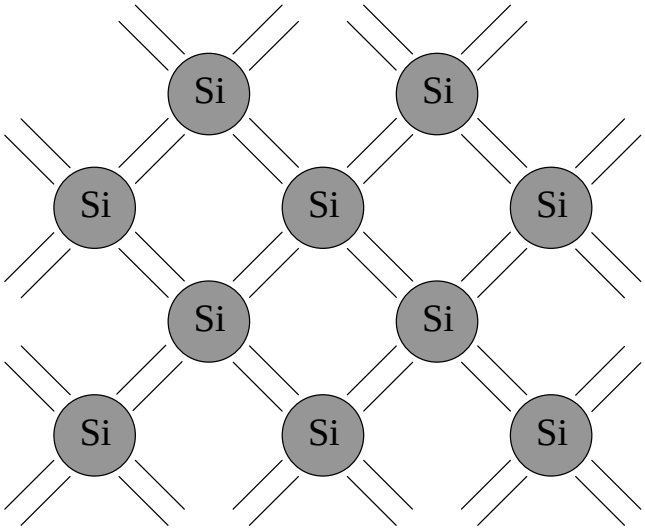
University of Utah

Semiconductors – A Review

Most VLSI circuits today are based on silicon (a semiconductor) and silicon dioxide (SiO_2, a near-perfect insulator).

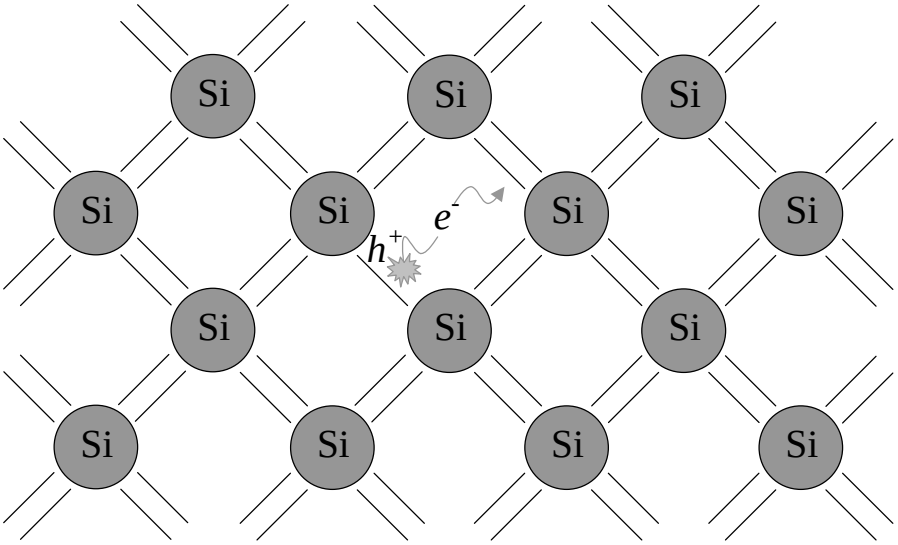
A silicon atom has four electrons in its outer shell. Since this shell has an 8-electron capacity, silicon atoms can fill their outer shell by forming bonds with four adjacent atoms.





When silicon forms a crystal lattice, all the silicon atoms are "happy" because they each have full outer shells (8 electrons). However, there are no free charge carriers, so no current can flow! All electrons are bound to nuclei.

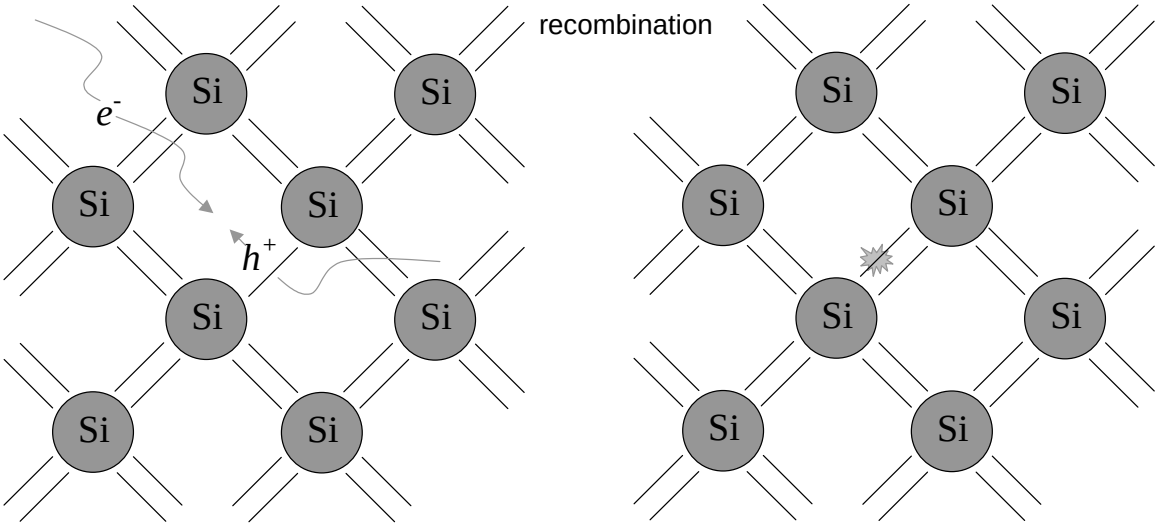
Then why is silicon a semiconductor and not an insulator? Remember, at temperatures above absolute zero, atoms are being vibrated and shaken about. Occasionally, these atomic-scale vibrations (thermal energy) will shake an electron out of the crystal lattice. (An incident photon can also cause this to happen.) This is called **generation**, since a pair of charge carriers are generated.



In the absence of an electric field, this electron moves through the lattice in a "random walk," bumped around by thermal energy. We call this electron a "carrier" since it is now free to move and thus carry current.

The "hole" left by escaped electron has a positive charge since there is one less electron than the surrounding protons in the silicon nuclei. It is very easy for a nearby electron to shift over into this hole. Of course, this just moves the hole. Thus, the hole is mobile (just like an air bubble in water) and acts as a positively-charged carrier.

So carriers are generated in pairs: one electron and one hole. If a hole and electron meet, the electron "falls into the hole", filling it. The hole is gone because it has been filled without creating another hole, and the free electron is no longer a carrier because it is now reattached to the lattice. This electron-hole annihilation is called **recombination.**

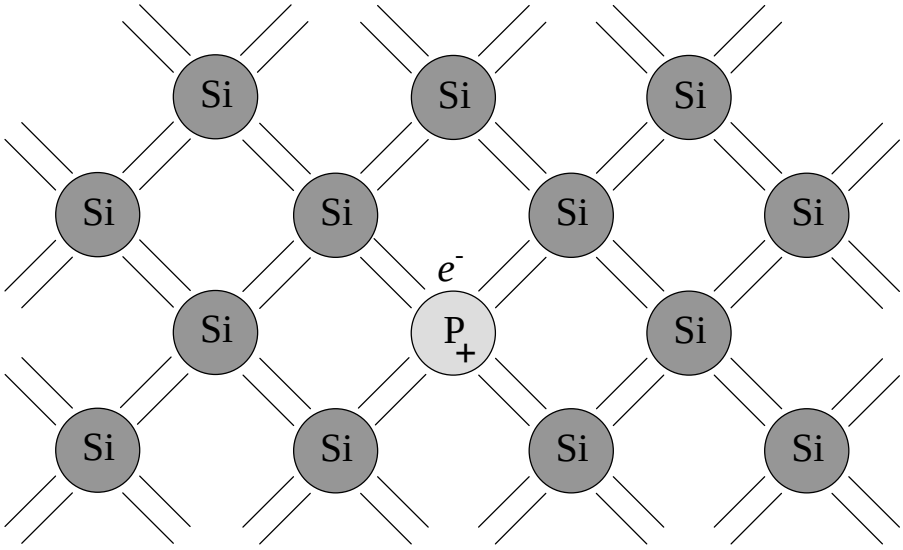


At room temperature (300°K), a pure ("intrinsic") silicon crystal has around 1.5×10^{10} "free" electrons per cubic centimeter, and an identical number of holes. This is why silicon is a semiconductor – the electrons in its crystal lattice can be shaken free fairly easily.

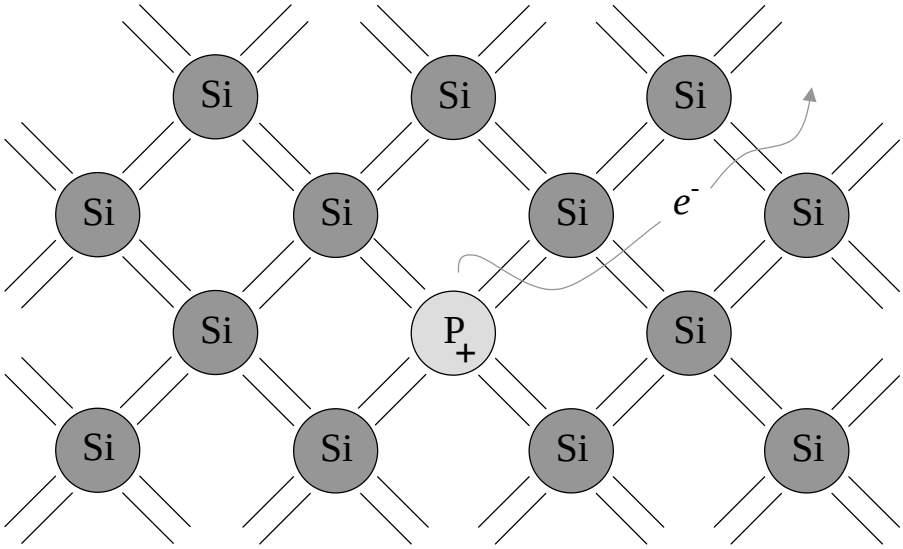
This may sound like a lot of electrons running around, but consider this: There are 5×10^{22} silicon atoms per cubic centimeter. This means that only one out of every 3 trillion atoms has an missing electron at any point in time!

So far, all we have is a resistor whose resistance decreases as we raise the temperature. Things get more interesting if we **dope** a silicon crystal – add impurities.

Suppose we add a few phosphorous (P) atoms to our silicon crystal. Phosphorous atoms are similar in size to silicon atoms, but they have **five** outer-shell electrons. This extra electron doesn't fit neatly into the lattice, and it is very easily knocked around by thermal energy. Phosphorous is a **donor,** since it "donates" extra electrons to the crystal.

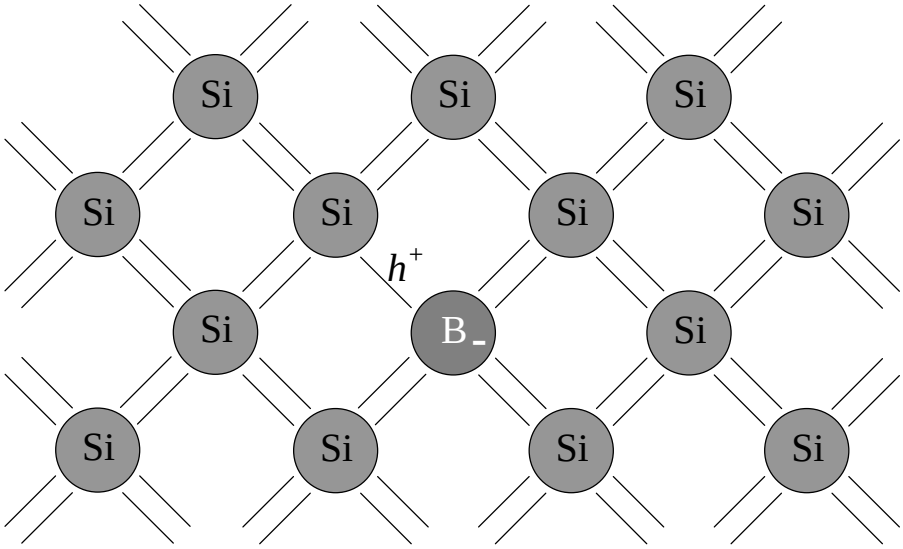


If this electron moves away from the donor atom, there will be a net positive charge on the phosphorous nucleus, since it needs five electrons in its outer shell to maintain charge neutrality. Unlike a hole, however, this positive charge due to the phosphorous nucleus is fixed to the lattice and cannot move.



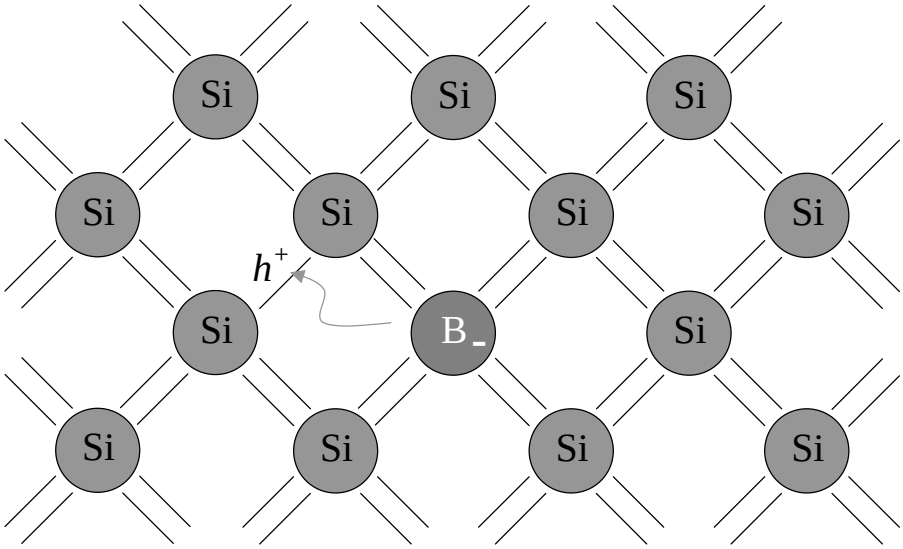
Silicon doped with donor atoms in known as **n-type silicon** because of the added **n**egative charge carriers. Typical doping levels are 10^{15} – 10^{20} dopants per cubic centimeter (one dopant atom per 10^2-10^7 silicon atoms), which greatly overwhelms the intrinsic thermally-generated carriers ($1.5 \times 10^{10} cm^{-3}$).

What if we dope silicon with an atom having three electrons in its outer shell, such as boron (B)?



Boron doesn't fit cleanly into the silicon lattice either; it has one too few electrons, so there is a hole next to the boron nucleus. Boron is called an **acceptor** because it accepts electrons (into this hole). Equivalently, it donates extra holes to the crystal.

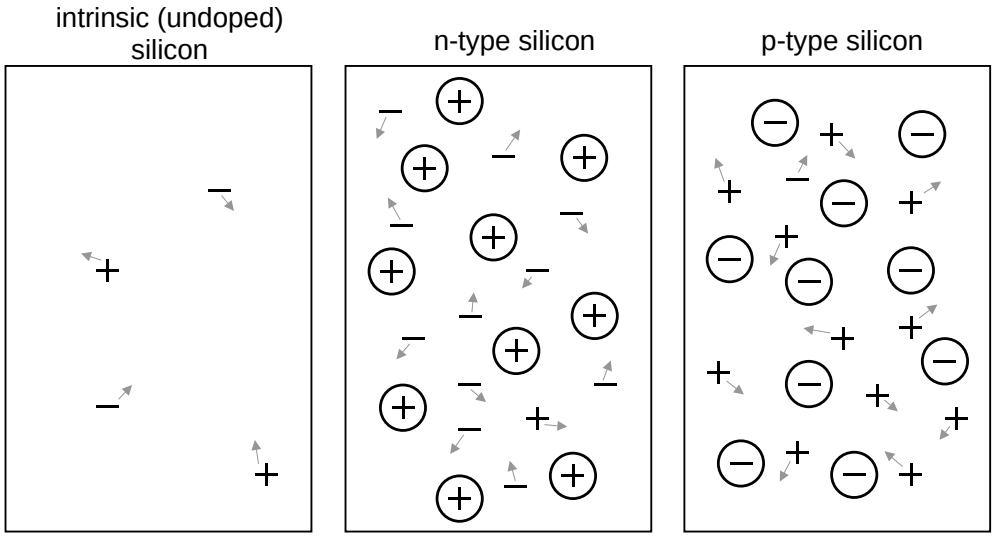
If this hole moves away from the acceptor atom, there will be a net negative charge on the boron nucleus, since it needs three electrons in its outer shell to maintain charge neutrality and now has four. Unlike an electron, however, this negative charge due to the boron nucleus is fixed to the lattice and cannot move.



Silicon doped with acceptor atoms in known as **p-type silicon** because of the added **p**ositive charge carriers (i.e., holes). As in n-type silicon, typical doping levels greatly overwhelm the intrinsic thermally-generated carriers.

Now let's look at a different way of graphically representing doped (extrinsic) silicon:

- We will stop drawing all the silicon atoms and lattice bonds, because these objects have no net charge and cannot move.
- We will draw mobile electrons and holes as plus and minus signs (+ and -)
- We will draw donors atoms as a plus sign with a circle around it, and acceptor atoms as a minus sign with a circle around it. The circle reminds us that these charges **cannot move**.



Notice that each piece of silicon is electrostatically **neutral** on the macroscopic level; there are equal numbers of positive and negative charges.

Also notice that even the n-type silicon has a very small concentration of holes due to thermal generation. The same goes for electrons in the p-type silicon.

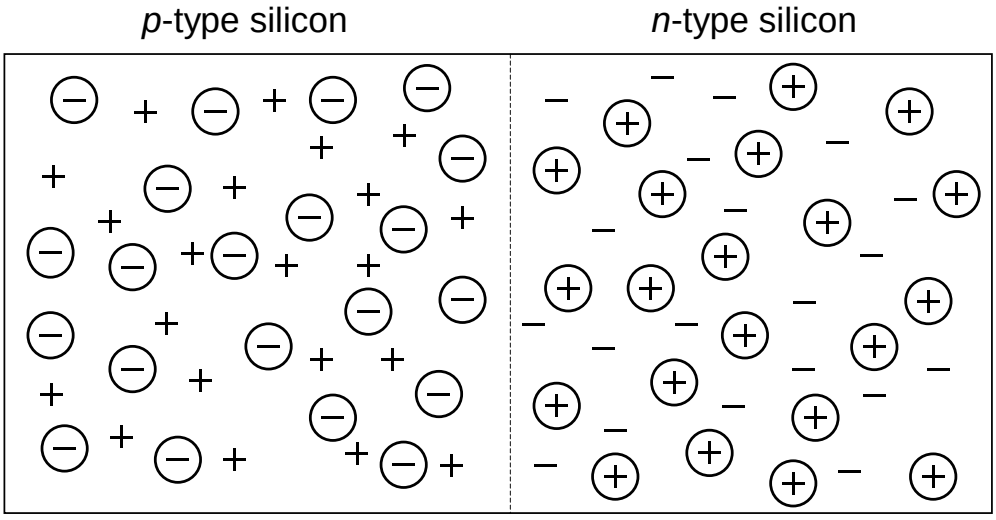
What happens if we take a piece of p-type silicon and attach it to a piece of n-type silicon?

p-n Junctions

An n-type piece of silicon has many free negative charges (electrons) and an equal number of fixed positive charges due to the dopant atoms. There will also be a tiny number of free electrons and holes generated thermally. The positive and negative charges balance each other out.

Conversely, p-type silicon has many free positive charges (holes) and an equal number of fixed negative charges.

What happens if we take a piece of p-type silicon and attach it to a piece of n-type silicon, forming a ***p-n* junction**?



Both pieces of silicon were (and still are) electrostatically neutral, so there is no electric field to pull electrons or holes one way or the other. The carriers are moving around randomly with thermal energy. So what happens?

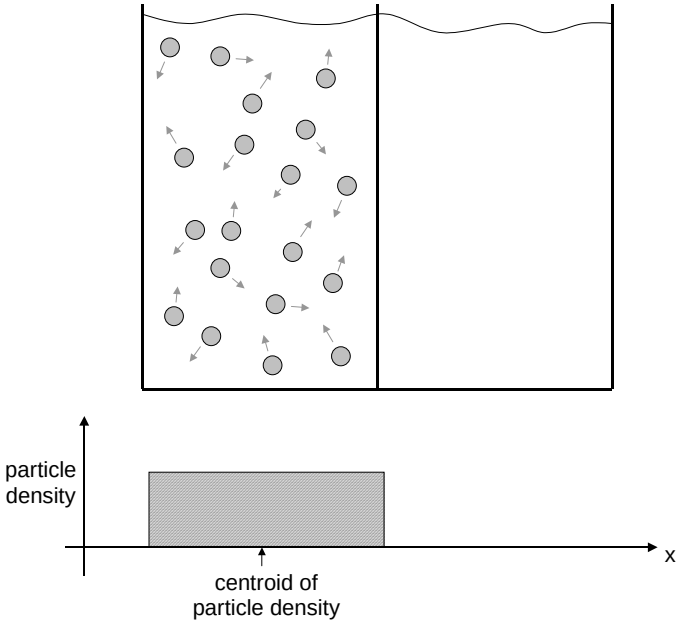
There are two mechanisms for carrier (electron or hole) motion: **drift** and **diffusion**.

- **Drift** requires an **electric field.**
- **Diffusion** requires a **concentration gradient.**

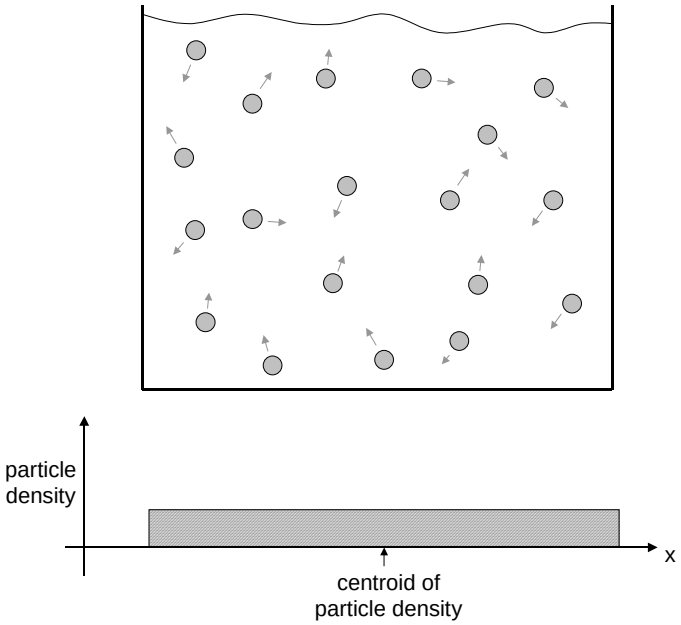
Diffusion

What are the mechanisms of carrier (electron or hole) motion in the absence of an electric field? Don't they just move randomly in an unpredictable manner? Random: yes. Unpredictable: no.

Consider the divided tank shown below. It is filled with water on both side, and particles are mixed into one side.



Now we remove the barrier…



Observation: Although each particle moves in an unpredictable fashion, as a group their behavior is predictable. The particles move from an unlikely arrangement (all the particles on one side of the tank) to a much more likely arrangement (fairly evenly distributed). This is just the second law of thermodynamics – entropy must increase!

The Diffusion Equation in 1-D

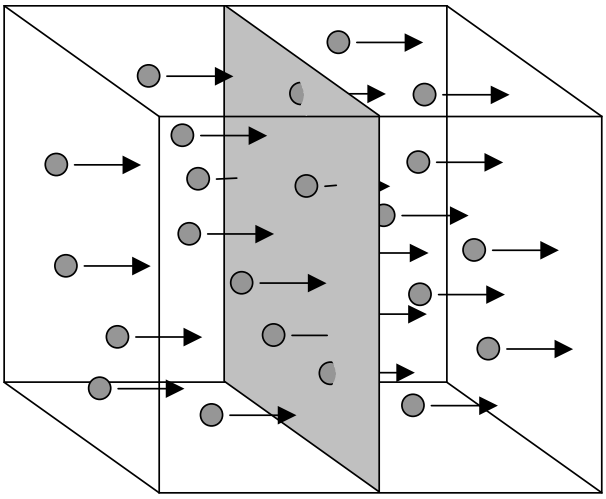
On average, particles move away from areas of high concentration and into areas of low concentration.

If we express particle concentration (density) at point x as $c(x)$, then the particle flux (particles per second per unit area passing point x) is given by:

$$\phi(x) = -D\frac{dc(x)}{dx}$$

where the constant D is called the **diffusion coefficient.**

The concept of particle flux is illustrated below. If we count the number of particles passing through the shaded region each second and divide by the area of the region, this gives us flux. The unit for flux is $cm^{-2}s^{-1}$ (particles per second per unit area).



The electron flux (electrons per area per second passing point x) $\phi_n(x)$ is given by:

$$\phi_{n-diffusion}(x) = -D_n \frac{dn(x)}{dx}$$

where $n(x)$ is the electron density at position x and D_n is the **diffusion coefficient** for electrons. The quantity $dn(x')/dx$ represents the **electron concentration gradient** in the x direction at point x'.

The hole flux is given by:

$$\phi_{p-diffusion}(x) = -D_p \frac{dp(x)}{dx}$$

where $p(x)$ is the hole density at position x and D_p is the diffusion coefficient for holes. The quantity $dp(x')/dx$ represents the **hole concentration gradient** in the x direction at point x'.

Since $p(x)$ and $n(x)$ have units of cm^{-3} and flux has units of $cm^{-2}s^{-1}$, the unit of the diffusion coefficient is cm^2/s.

Drift

In the presence of an electric field, charged particles exhibit **drift**. If a particle with charge q is placed in an electric field E, it experiences a force:

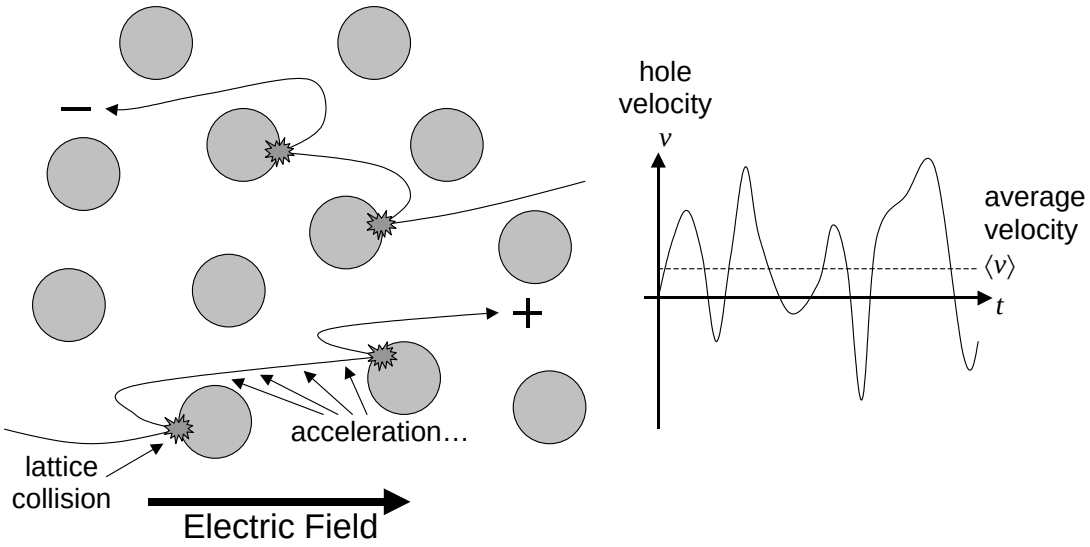
$$F = qE$$

Using Newton's second law ($F = ma$), we find that the particle undergoes an acceleration:

$$a = \frac{q}{m} E$$

So a charged particle in a constant electric field should accelerate continuously…

Inside the crystal lattice of silicon, however, the story is different. Accelerating particles are always colliding with the lattice and loosing energy. The situation is similar to the microscopic collisions that result in mechanical friction. When we apply a constant, horizontal force to a block of wood on a table, it doesn't accelerate without limit, even though Newton's law predicts this will happen. Rather, it quickly reaches a constant velocity where the frictional forces balance our applied force. In other words, when friction (millions of microscopic collisions) is involved, a constant force produces a constant **velocity**, roughly (no pun intended).



This is certainly the case for charged particles in a solid piece of silicon. While each electron is bouncing around, colliding with the lattice, it is observed experimentally that electrons travel at an **average velocity** proportional to the electric field:

$$\langle v_x \rangle = -\mu_n E_x$$

Here, E_x represents the electric field in the x direction, and $\langle v_x \rangle$ is the electron's average velocity in the x direction. The constant of proportionality μ_n is called the **electron mobility,** and is obtained empirically.

The average velocity of holes is observed to be:

$$\langle v_x \rangle = \mu_p E_x$$

Notice that two things are different here from the previous equation. First, the sign has reversed, indicating that holes have a positive charge and thus travel in the opposite direction of negatively-charged electrons. Second, there is a different constant of proportionality, μ_p, called the **hole mobility.**

Of course, this tells us that the mobility of electrons and holes is different. If it was the same, we wouldn't bother with the subscripts. As you might expect, free electrons squeezing through the crystal lattice are more mobile than holes, which move when a bond switches from one atom to another.

In silicon, electron mobility is typically three times larger than hole mobility.

Mobility is a function of doping density. Ionized dopant atoms are very effective at interfering with the path of a drifting carrier. Mobility decreases with increasing dopant concentration (in a rather complex way).

Mobility is also a function of temperature. The lattice offers more resistance to drifting carriers when it has more thermal energy. Mobility drops with increasing temperature. (It is observed that mobility is proportional to $T^{-3/2}$ at room temperature and higher.)

Since electric field strength has units of V/cm and velocity has units of cm/s, the unit for mobility is $cm^2/V{\cdot}s$.

Electron and hole flux due to drift can be written as:

$$\phi_{n-drift}(x) = -\mu_n n(x) E(x)$$

$$\phi_{p-drift}(x) = \mu_p p(x) E(x)$$

where $n(x)$ and $p(x)$ are the densities of electrons and holes, respectively, at point x.

The Einstein relation

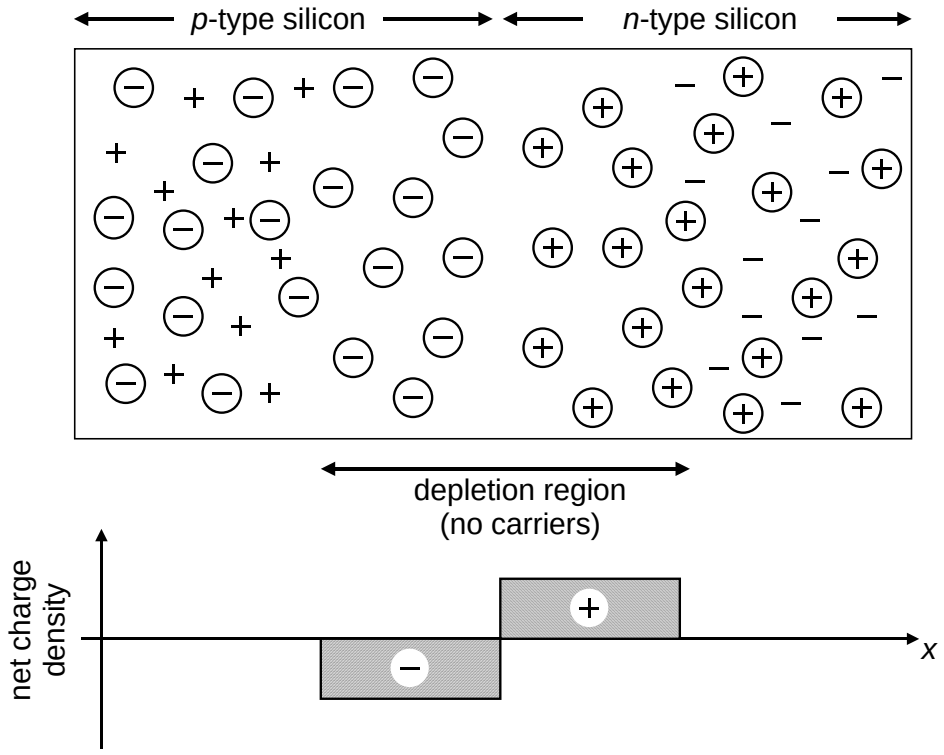
Mobility and diffusion constants are not independent quantities. If we know one, we can compute the other:

$$D = \frac{kT}{q}\mu \quad \text{(the Einstein relation)}$$

where k is Boltzmann's constant (k = 1.38 x 10^{-23} J/K), and the quantity kT/q, often called the **thermal voltage**, is approximately 26 mV at room temperature.

Back to the *p-n* Junction…

Now we have the concepts necessary to understand the behavior of carriers in a *p-n* junction. While there is no drift initially, there are large hole and electrons concentration gradients, giving rise to carrier diffusion. However, as these opposite charge carriers diffuse together, they annihilate one another (recombination), leaving behind a region depleted of carriers – a **depletion region**.



Inside the depletion region, the fixed dopant ions are "uncovered," revealing a net, fixed charge. Now we have an electric field! What happens next?

Electrostatics in 1-D

We need to get a bit more quantitative with regards to charge density and the resulting electric fields. The following two equations take us from charge density ρ to electric field strength E to potential ψ:

$$\frac{dE}{dx} = \frac{\rho(x)}{\varepsilon} \quad \text{(Poisson's equation)}$$

$$\frac{d\psi}{dx} = -E(x)$$

where

$$\varepsilon = k_{Si}\varepsilon_0 \quad \text{(permittivity of silicon)}$$

$$\varepsilon_0 = 8.85\times10^{-14}\ \text{F/cm} \quad \text{(permittivity of free space)}$$

$$k_{Si} = 11.8 \quad \text{(dielectric constant of silicon)}$$

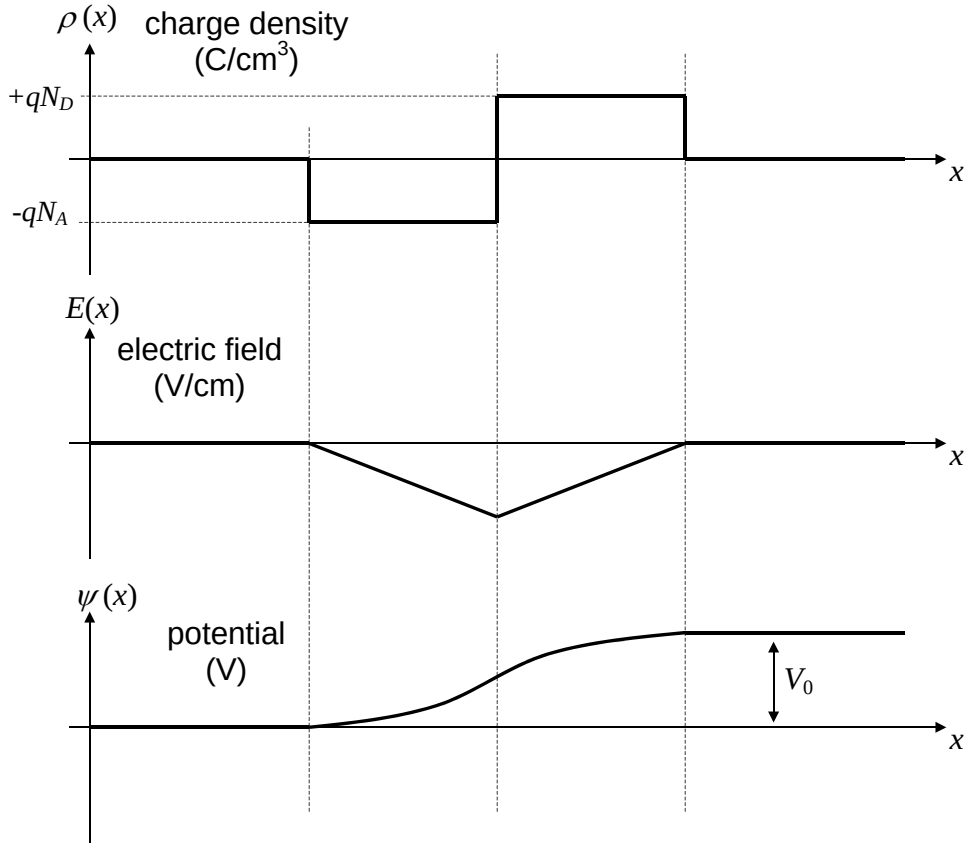
Charge density ρ has units of C/cm^3, and electric field strength E has units of V/cm. We see from Poisson's equation that charge density divided by permittivity gives us units of C/F·cm^2. Since 1 F = 1 C/V, we end up with units of V/cm^2, which is correct for electric field gradient dE/dx.

We can also express these two equations in integral form:

$$E(x_2) - E(x_1) = \frac{1}{\varepsilon}\int_{x_1}^{x_2} \rho(x)dx$$

$$\psi(x_2) - \psi(x_1) = -\int_{x_1}^{x_2} E(x)dx \quad (\Delta Energy = Work = \int Force)$$

Here's a graphical example of how electric field strength and potential (voltage) is computed for a particular charge distribution. Let's assume we build a *p*-*n* junction from *p*-type silicon having an acceptor doping density of N_A and *n*-type silicon having a donor doping density of N_D (units of cm^{-3}).

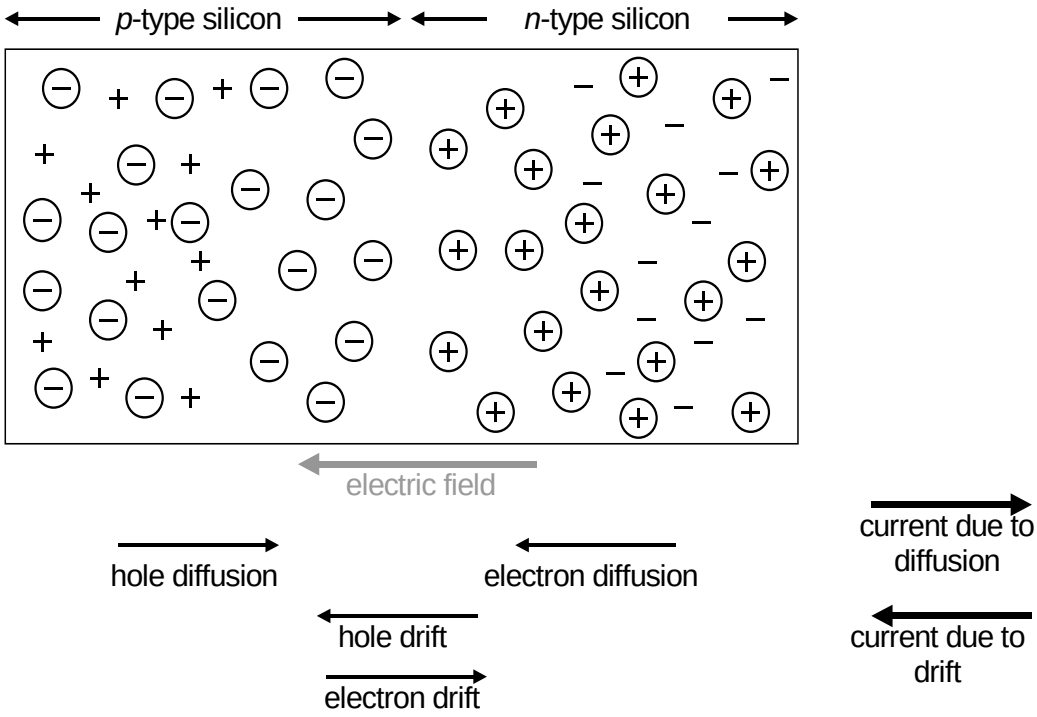


We'll assume that a *p*-*n* junction gives us a "rectangular" distribution of charge as shown above.

The electric field is the integral of the charge density. It decreases linearly in regions of constant negative charge density and increases linearly in regions of constant positive charge density. Another way of looking at this is noticing that the slope of the electric field is given by the charge density.

The potential is the integral of the electric field (with a negative sign). It increases more rapidly when the electric field is stronger. Another way of looking at this is noticing that the slope of the potential is given by the electric field (multiplied by –1).

Now we can identify the diffusion and drift current associated with a p-n junction. From the previous diagram, we see that while the electric field strength varies in the depletion region, it is always negative. This sets up the following situation:



With no external voltage applied to the *p-n* junction, the diffusion and drift currents balance exactly, and there is no net current flow.

We can quantify these various currents using the previous equations for drift and diffusion flux, and calculating current from flux:

$$I_p = qA\phi_p$$

$$I_n = -qA\phi_n$$

Holes generate current in the direction of hole flux. Electrons generate current opposite to the direction of electron flux, since current is defined as the movement of positive charge.

Remember, the unit of flux is $cm^{-2}s^{-1}$. If we multiply by charge (*C*) and area (cm^2), we get *C/s*, or amps.

Thus, the currents can be expressed as:

$$I_{diffusion}(x) = qA\left[\phi_{p-diffusion}(x) - \phi_{n-diffusion}(x)\right]$$

$$= -qAD_p \frac{dp(x)}{dx} + qAD_n \frac{dn(x)}{dx}$$

$$= qA\left[D_n \frac{dn(x)}{dx} - D_p \frac{dp(x)}{dx} \right]$$

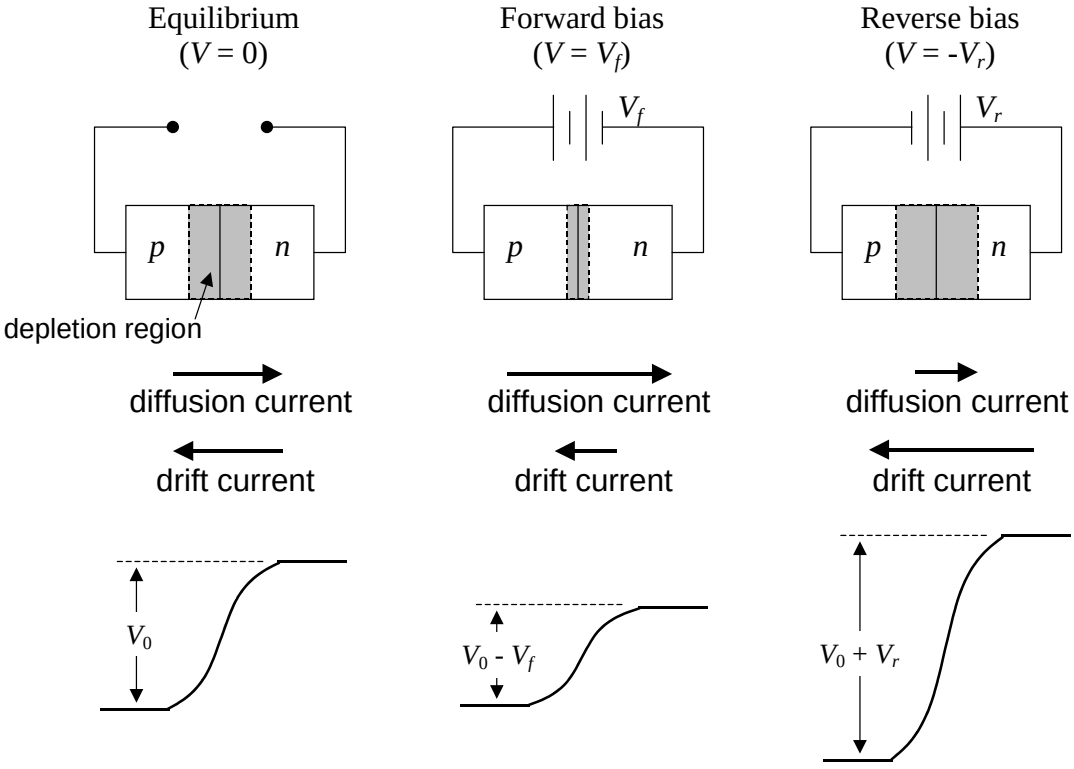
$$I_{drift}(x) = qA\left[\phi_{p-drift}(x) - \phi_{n-drift}(x)\right]$$

$$= qA\mu_p p(x)E(x) + qA\mu_n n(x)E(x)$$

$$= qAE(x)\left[\mu_p p(x) + \mu_n n(x)\right]$$

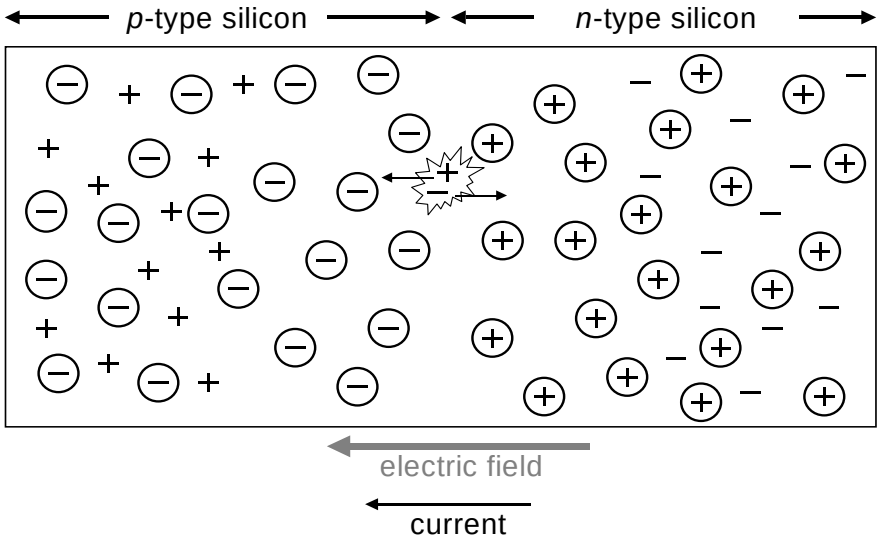
$$I_{total}(x) = I_{diffusion}(x) + I_{drift}(x)$$

By applying an **external voltage** (a bias), we can modulate the width of the depletion region and cause diffusion or drift current to dominate. The result is a net current flow:



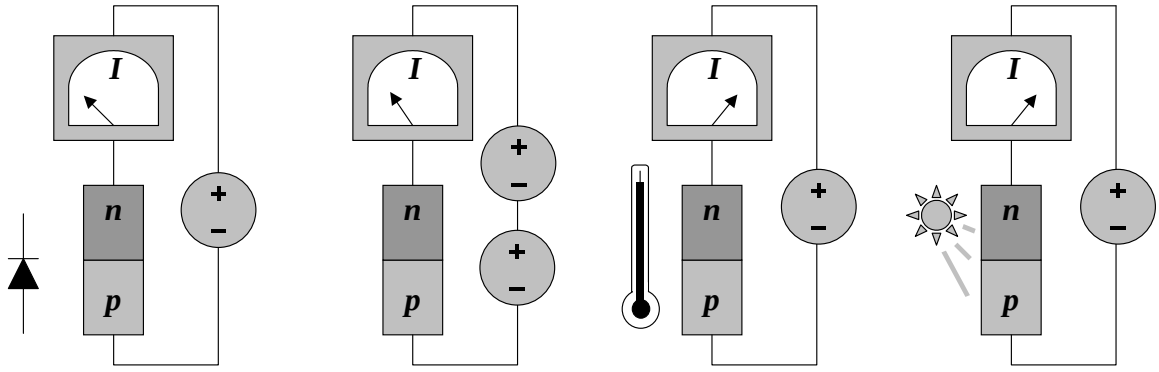
Leakage Current in a *p-n* Junction

What if an electron-hole pair is generated (thermally, or by a photon) **inside** the depletion region?



Any carriers finding themselves in the depletion region are quickly swept away by the strong electric field, leading to drift-based current. When the *p-n* junction is reverse-biased, this tiny drift current dominates and is commonly called **leakage current**.

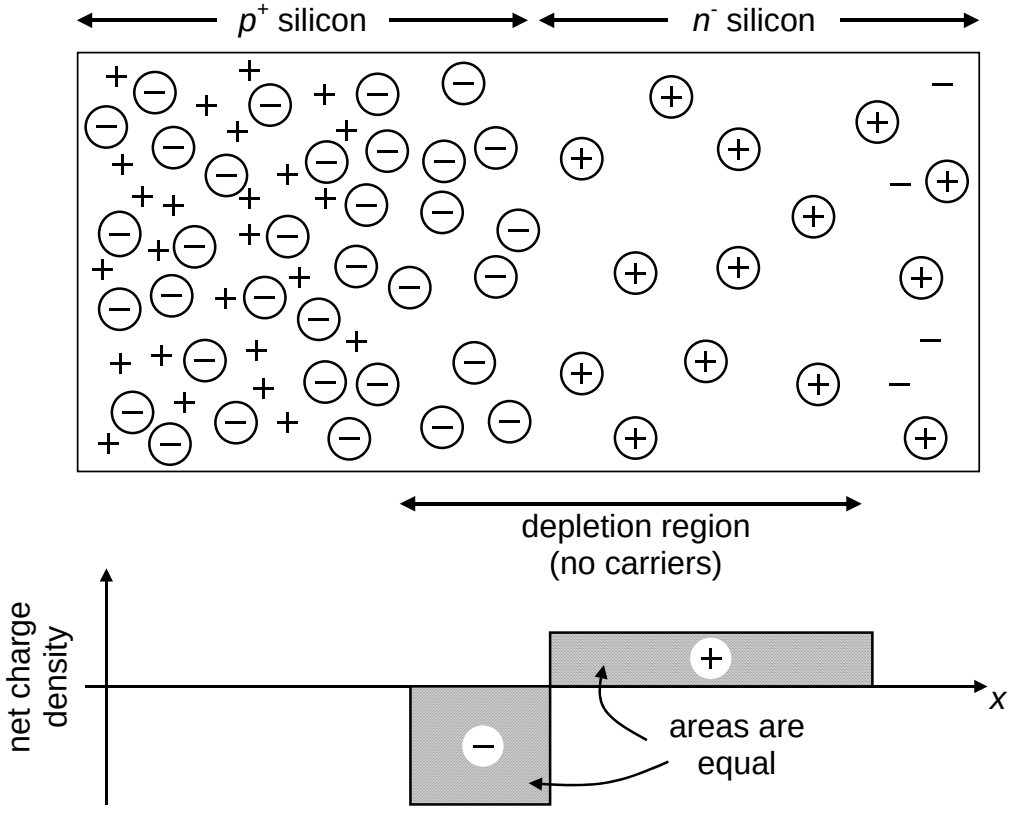
Leakage current is weakly affected by the reverse bias voltage, which increases the strength of the electric field in the depletion region. It is strongly affected by temperature and light, both of which generate more carriers in the depletion region. Leakage current doubles with every 11°C (20°F) increase in temperature. Leakage current is directly proportional to light intensity, and this is the operational basis for photodiodes.



Asymmetrical Junctions

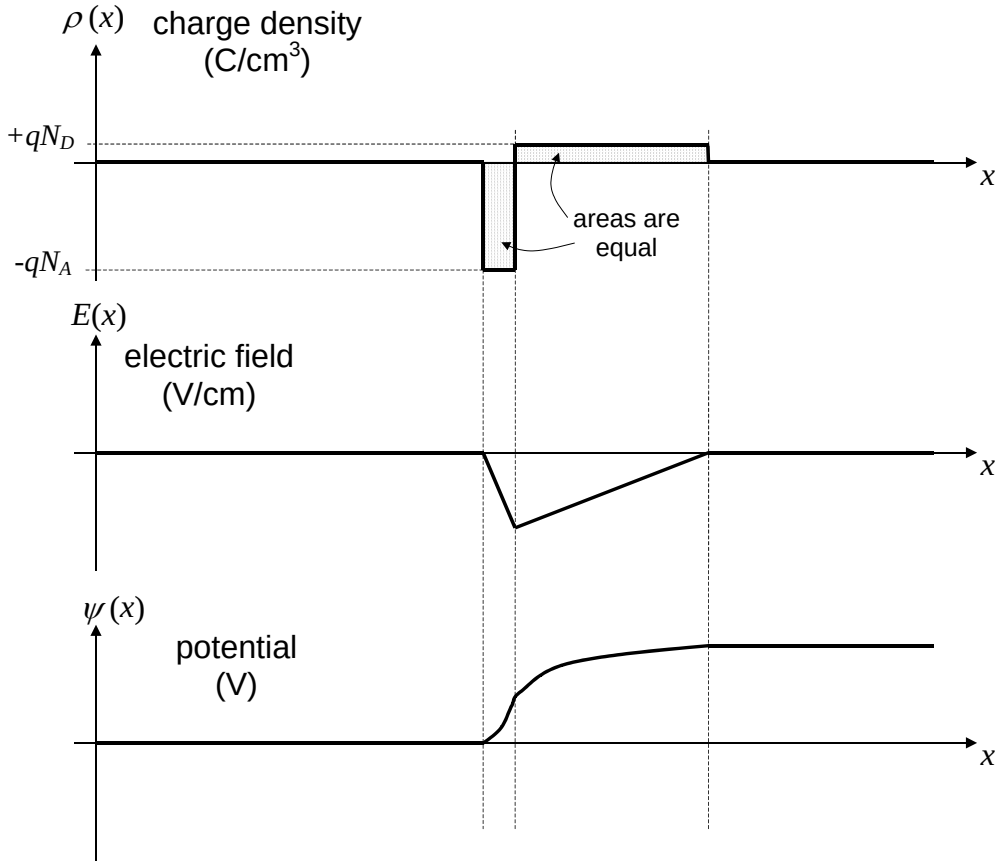
Heavily doped silicon is denoted n^+ or p^+, and weakly doped silicon is denoted n^- or p^-. Notice that the superscripts have nothing to do with positive or negative charge; they only indicate dopant concentrations above or below "average," whatever we define average to be.

In CMOS transistors, we will often encounter asymmetrical *p-n* junctions: either p^+-n^- or n^+-p^- junctions. Let's take a look at one such junction:



An equal number of dopant ions are uncovered on each side of the junction, maintaining charge neutrality. Notice that the depletion region extends further into the more weakly-doped silicon. In the extreme case where the higher dopant concentration is orders of magnitude higher than the lower dopant concentration, the depletion region will exist almost entirely in the weakly-doped region. Only a thin sliver of the heavily-doped region will be depleted.

Charge density, electric field, and potential profile for an asymmetric *p-n* junction:



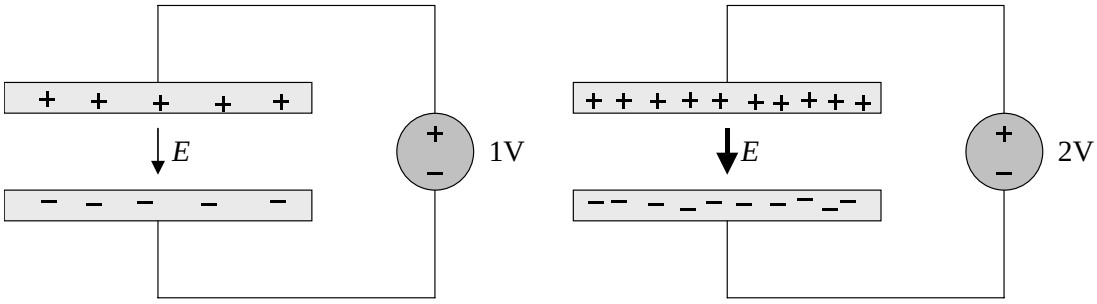
Charge Storage: Capacitance

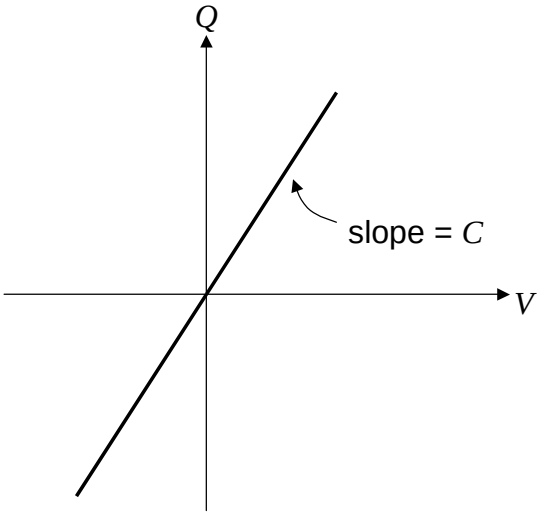
Capacitance is the relationship between **charge** and **voltage**. Capacitance is defined as the incremental change in stored charge resulting from an incremental change in voltage:

$$C \equiv \frac{\partial Q}{\partial V}$$

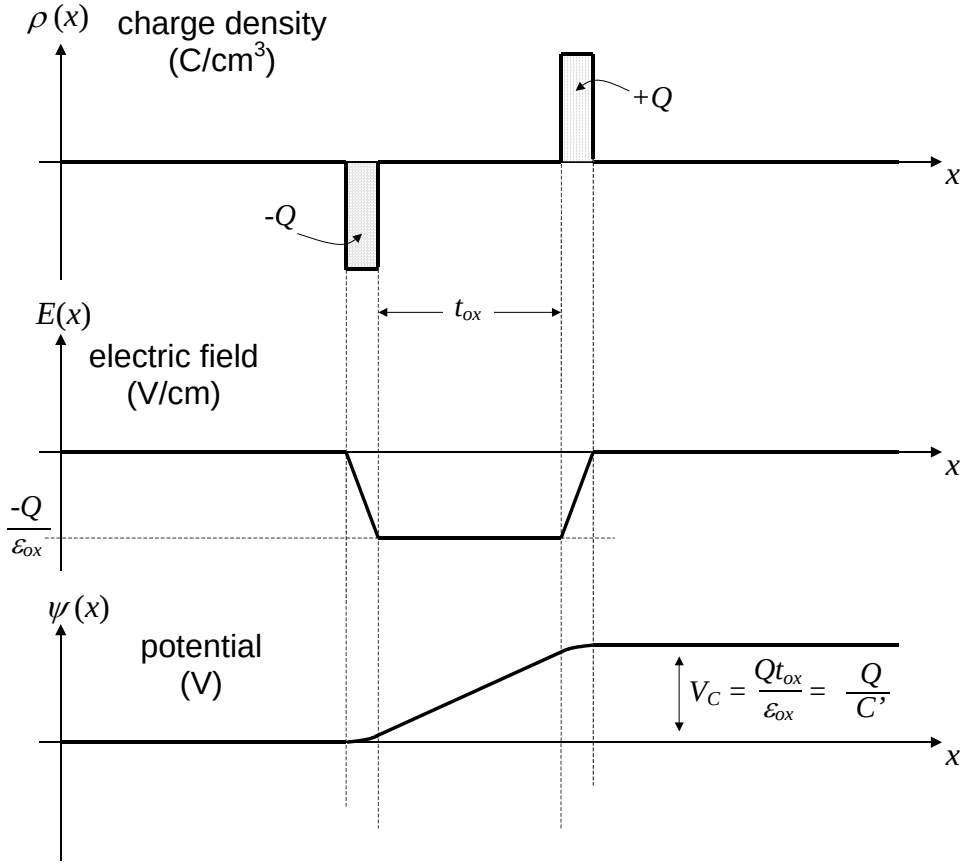
A parallel-plate capacitor is a **linear capacitor**. That is, it's capacitance is constant, and does not depend on the voltage across it or the charge it is holding.

For a linear capacitor, $Q = CV$; the charge stored on each plate is directly proportional to the voltage across the plates.





Let's take a look at the fields and potentials in a parallel-plate capacitor:



In CMOS technology, almost all capacitors use silicon dioxide (SiO_2) as the insulator. The capacitance of an "oxide capacitor" is given by:

$$C_{ox} = A\frac{\varepsilon_{ox}}{t_{ox}}$$

where

$\varepsilon_{ox} = k_{ox}\varepsilon_0$ (permittivity of oxide [SiO_2])

$\varepsilon_0 = 8.85\times10^{-14}$ F/cm (permittivity of free space)

$k_{ox} = 3.9$ (dielectric constant of SiO_2)

Sometimes it is handy to use **capacitance per unit area,** C':

$$C'_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

Example: In a modern 0.35μm CMOS process, the gate oxide thickness is around 80Å = 8nm (1Å = 10^{-10}m). This gives us a capacitance per unit area of 430nF/cm^2. Since we will build structures on the micron level, it is more useful to convert this number to 4.3fF/μm^2.

Remember:

1 pF (picofarad) = 10^{-12}F
1 fF (femtofarad) = 10^{-15}F
1 aF (attofarad) = 10-^{18}F

If we build a 20μm x 20μm capacitor using gate oxide, we can obtain a capacitance of 1.7pF.

Typical values of integrated (on-chip) capacitors range from 50fF (0.05pF) to 50pF.

Below 100fF, parasitic capacitances can dominate. (More on these later.) Capacitors over 5pF begin to consume quite a bit of area.

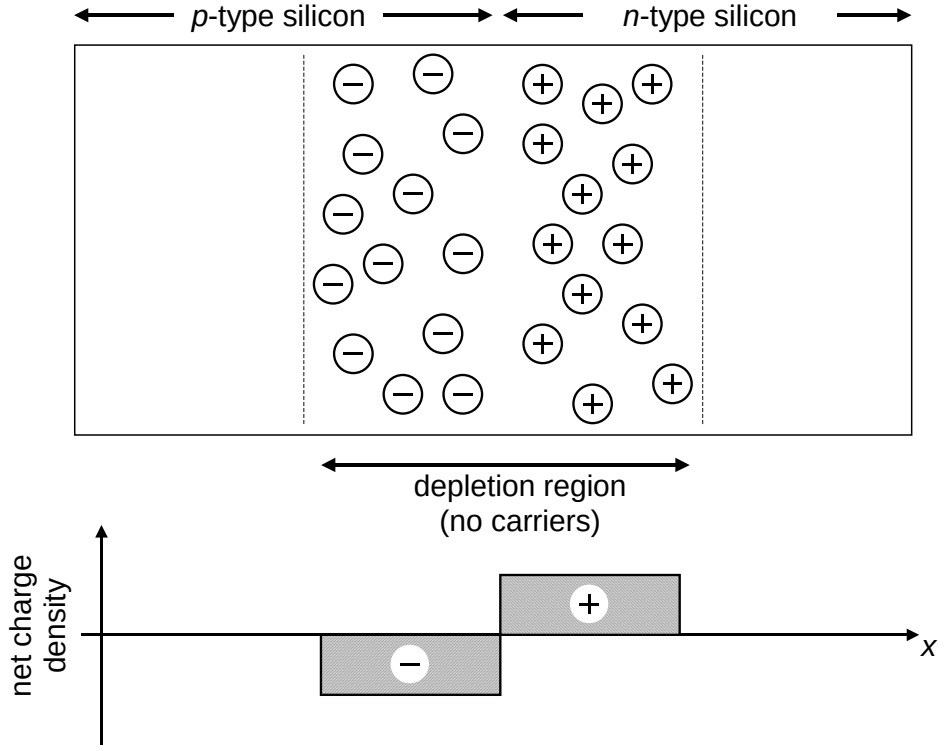
For example, if we build a minimum-length transistor with moderate width (W/L = 10) in 0.35μm technology, its gate area will be 3.50μm x 0.35μm = 1.2μm^2. A good rule of thumb when designing VLSI circuits is that the total area of the circuit will be approximately 2-3 times the total gate area of all the transistors. This extra area is consumed by source/drain regions, wires, and spacing between transistors. So let's assume that each transistor will consume about 3μm^2 of silicon. How big is a capacitor, in terms of transistors? Since our gate capacitance is 4.3fF/μm^2, this works out to 13fF/(transistor area). So a 10pF capacitor takes up as much room as 770 transistors! Capacitance is **expensive** in terms of silicon area!

The situation gets worse if we can't use gate oxide to build capacitors. Gate oxide capacitors have a semiconductor as their bottom plate (they are often called MOS capacitors or MOScaps), and this can result in nonlinear behavior. Linear capacitors are often built between two layers of polysilicon ("poly1" and "poly2"), but the oxide between these layers is thicker than the gate oxide. In the 0.35μm process mentioned above, the gate oxide capacitance is 4.3fF/μm2, while the poly1/poly2 capacitance is only 0.8fF/μm^2! Using our previous assumptions, this works out to 2.4fF/(transistor area), so a 10pF capacitor takes up as much room as 4200 transistors!

The Capacitance of a *p-n* Junction

A *p-n* junction stores charge and has a potential difference across it, so it must have a capacitance. Let's consider the case of a **reverse-biased** *p-n* junction – the kind we will encounter most often in CMOS design.

Below is the familiar picture of the charge stored in a p-n junction. Now we are using even more shorthand when we draw our cartoons: We will no longer show the *p* and *n* regions outside of the depletion region, because these regions are neutral (have no net charge):



We've also added dotted lines to show the boundaries of the depletion region. This will come in handy later.

To calculate the capacitance of this junction, we have to find an expression relating charge to voltage. Since this junction is symmetric, we can make things easier by finding the voltage drop across one half of the junction and then multiplying by two.

Let's assume the total charge in the *p* part of the junction is doped with an acceptor density of N_A (units cm^{-3}). This means the charge density in the depletion region is qN_A. The junction has a cross-sectional area of A.

We can calculate the electric field $E(x)$ and the potential $\psi(x)$:

$$E(x) = \int \frac{\rho(x)}{\varepsilon} dx = \int \frac{-qN_A}{\varepsilon} dx = -\frac{qN_A}{\varepsilon} x$$

$$\psi(x) = -\int E(x) dx = \int \frac{qN_A}{\varepsilon} x dx = \frac{qN_A}{2\varepsilon} x^2$$

Assuming the width of this half of the depletion region is W, we can calculate the voltage drop across the junction, which is equal to the junction's built-in potential V_0 plus the reverse bias V_r:

$$V_0 + V_r = \psi(W) = \frac{qN_A}{2\varepsilon} W^2$$

The total depletion charge on one side of the junction is given by:

$$Q = -qN_A AW$$

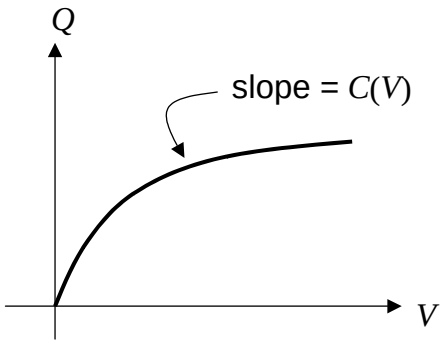
which we can rearrange to give us width as a function of charge:

$$W = -\frac{Q}{qN_A A}$$

Now we can find an expression relating charge to voltage:

$$V_0 + V_r = \frac{qN_A}{2\varepsilon} W^2 = \frac{Q^2}{2q\varepsilon N_A A^2}$$

$$Q = A\sqrt{2q\varepsilon N_A (V_0 + V_r)}$$



Now we can find the junction capacitance:

$$C_j \equiv \frac{\partial Q}{\partial V_r} = \frac{A}{2}\sqrt{\frac{2q\varepsilon N_A}{V_0 + V_r}}$$

The capacitance changes with the voltage across it, so this is a **nonlinear** capacitor. The junction capacitance per unit area is given as:

$$C_j' = \frac{1}{2}\sqrt{\frac{2q\varepsilon N_A}{V_0 + V_r}}$$

For the more general case where $N_A \neq N_D$, it can be shown that:

$$C_j' = \frac{1}{2}\sqrt{\frac{2q\varepsilon}{V_0 + V_r}\frac{N_D N_A}{N_D + N_A}}$$

Often, this is written as:

$$C_j' = \frac{C_{j0}'}{\sqrt{1 + \frac{V_r}{V_0}}}$$

where C'_{j0} is the capacitance per unit area when V_r = 0:

$$C_{j0}' = \frac{1}{2}\sqrt{\frac{2q\varepsilon}{V_0}\frac{N_D N_A}{N_D + N_A}}$$

In the case of an asymmetric p^+-n^- junction, where $N_A \gg N_D$, we get:

$$C_j' = \frac{1}{2}\sqrt{\frac{2q\varepsilon N_D}{V_0 + V_r}}$$

Incidentally, we can compute the *p-n* junction's built-in potential V_0 (sometimes called Φ_0) from doping parameters:

$$V_0 = \Phi_0 = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2}$$

where n_i is the intrinsic carrier concentration for silicon: $1.5 \times 10^{10} cm^{-3}$ at room temperature (and highly dependant on temperature). V_0 is typically in the range of 0.7-0.9V.

Example: The n^+ drain region of an nMOS transistor forms an asymmetric junction with the p^- substrate. The drain has dimensions of 7μm x 2μm. The substrate doping level is $N_A = 10^{16} cm^{-3}$, $N_D \gg N_A$, and $V_0 = 0.8V$. The substrate is at ground. What is the drain-to-substrate "parasitic" capacitance for V_{drain} = 0V, 1V, and 5V?

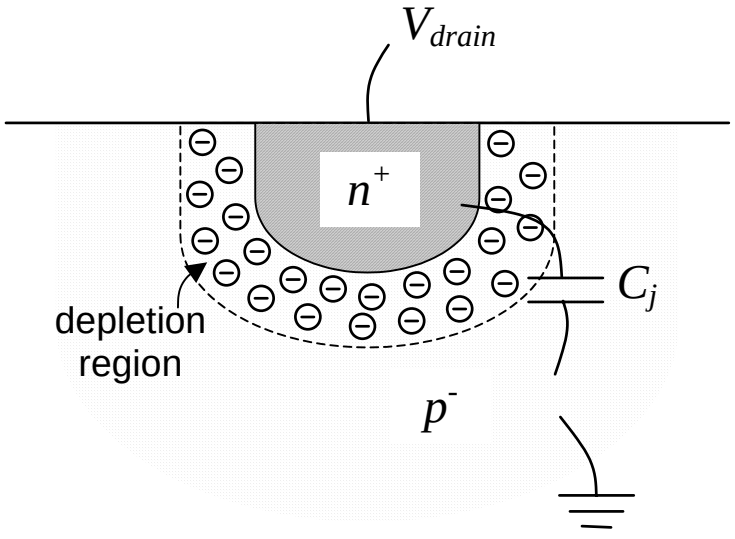
$$A = 7\mu m \times 2\mu m = 14\mu m^2$$

$$C_j = \frac{A}{2} \sqrt{\frac{2q\varepsilon N_A}{V_0 + V_r}}$$

$C_j(V_{drain} = 0V) = 4.5fF$

$C_j(V_{drain} = 1V) = 3.0fF$

$C_j(V_{drain} = 5V) = 1.7fF$

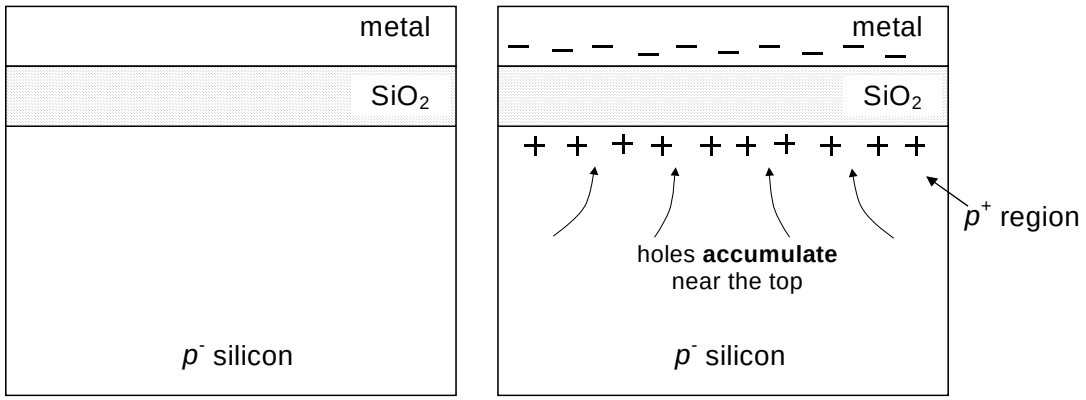


The MOS Capacitor

Suppose we build a parallel-plate capacitor where one plate is metal, one plate is a semiconductor (e.g., weakly-doped silicon), and the insulator is SiO_2. Such a device is called a MOS (metal-oxide-semiconductor) capacitor. The metal plate is called the **gate,** and is not always built out of metal. Nowadays, gates are made from heavily-doped polycrystalline silicon (or "polysilicon," or just "poly"). Polysilicon does not have a rigid crystal lattice, and conducts current freely, acting almost like a metal.

The MOS capacitor has several distinct regions of operation. For our example, we will consider a MOS capacitors build with p^- silicon.

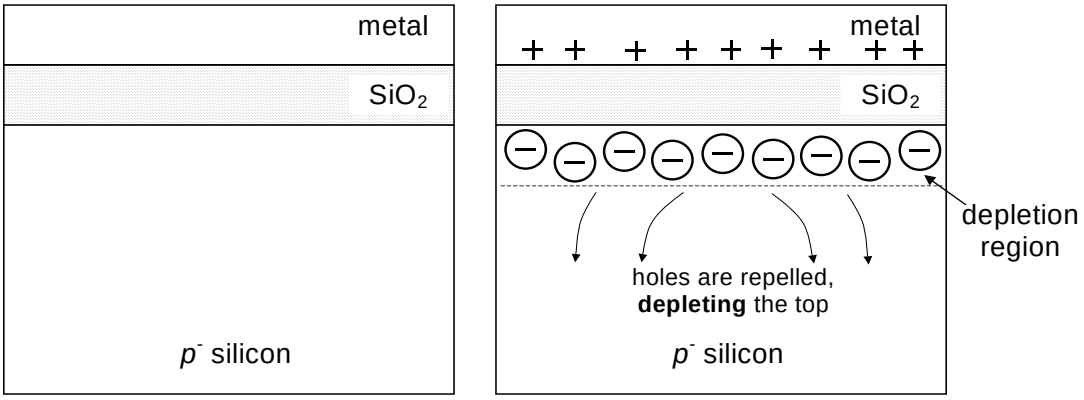
With no external voltage placed across it, the MOS capacitor is in the **flat band** region (ignoring work functions and implanted fixed charge). If we lower the voltage on the gate (by introducing negative charges), we attract "extra" holes (above the background level of holes in the p^- silicon) which **accumulate** on the surface of the silicon. This creates a thin p^+ region near the silicon-oxide interface.



flat band accumulation

In the accumulation regime, MOS capacitors act as linear capacitors since the p^+ region acts as a highly conductive "bottom plate."

If instead we raise the gate voltage (by introducing positive charges), we "scare away" holes in the p^- silicon. This **depletes** the surface of holes, creating a depletion region with exposed negative dopant ions.

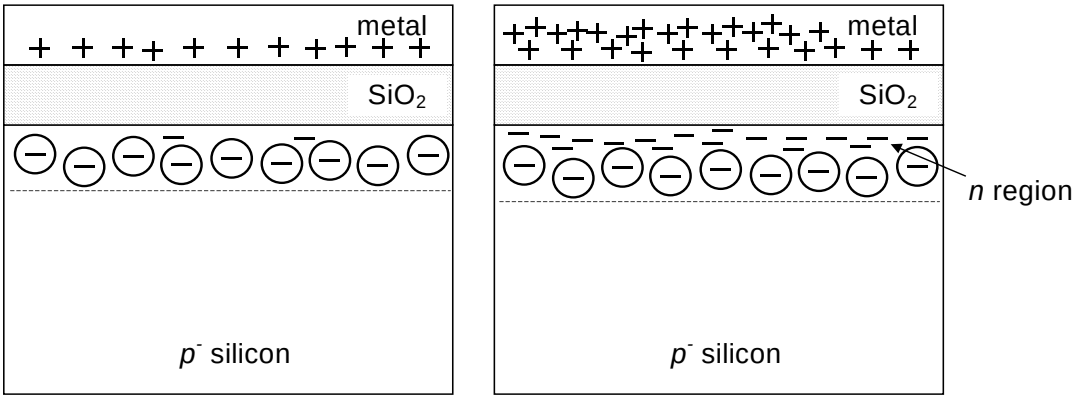


flat band depletion

As we have seen, the depletion capacitance is nonlinear, so a MOS capacitor operating in this regime is not very linear. It can be modeled as two capacitors in series: a linear oxide capacitance and the nonlinear depletion capacitance. Thus, the total capacitance is **less** than the oxide capacitance.

Due to contact potentials between the substrate and gate, MOS capacitors are typically in the depletion region when the gate and substrate are at the same potential.

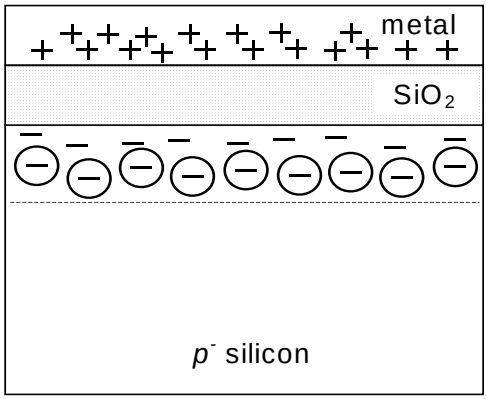
As we continue to raise the gate voltage, the depletion region cannot provide enough negative charges to match all the positive charges we are putting on the gate. Eventually, these positive charges begin to attract thermally-generated electrons (and repel thermally-generated holes), producing mobile electrons which balance the charge.



weak inversion strong inversion

The production of mobile electrons begins to **invert** the silicon at near the surface – it changes it from *p*-type silicon to *n*-type silicon! When the number of mobile electrons (the **inversion charge**) is much lower than the number of exposed dopant ions in the depletion region (the **depletion charge**), we are in the regime of **weak inversion**. When the inversion charge greatly exceeds the depletion charge, we have **strong inversion**, and a conductive n^+ layer forms at the surface of the semiconductor.

When the inversion charge and depletion charge are comparable, we are in a regime know as **moderate inversion**.

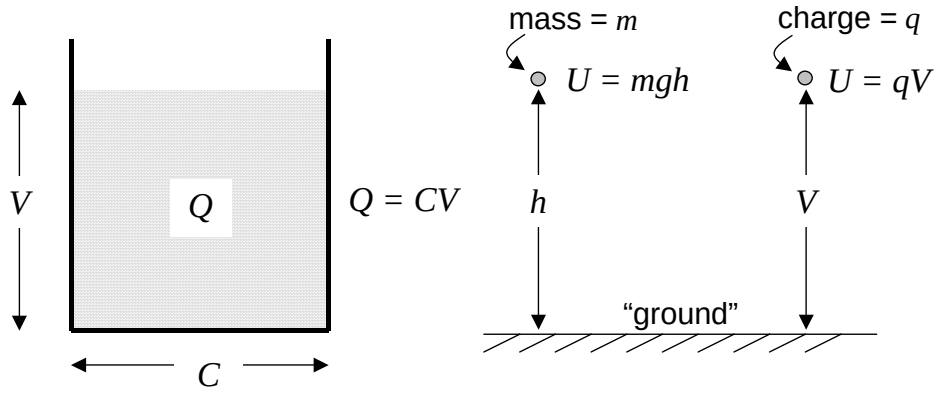


moderate inversion

The Hydraulic Analogy of Charge

As we explore the operation of transistors, it will be useful to observe the analogy between electricity and water. While this analogy shouldn't be taken too far, it is surprisingly versatile.

For example, we can draw an analogy between charge storage devices (capacitors) and water storage devices (reservoirs).



The energy stored in a capacitor can be calculated using the hydraulic analogy. A water molecule with mass m at height h above the ground has a potential energy of $U = mgh$. Analogously, a particle with charge q and potential V has an energy of $U = qV$. Calculating the total energy in a capacitor is analogous to calculating the total potential energy of water in a reservoir.

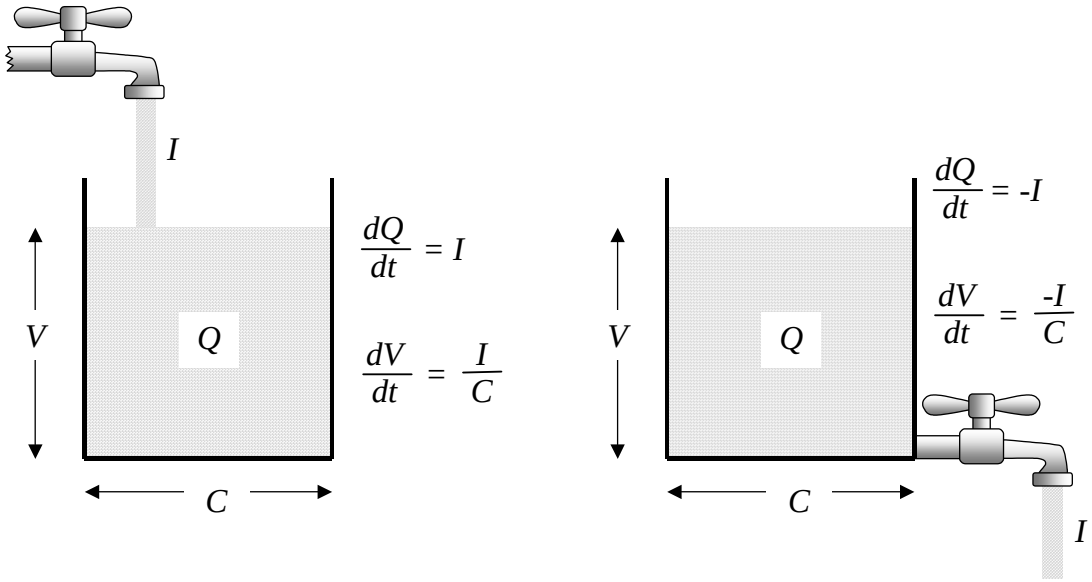
In a "charge reservoir" with capacity C, a thin slice of charge (thickness = dv) at height v contains a total charge equal to Cdv, and thus has a potential energy of $Cdv \cdot v$. Integrating this from $v = 0$ to $v = V$, we get

$$U_{\text{linear capacitor}} = \int_0^V Cv \cdot dv = \frac{1}{2}Cv^2\Big|_0^V = \frac{1}{2}CV^2$$

which is the familiar expression for energy stored in a capacitor.

If we think of current as a stream of water, we can find an analogy for the familiar equation describing capacitor charging and discharging:

$$i = C\frac{dV}{dt}$$



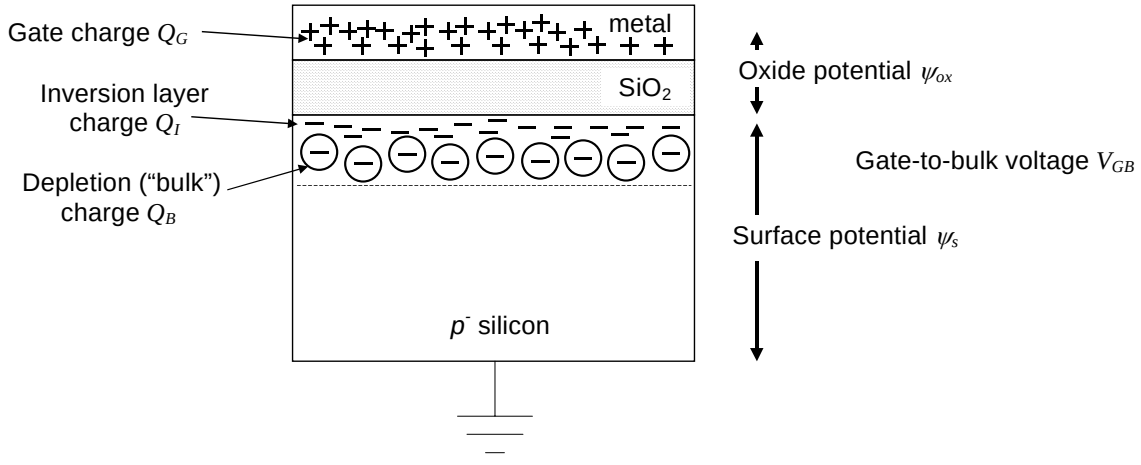
It's easy to see that a larger capacitor (larger reservoir area) will take longer to charge (fill) given a fixed current.

Now we understand the building blocks of CMOS transistors:

- *p-n* junctions
- MOS capacitors

The MOS Transistor in Strong Inversion

In this section, we shall explore the behavior of the MOS transistor when the area under the gate – the **channel** – is strongly inverted. Strong inversion is shown below for a MOS capacitor.



The gate-to-bulk voltage V_{GB} can be decomposed into the potential across the oxide (ψ_{ox}) and the **surface potential** of the silicon substrate (ψ_s).

Ignoring any implanted charge or contact potential effects, the charge on the gate (Q_G) must be balanced out by the charge in the channel. The channel charge consists of the fixed depletion or "bulk" charge Q_B and the mobile inversion layer charge Q_I:

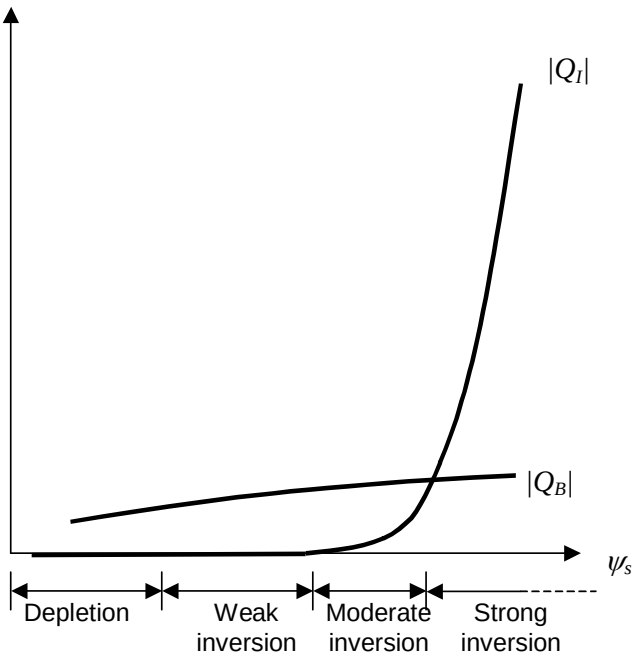
$$|Q_G| = |Q_I + Q_B|$$

As we saw last time, the charge in a uniformly-doped depletion region is proportional to the square root of the voltage across it:

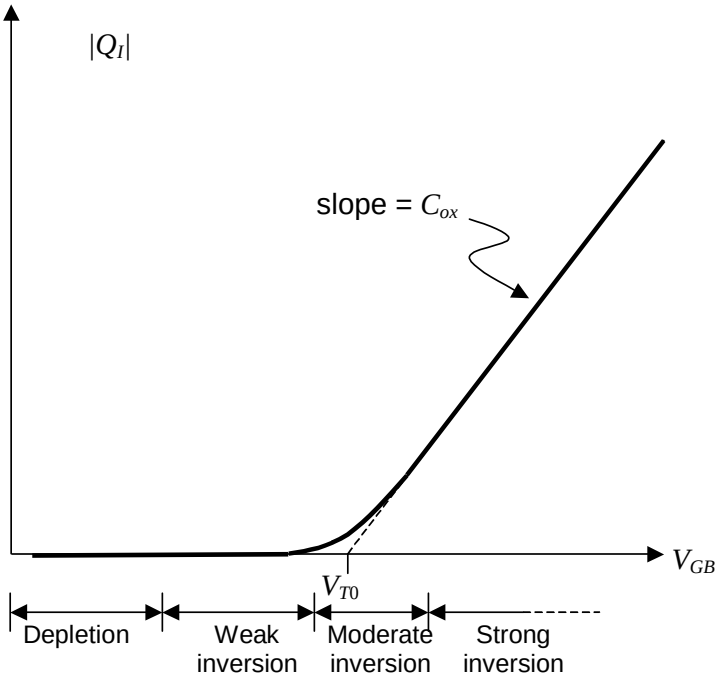
$$Q'_B = -\sqrt{2q\varepsilon N_A}\sqrt{\psi_s}$$

In strong inversion, the inversion layer charge far surpasses the depletion charge.

$$|Q_I| >> |Q_B| \text{ in strong inversion}$$



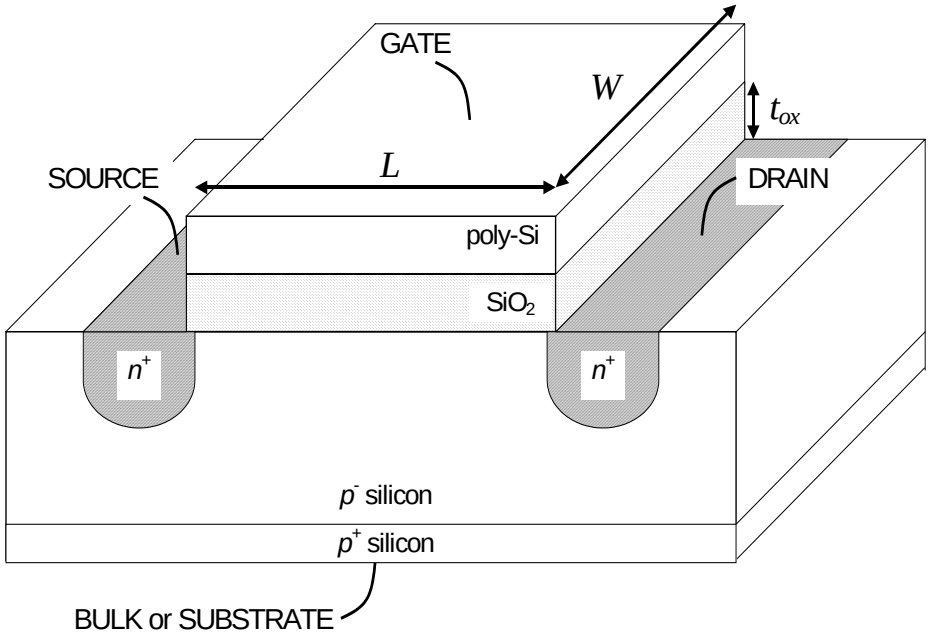
In strong inversion, the gate charge is balanced out primarily by the inversion layer charge. The voltage at which inversion layer charge dominates is called the **threshold voltage** V_T. The symbol V_{T0} will often be used, and this indicates the threshold voltage when the source voltage equals zero (more on this later).



$$Q_I' \cong -C_{ox}'\left(V_{GB} - V_{T0}\right) \text{ in strong inversion}$$

Of course, to make a transistor we need more than a gate; we also need a source and a drain. Charge carriers flow through the channel (under the gate) from the source to the drain.

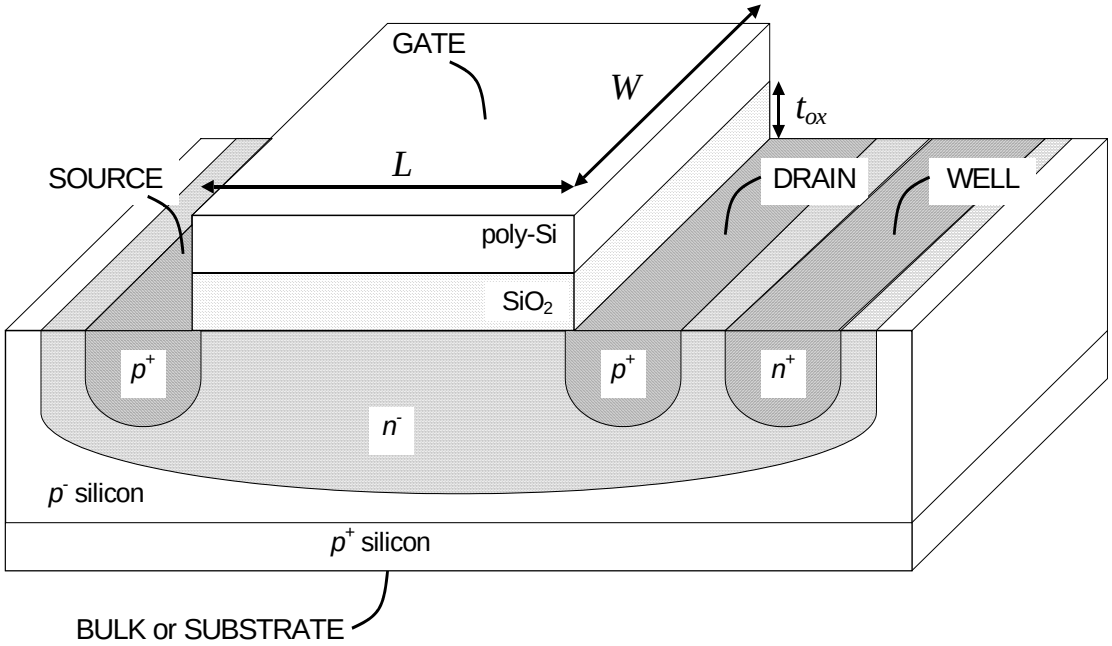
An *n*-channel MOSFET ("nMOS transistor" or "nFET") is shown below. Notice that an *n*-channel FET actually has a channel made of weakly-doped *p*-type silicon. However, when the transistor conducts current, this channel is inverted, and is thus *n*-type.



Electrons carry charge in nFETs (since the inverted channel is *n*-type). Since electrons flow from the source to the drain, current flows from the drain to the source.

The channel has a width W and a length L. The width-to-length ratio ("W/L ratio") is an important parameter in MOSFET operation, as we shall see.

The complementary type of MOSFET is a *p*-channel MOSFET ("pMOS transistor" or "pFET"). This is shown below. Although the channel is made from *n*-type silicon, it becomes *p*-type when inverted.

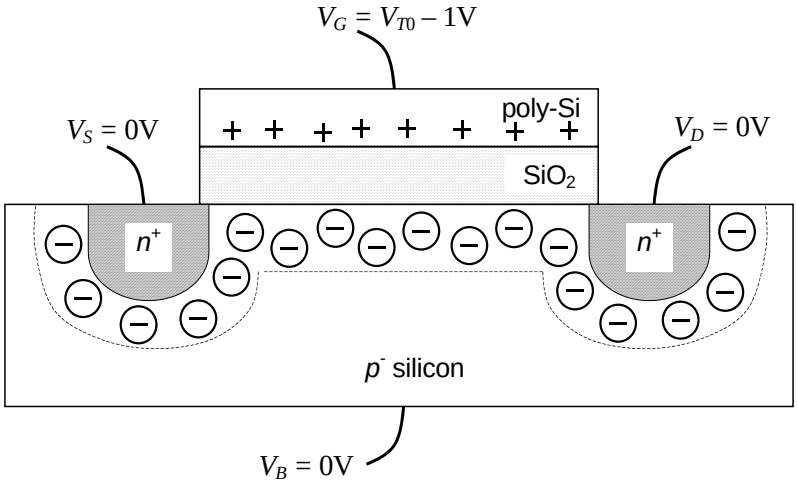


In a complementary-MOS (CMOS) process where both nFETs and pFETs can be built, one of the two devices must be built in a well whose doping is opposite that of the substrate. In most processes nowadays, p^- substrates and n^- wells are used, so that pFETs reside inside wells.

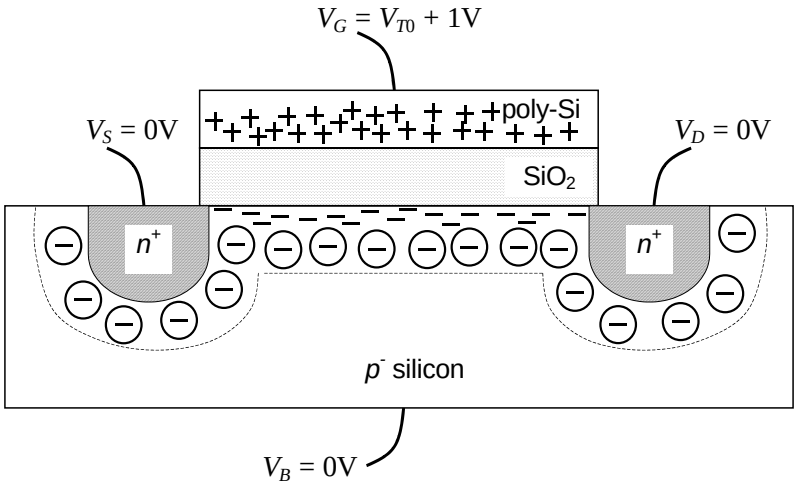
The substrate is always set at the lowest potential used in the circuit (usually called ground or V_{SS}) so that the nFET source/drain regions stay reverse-biased. Likewise, the wells are usually set to the highest potential used in the circuit (usually called V_{DD}) so that the pFET source/drain regions stay reverse-biased.

If we place a metal wire (aluminum is used in most VLSI processes) directly against a semiconductor, we get a rectifying contact – a **Schottky barrier diode**. This is obviously not what we want when we tie a wire to our transistor. One way to avoid this diode is to connect metal only to heavily-doped (n^+ or p^+) regions. This forms a nonrectifying **ohmic contact** with a typical resistance of a few tens of ohms. This explains why we connect to the well via a n^+ **well contact** region and connect to the substrate via a p^+ **substrate contact** region.

Now let's take a look at an nFET with a gate voltage below V_T (subtheshold). The only charge in the channel is due to the depletion region, which also extends around the source and drain *p-n* junctions. (For now, we will neglect the small number of mobile electrons present in this condition.)



Now, if we raise the gate voltage above the threshold voltage V_T, we create an inversion layer of mobile electrons that balance out most of the gate charges:

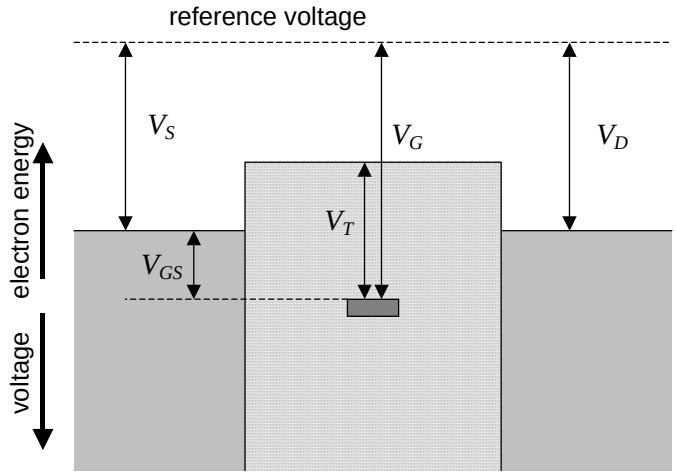


The charge in the inversion layer is proportional to the **effective voltage** V_{eff}.

$$V_{eff} = V_{GS} - V_T$$

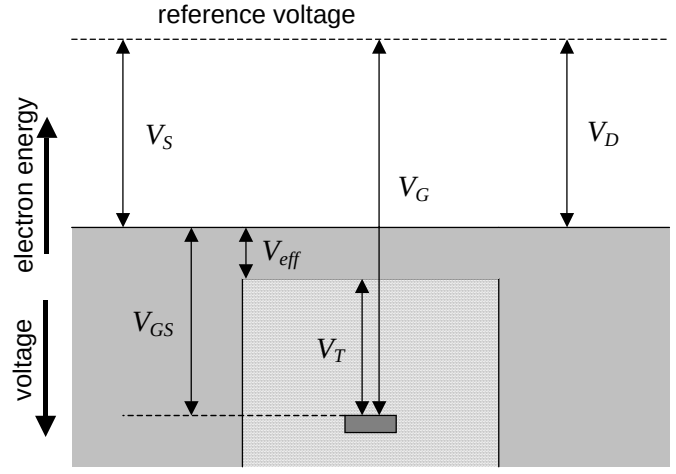
$$Q_I' \cong -C_{ox}' V_{eff} \quad \text{in strong inversion}$$

We can represent the channel charge with a simple fluid model:



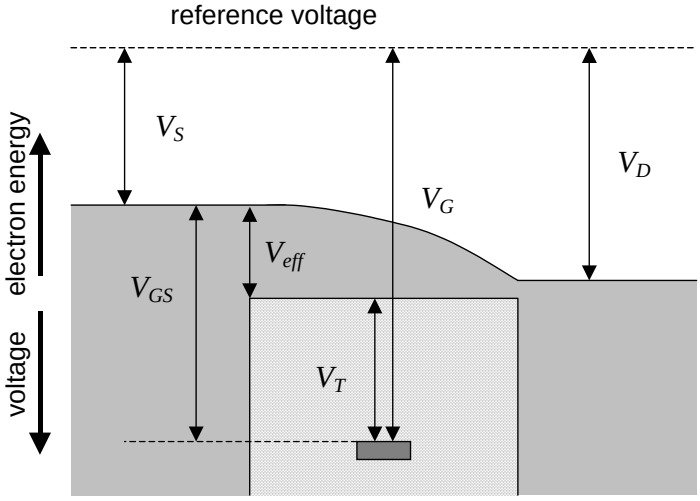
Here, the channel is represented by a barrier whose width corresponds to the channel's length. The barrier can be moved up and down by means of a handle placed some distance below the top. Water on either side of the barrier represents the electrons in the source and drain. Since electrons have negative charge, the voltage scale is inverted; increasing voltage is in the downward direction.

Let's consider the case where $V_S = V_D$. When the barrier is above the water level (i.e., $V_{GS} < V_T$), the source and drain tanks are separated, and no water flows. There is no channel charge (water above the barrier).

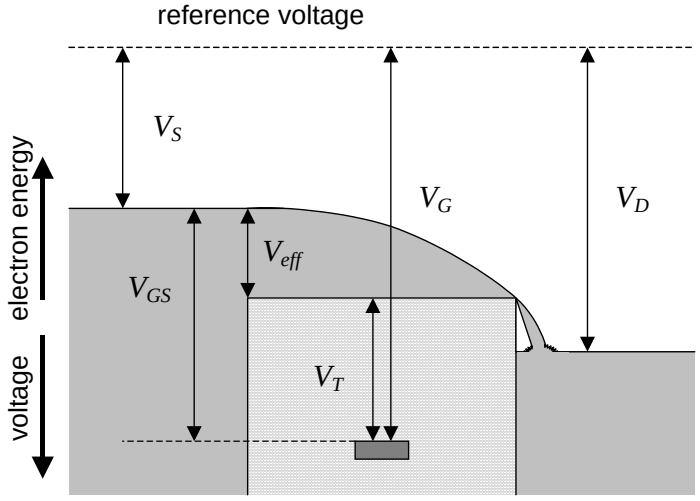


When the barrier is pulled below the source/drain water level (i.e., $V_{GS} > V_T$), water flows in and fills the channel. The water in the channel (representing the channel charge) is proportional to V_{eff}. Of course, there is no net water flow here since the source and drain water levels are equal.

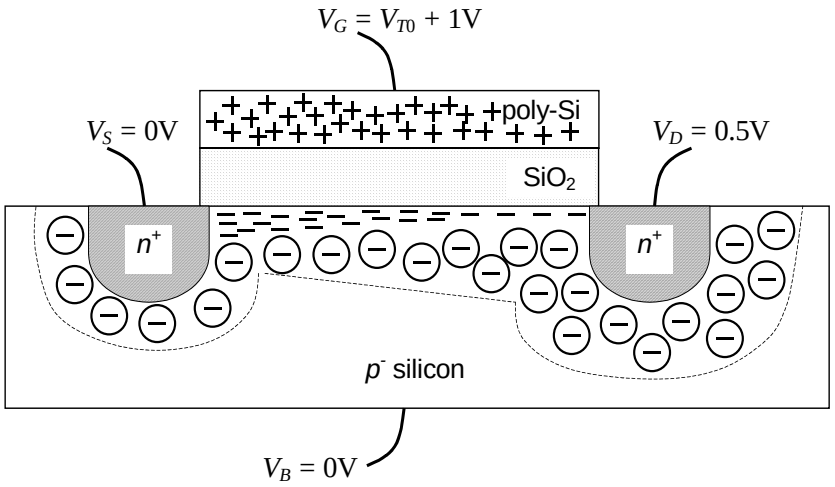
Now, if we lower the level of the drain tank (i.e., raise the drain voltage V_D), water flows from the source to the drain. It is clear to see that two things would allow more water to flow: lowering the barrier (i.e., increasing the gate voltage V_G) or increasing the difference between the source and the drain water levels (i.e., raising the drain-to-source voltage V_{DS}).



What if we continue to lower the drain water level (i.e., raise the drain voltage) to the point where $V_{DS} > V_{eff}$? In this case, the water over the channel goes almost to zero at the drain edge, and we create a "waterfall" that spills into the drain tank. Notice that the rate of water flowing across the barrier (i.e., the channel current) is independent of the drain water level! We can still increase the channel current by raising the source level (i.e., lowering the source voltage) or lowering the barrier (i.e., raising the gate voltage).



Now let's return to our cross-section of an actual nFET and observe the channel in strong inversion when $V_D > V_S$:



Notice that the depletion layer widens towards the drain since the drain *p-n* junction is more strongly reverse-biased than the source *p-n* junction. More importantly, the inversion layer contains more charge towards the source. Why is this?

Near the source, the channel potential (surface potential) V_C is approximately equal to V_S. Near the drain, the channel potential is approximately equal to V_D. Since V_S is lower than V_D, the voltage across the oxide is greater near the source, and charge is proportional to voltage in a parallel-plate capacitor. We can write the charge at position x under the channel as:

$$Q_I(x) = WC'_{ox}\left[V_{GS} - V_C(x) - V_T\right]dx$$

Where the source is at $x = 0$ and the drain is at $x = \mathrm{L}$.

This non-uniform distribution of charge creates an electric field along the channel, sweeping electrons from the source to the drain. The current at each point in the channel is given by:

$$I_D = Q_I(x)\mu_n E(x)$$

The current is just the charge multiplied by its velocity, which is the electric field strength multiplied by the mobility. The electric field is spatial derivative of the channel voltage, so we can write the channel current as:

$$I_D = WC'_{ox}\left[V_{GS} - V_C(x) - V_T\right]\mu_n \frac{dV_C(x)}{dx}$$

Now we solve for I_D using the boundary conditions $V_C(0) = V_S$ and $V_C(L) = V_D$. Assuming for now that $V_S = 0$, we get:

$$\int_{x=0}^{L} I_D dx = \int_{V-0}^{V_{DS}} W\mu_n C'_{ox}\left[V_{GS} - V_C(x) - V_T\right]dV_C$$

$$I_D = \mu_n C'_{ox}\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

This is the familiar MOS transistor equation valid in strong inversion ($V_{GS} > V_T$).

For small V_{DS} [$V_{DS} << 2(V_{GS} - V_T)$], the V_{DS}^2 term is very small, and we can make the following approximation:

$$I_D \approx \mu_n C'_{ox}\frac{W}{L}(V_{GS} - V_T)V_{DS}$$

$$\approx \frac{V_{DS}}{R_{on}}$$

where

$$R_{on} = \frac{1}{\mu_n C'_{ox}\frac{W}{L}(V_{GS} - V_T)}$$

In other words, for small drain-to-source voltages, the MOSFET behaves like a resistor. The resistor's value decreases as we increase the gate voltage. This is not at all surprising since that increases the amount of inversion charge. This also makes sense when we look back at the fluid model.

Now let's return to the original MOSFET equation:

$$I_D = \mu_n C'_{ox}\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right]$$

When $V_{DS} = V_{GS} - V_T$, I_D is maximum:

$$I_{D,\max} = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}(V_{GS} - V_T)^2$$

Here, Q_I goes to zero at the end of the channel since $V_D = V_{eff}$. Wait a minute: How can the transistor conduct if there is zero charge at the end of the channel? Well, the charge doesn't go *exactly* to zero, but it is quite small, which means the charge is moving extremely fast – the electric field is very strong here.

Look back at the fluid model where $V_{DS} > V_{eff}$. The slope of the water level corresponds to the electric field strength. The water is moving fastest at the "pinch off" point adjacent to the drain. In fact, this condition in transistors is called **pinch off**.

What happens if we increase V_D further? When $V_D > V_{eff}$, do positive charges accumulate in the channel? No. In fact, very little changes beyond this point, as the fluid model makes clear. The drain becomes "disconnected" from the channel, meaning that changes no in drain voltage no longer affect channel current (to first order).

So the equation we derived for MOSFET operation is only valid up to this point. After this point, the current is "frozen" at that maximum value; it does not follow the parabola described by the equation. So we can describe the current flowing into the drain of an *n*MOS transistor using two equations (plus one useful approximation):

$$I_D = \mu_n C'_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] \qquad V_{DS} < V_{GS} - V_T \text{ (**triode** region)}$$

$$I_D \approx \mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)V_{DS} \qquad V_{DS} << V_{GS} - V_T \text{ (deep triode region)}$$

$$I_D = \frac{1}{2}\mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)^2 \qquad V_{DS} \geq V_{GS} - V_T \text{ (**saturation** region)}$$

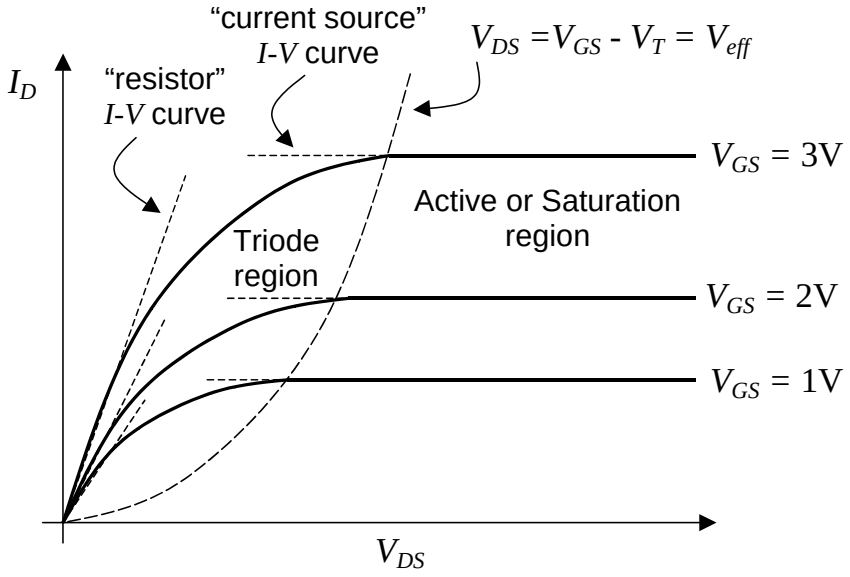
For pMOS transistors, we just use the hole mobility instead of electron mobility, and insert a minus sign since the charge carriers are now holes (or you can just think of the current as flowing *out* of the drain instead of into it). Also, pFETs have negative threshold voltages (but V_{GS} will also be negative in normal operation).

$$I_D = -\mu_p C'_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] \qquad |V_{DS}| < |V_{GS} - V_T| \text{ (**triode** region)}$$

$$I_D \approx -\mu_p C'_{ox} \frac{W}{L}(V_{GS} - V_T)V_{DS} \qquad |V_{DS}| << |V_{GS} - V_T| \text{ (deep triode region)}$$

$$I_D = -\frac{1}{2}\mu_p C'_{ox} \frac{W}{L}(V_{GS} - V_T)^2 \qquad |V_{DS}| \geq |V_{GS} - V_T| \text{ (**saturation** region)}$$

This behavior is shown graphically below. When a MOSFET has a low voltage across it (low V_{DS}), it acts like a voltage-controlled resistor. The resistance is inversely proportional to V_{eff}. When a MOSFET has a sufficiently high voltage across it (high V_{DS}), it acts like a voltage-controlled current source. The current is proportional to the square of V_{eff}.

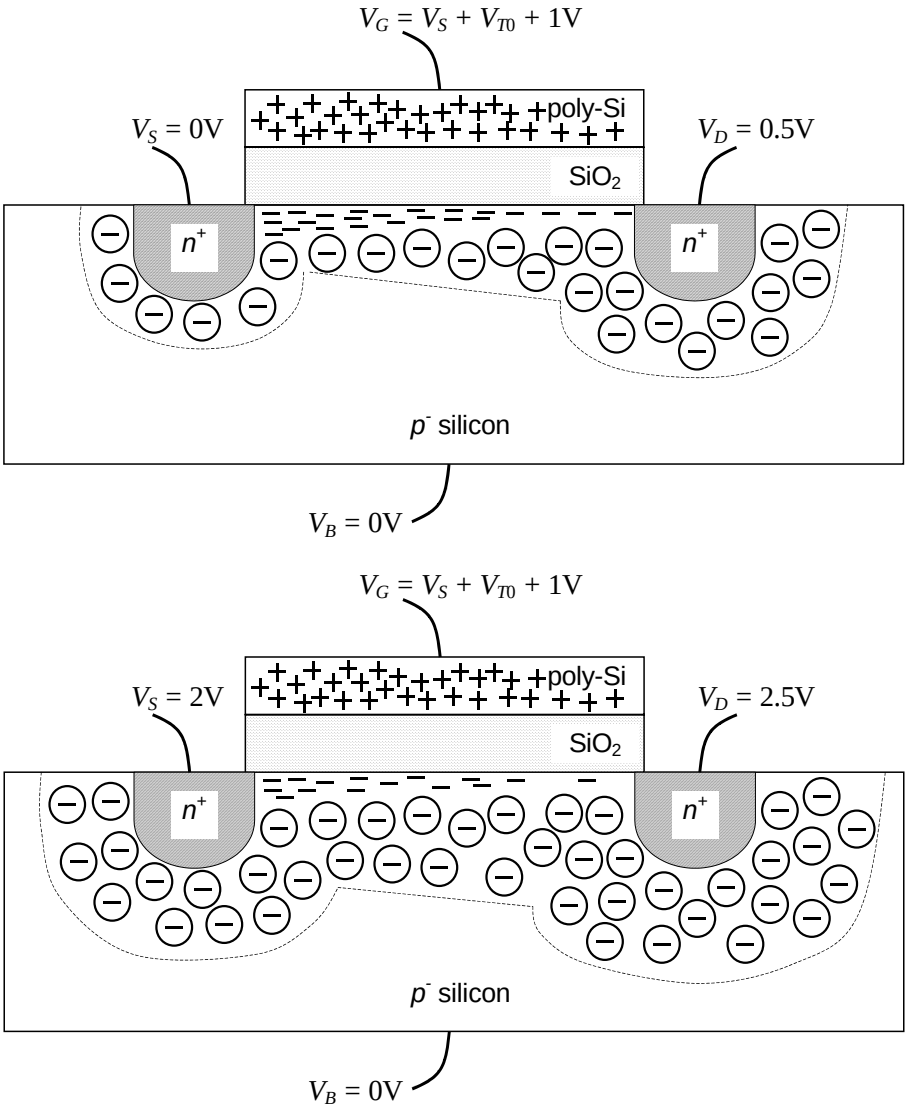


The "current source" behavior of a MOSFET in the saturation region is apparent from the fluid model. The "waterfall" into the drain corresponds to a current source. Why does the current depend on the *square* of the effective gate voltage? Because when we increase the gate-to-source voltage, we affect the current in two ways: we increase the channel charge and we increase the electric field in the channel. (The slope of the water increases.)

Notice that much of our voltage supply can be used up just keeping MOSFETs in the active region!

The Body Effect

What happens if the source is not at the same potential as the substrate? The figure below shows the situation as we keep V_{DS} constant but raise $V_S > 0$.



Now that the source and drain are more strongly reverse-biased with respect to the substrate, the depletion region widens and the depletion charge Q_B increases. Now that there is more channel charge from the depletion region, there need not be as must inversion layer charge (Q_I) to balance the gate charge Q_G. Thus, the inversion layer weakens and the current drops.

This effect can be modeled as a slight increase in the threshold voltage V_T. The depletion charge is given by:

$$Q'_B = -\sqrt{2q\varepsilon N_A}\sqrt{\psi_s}$$

The surface potential ψ_s can be expressed in terms of the resting surface potential ϕ_0 and the source-to-bulk potential V_{SB}:

$$\psi_s = \phi_0 + V_{SB} = |2\Phi_F| + V_{SB}$$

where

$$\Phi_F = \frac{kT}{q} \ln \frac{N_{sub}}{n_i}$$

where N_{sub} is the substrate doping level. Typical values of $|2\Phi_F|$ range between 0.7V and 0.9V, and are denoted "PHI" in SPICE models.

The "correction factor" in the threshold voltage is given by:

$$V_T = V_{T0} + \gamma\left(\sqrt{|2\Phi_F| + V_{SB}} - \sqrt{|2\Phi_F|}\right)$$

where

$$\gamma = \frac{\sqrt{2q\varepsilon N_{sub}}}{C'_{ox}}$$

Typical values of γ range from $0.4V^{1/2}$ to $0.8V^{1/2}$. Notice that both γ and $|2\Phi_F|$ increase with doping levels. Smaller, more advanced VLSI processes tend to be more heavily doped, so the body effect is becoming more significant.

Example: For nMOS transistors in a particular VLSI fabrication process, γ = $0.45V^{1/2}$ and $|2\Phi_F|$ = 0.9V. The basic threshold voltage V_{T0} = 0.70V. What is the threshold voltage for V_{SB} = 0, 1, 2, 3, and 4V?

$V_T(V_{SB} = 0\text{V}) = V_{T0} = 0.70\text{V}$

$V_T(V_{SB} = 1\text{V}) = 0.89\text{V}$

$V_T(V_{SB} = 2\text{V}) = 1.04\text{V}$

$V_T(V_{SB} = 3\text{V}) = 1.16\text{V}$

$V_T(V_{SB} = 4\text{V}) = 1.27\text{V}$

Another way to think of the body effect is to imagine the substrate as a "back gate" that modulates the channel in the same way that the normal "front" gate does. The front gate acts thorough the oxide capacitance, and the substrate, or back gate, operates through the depletion capacitance. Since the substrate, or **body**, will always have a negative voltage with respect to the source, it acts in opposition to the front gate. Since the depletion capacitance is smaller than the oxide capacitance, the back gate has a weak effect on the channel.

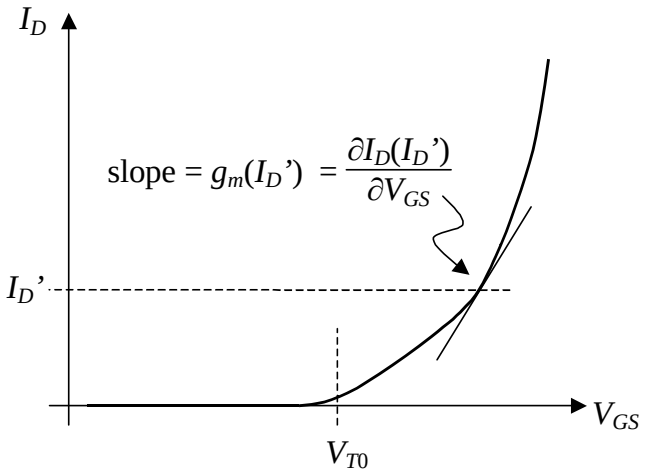
Small-Signal Parameters

The **large-signal equations** that describe MOSFET behavior are nonlinear. A circuit represents a system of nonlinear equations. Many times we are interested in building linear circuits, like filters that select certain frequencies but do not distort signals. How do we get linear behavior out of nonlinear devices?

If we zoom in on a nonlinear curve, it will eventually begin to look linear over a small range. This is similar to performing a Taylor expansion around a certain point on the curve, and then only keeping the first-order term. If we **bias** a transistor around an certain **operating point** and make only small perturbations about this point, we can approximate the device as a linear system. This is called **small-signal analysis**, because the amplitude of our signals must remain small for this analysis to remain valid.

The most important **small-signal parameter** of a MOSFET is its **transconductance**, g_m. If we make a *small* change in gate voltage, the transconductance tells us how much that affects the drain current:

$$g_m \equiv \frac{\partial I_D}{\partial V_{GS}}$$



We can express the transconductance of an above-threshold MOSFET in saturation in three equivalent ways:

In strong inversion:

$$g_m = \mu C'_{ox} \frac{W}{L} V_{eff}$$

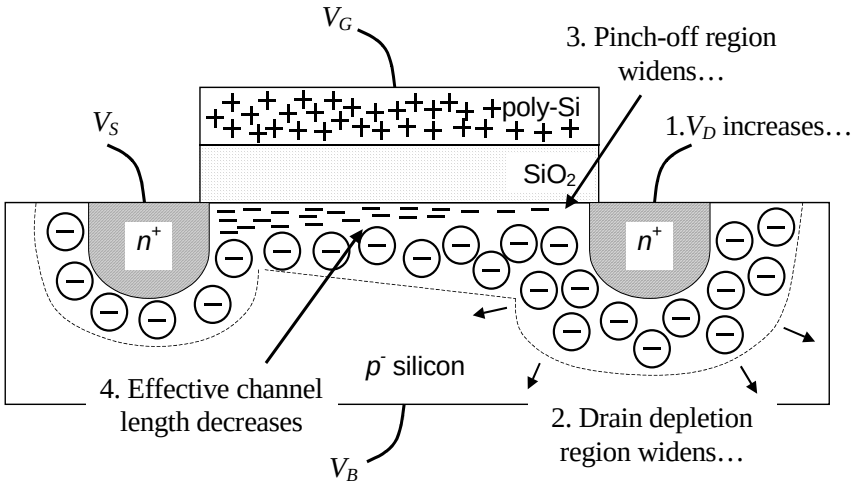
$$g_m = \sqrt{2\mu C'_{ox} \frac{W}{L} I_D}$$

$$g_m = \frac{2I_D}{V_{eff}}$$

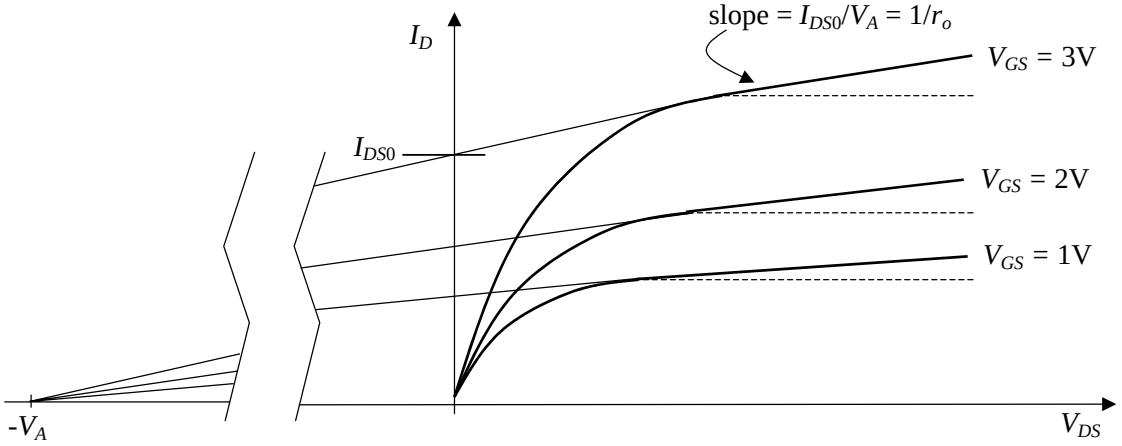
Channel-Length Modulation

When we discussed pinch-off and the saturation region, we said that after pinch-off occurred, the drain current was constant. (Remember the waterfall in the fluid model?) Well, that's not quite true due to an effect called **channel-length modulation**.

As we increase the drain voltage in the saturation region, the drain depletion region widens. This has the effect of widening the pinch-off region and thus shrinking the channel by a small amount.



Since the current in a MOSFET is proportional to W/L, a shrinking channel length increases the current through the device. In "long-channel" devices (which usually means at least 3 times the minimum length allowed by the process), the slight increase in current as the drain voltage increases is nearly linear. If we plot I_D vs. V_{DS} for several values of V_{GS} and extrapolate all the current traces backwards, they tend to converge at the same point on the V_{DS} axis. This voltage is called the **Early voltage** (V_A), and is the characteristic voltage in the first-order model of channel length modulation.



We can thus modify our expression for current in the saturation region by adding an extra term:

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}\left(V_{GS}-V_T\right)^2\cdot\left(1+\frac{V_{DS}}{V_A}\right)$$

We can define a small-signal parameter called **output resistance** to describe this effect:

$$r_o \equiv \frac{\partial V_{DS}}{\partial I_D} = \frac{V_A}{I_D}$$

Since longer channels are affected less severely by drain depletion region widening, they show behavior closer to that of an ideal current source:

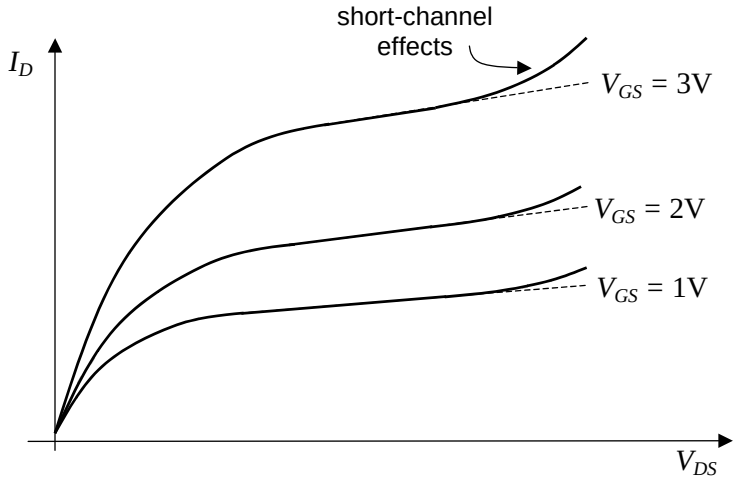
$$V_A, r_o \propto L \text{ for long-channel devices}$$

Some people use the **channel length modulation coefficient** λ instead of V_A:

$$I_D = \frac{1}{2}\mu_n C'_{ox}\frac{W}{L}\left(V_{GS}-V_T\right)^2\cdot\left(1+\lambda V_{DS}\right)$$

where $\lambda = 1/V_A$.

In modern **short-channel devices** (lengths near the minimum length allowed by the process), the effect is more severe. Several additional effects contribute to drain current increasing faster than linearly. One of these effects is called **drain-induced barrier lowering**, or **DIBL** ("dibble"). In very short devices, the drain can act as an additional gate ("side gate?"), coupling to the channel through the drain depletion capacitance. This can lead to dramatic increases in drain current (and dramatic decreases in r_o) as the drain voltage is increased.



Most models of channel-length modulation are **empirical models**. They basically fit the observed behavior with first-order or higher-order models. The physical details – even in long-channel devices – are complex and difficult to model reliably.

In other words, the Early voltage is mostly a hack to fit the data.

Transistor Summary

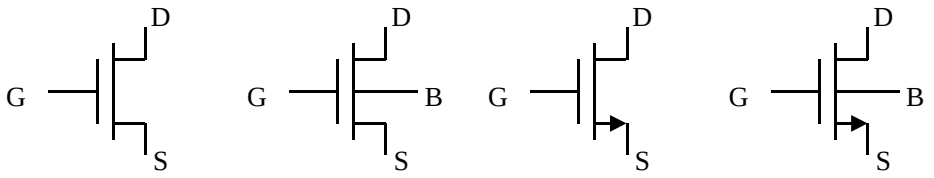
Strong Inversion

Parameter	Expression
Drain current – triode region	$I_D = \mu_n C'_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right]$ where $V_T = V_{T0} + \gamma\left(\sqrt{\lvert 2\Phi_F\rvert + V_{SB}} - \sqrt{\lvert 2\Phi_F\rvert}\right)$
Drain current – deep triode region (linear region)	$I_D \approx \mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)V_{DS}$
Drain current – saturation region	$I_D = \frac{1}{2}\mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)^2$
Condition for saturation	$V_{DS} \geq V_{GS} - V_T$
Drain current – saturation region (with channel-length modulation)	$I_D = \frac{1}{2}\mu_n C'_{ox} \frac{W}{L}(V_{GS} - V_T)^2 \cdot \left(1 + \frac{V_{DS}}{V_A}\right)$
Transconductance	$g_m = \sqrt{2\mu C'_{ox} \frac{W}{L} I_D}$ $= \mu C'_{ox} \frac{W}{L}(V_{GS} - V_T)$ $= \frac{2I_D}{V_{GS} - V_T}$
Output resistance	$r_o = \frac{V_A}{I_D}$

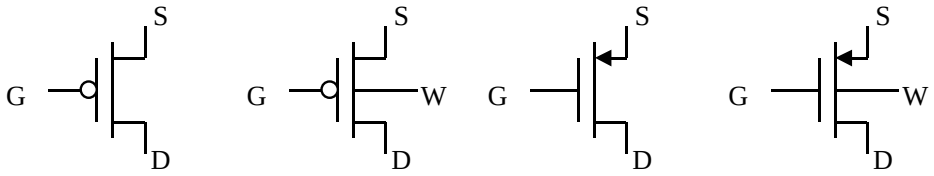
For a *p*MOS device with well voltage V_W, replace V_G, V_S, and V_D with $V_W - V_G$, $V_W - V_S$, and $V_W - V_D$ respectively.

The following circuit symbols are commonly used to represent nMOS and pMOS transistors in circuit diagrams. In this class we will use the first symbols, but the book uses the later symbols (with the arrows). The arrows make a rather artificial distinction between the source and drain in a completely symmetric device.

_n_FET symbols



_p_FET symbols



(G = gate; S = source, D = drain; B = bulk; W = well)