Matthew Schwartz
Statistical Mechanics, Spring 2025

Lecture 14: Semiconductors

1 Introduction

We saw how the Sommerfeld free electron model can explain many properties of metals. It postulates that to a leading approximation, electrons in metals roam freely in the metal like particles in a box. These electrons are fermions so this is an example of a free-electron gas. The free electron gas gives a number of very reasonable predictions. It predicts Fermi energies for metals in the $\sim$10 eV range and Fermi temperatures in the $\sim$50,000 K range. Thus Fermi gases in metals are highly degenerate, with very few states excitable at room temperature.

We used the free electron model to show that the heat capacity of electrons in metals is linear in T. Thus electrons dominate over the phonon contribution to C_V at low T but at high T, the electrons' contribution is negligible. The low T prediction is in agreement with data and the high T prediction explains why the electronic contribution can be ignored in the law of Dulong and Petit. The free electron model also gives very good predictions for the bulk modulus of metals.

In computing the heat capacity of metals we noted that there is an alternative language for discussing properties of degenerate Fermi gases where instead of thinking of all the energy levels as being excitations of the ground state, we think of the excited electrons as excitations above the Fermi level ε_F and the states that get excited as holes below the Fermi level. Although the holes are the absence of electrons, an excitation of a hole gives a positive contribution to the energy of the gas. Thus we can think of a Fermi gas as a collection of excitations and holes. This picture gives a powerful way to understand properties of semiconductors, as we will see in this lecture.

We also saw how the nearly free electron model, where the free theory is perturbed with a weak periodic potential leads to the emergence of bands. The band structure is crucial to understanding properties of real metals. Unfortunately, the free-electron model is very bad at incorporating differences among elements. A much more flexible approach is the **tight-binding model** which we explore in this lecture. The tight-binding model constructs allowed electronic states by combining atomic orbitals, similar to molecular orbital theory. After explaining the relevant concepts, we will use the tight-binding model to understand one of the most important technological innovations of the 20th century: semiconductors. Semiconductors are essential to every aspect of everyday life: they allow for very efficient and powerful computers. In Section 3 we use the band picture developed from the tight-binding model to understand how pn junctions and transistors work, and then discuss how transistors lead to computers.

Although this lecture seems long, it has very few equations, so it should be relatively quick to read. In fact, the lecture is really two lectures in one: Section 2 explains how to understand why different elements form different kinds of solids. The rest of the lecture explains why their band structure gives semiconductors such amazing properties. Section 2 is a bit more chemistry than physics, but it is important science that you should know. That being said, if you already have a qualitative understanding of the electronic properties of different elements then by all means feel free to skip it. In any case, please try to understand the material in this lecture, even if it takes multiple passes at the reading. As this lecture synthesizes chemistry, physics and technology, I think it contains a good amount material that could stick with you after the course is over.

2 The periodic table

To understand metals and semiconductors, we need a better understanding of the electron orbitals in elements than you might have gotten from your intro quantum mechanics class. In this section we'll first review the hydrogen atom then describe how to generalize to the other elements.

2.1 Hydrogen atom review

In quantum mechanics, you (hopefully) solved the Schrödinger equation to find the energy states of the hydrogen atom. These satisfy $\left(-\frac{1}{2m}\vec{\nabla}^2 - \frac{e^2}{4\pi\epsilon_0 r}\right)\psi_{nm\ell\sigma} = \varepsilon_n\psi_{nm\ell\ell\sigma}$. The eigenfunctions are separable: $\psi_{n\ell m\sigma} = R_{n\ell}(r)Y_{\ell m}(\theta, \phi)$ with $R(r)$ given by Laguerre polynomials and $Y_{\ell m}(\theta, \phi)$ are spherical harmonics. The energy levels depend only on the principle quantum number $n \geqslant 1$: $\varepsilon_n = -\mathrm{Ry}\frac{1}{n^2}$ where $\mathrm{Ry} = \frac{m_e e^4}{8h^2\varepsilon_0^2} = 13.6\mathrm{eV}$ is the Rydberg constant. The other quantum numbers are the the angular momentum ℓ and the projection m of angular momentum on the z axis. ℓ is a whole number from 0 to $n-1$ and $m = -\ell, -\ell+1, \cdots, \ell$. Thus there are $n-1$ values of ℓ for every n and $2\ell+1$ values of m for every ℓ. The final quantum number is the spin $\sigma = \pm\frac{1}{2}$. The energy levels of hydrogen don't depend on ℓ, m or σ, only on n.

The quantum number ℓ gives the shape of the orbital, and m its orientation. We associate letters to ℓ values: $\ell = 0$ is the letter s, $\ell = 1$ is the letter p, $\ell = 2$ is the letter d, $\ell = 3$ is the letter f. (By the way, these letters originated from properties of associated spectral lines: sharp, principle, diffuse and fundamental.) The s orbitals are spherically symmetric. There is only one s orbital for each n since $2\ell+1 = 1$ when $\ell = 0$. The p orbitals have two lobes, one with $\psi < 0$ and one with $\psi > 0$. There are three p orbitals for each n since $2\ell+1 = 3$ when $\ell = 1$. The three p orbitals have different orientations, with the lobes pointing in the x, y or z direction. There are five d orbitals and seven f orbitals. Here are some orbital shapes of the s, p and d orbitals

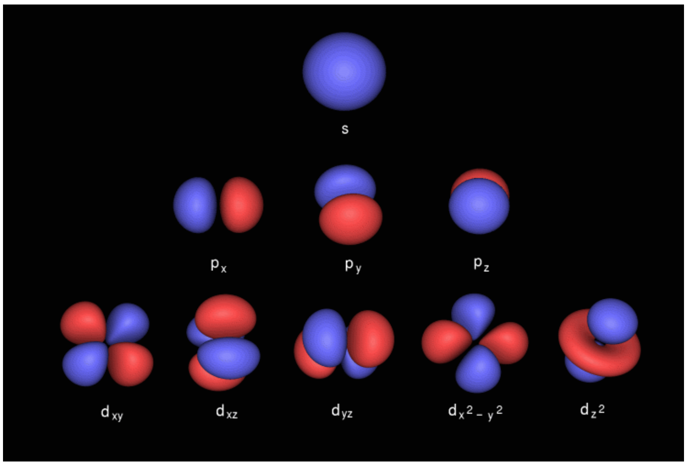


Figure 1. The shape of hydrogen atom orbitals are determined by spherical harmonics. These plots are like radiation patterns: the distance from the origin at the angle θ, ϕ is given by $Y_{\ell m}(\theta, \phi)$ with the color denoting the sign.

The radial wavefunction is determined mostly by the principle quantum number n. For $n = 1$, $R_{10}(r) \sim \exp\left(-\frac{r}{a_0}\right)$ with $a_0 = \frac{\hbar}{m_e c\alpha} = 5 \times 10^{-11}m$ the Bohr radius. Thus the size of the hydrogen atom itself is around a_0. With $n = 2$, $R_{20}(r) \sim \left(2 - \frac{r}{a_0}\right)\exp\left(-\frac{r}{2a_0}\right)$. So the $n = 2$ wavefunction also dies exponentially, but has a node at $r = 2\alpha_0$ after which it flips sign. In general, the radial dependence looks like $R_{n\ell}(r) \sim \exp\left(-\frac{r}{na_0}\right)$. So for bigger n the orbitals get bigger and bigger, with the atomic radius scaling as $\langle r\rangle \sim na_0$. The ℓ dependence of the radial wavefunctions is subleading – it doesn't affect the exponential, only a polynomial prefactor, e.g. $R_{21} \sim r\exp\left(-\frac{r}{2a_0}\right)$.

The spin σ doesn't affect the shape of the wavefunction. The only relevant fact is that since σ has two possible values, there are two degenerate orbitals (same energy) at each $n\ell m$. Thus for $n = 1$ there are two possible states, $1s^1$ and $1s^2$. For $n = 2$ there are two $2s$ states and six $2p$ states. For $n = 3$ there are ten $3d$ states, and so on.

2.2 Hydrogenic atoms

That may be where you left off in quantum mechanics. What do the orbitals of other elements look like? If there is only one electron and the nucleus has charge Z then we have the exact solution: the potential is Z times as big and states for the electron have energies $\varepsilon_n = -\mathrm{Ry}\,\frac{Z^2}{n^2}$ and sizes $\langle r\rangle \sim \frac{n}{Z}\alpha_0$. The challenge comes when there is more than one electron, since the previous electrons screen the nuclear charge, and can do so asymmetrically. The screening also breaks the degeneracy of the energy levels and distorts the wavefunctions away from the hydrogenic form.

To a first approximation, we can treat the orbitals of the additional electrons as similar to those of one-electron atoms. In fact, the orbital shapes for the electrons in multiple-electron atoms are often similar to those of one-electron atoms, even if the energies of those orbitals are very different. Thus we label the multi-electron atoms using the notation for hydrogen-atom orbitals. For example, we write boron as $1s^2 2s^2 2p^1$, meaning the $n=1, \ell=0$ orbital has two electrons in it (the $1s^2$ part), the $n=2, \ell=0$ orbital has two electrons, and the $n=2, \ell=1$ orbital has 1 electron.

Now consider carbon $(Z=6)$ which has another electron to come in. It will then be $1s^2 2s^2 2p^2$. But are the two electrons going to be the same p orbital with different spins, or different p orbitals? To figure out the answer, suppose the first $2p$ electron is in a p_x orbital. Since its electron cloud points in the x direction, the nuclear charge in x will be screened more in x than in the y or z directions. This means that the p_y and p_z orbitals will have lower energy than the other-spin p_x orbital, since they are less screened. So after p_x then either p_y or p_z gets filled. That is, for carbon, the two $2p$ electrons will have different values of m. In nitrogen, $1s^2 2s^2 2p^3$, the three $2p$ orbitals will be in the three different directions, p_x, p_y and p_z. It is not until oxygen, $1s^2 2s^2 2p^4$, that a second electron goes into the p_x state. Note here that x, y and z are arbitrary, the point is only that different m states are filled before two spin states go in to the same m. This effect leads to **Hund's rule:** every orbital is singly occupied with one electron before any is doubly occupied. When all the orbitals of a given ℓ are filled, we say the **shell** (all the m and s values for a given ℓ) is **closed**.

To understand the periodic table, we need more than Hund's rule. We have to go beyond the hydrogenic atom spectrum and understand how the electron screening moves the energy levels up and down. It turns out to be very complicated. The energy levels as a function of Z look like

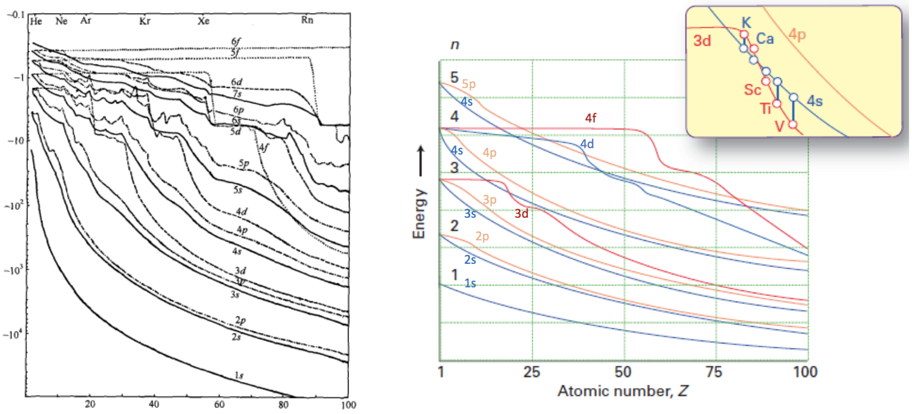


Figure 2. Energies of orbitals as a function of atomic number, including electron screening effects. Left is more realistic, from an old book, and right is more of a cartoon. Note that the 3d orbitals have energies above the 4s orbitals around $Z=19$ (potassium).

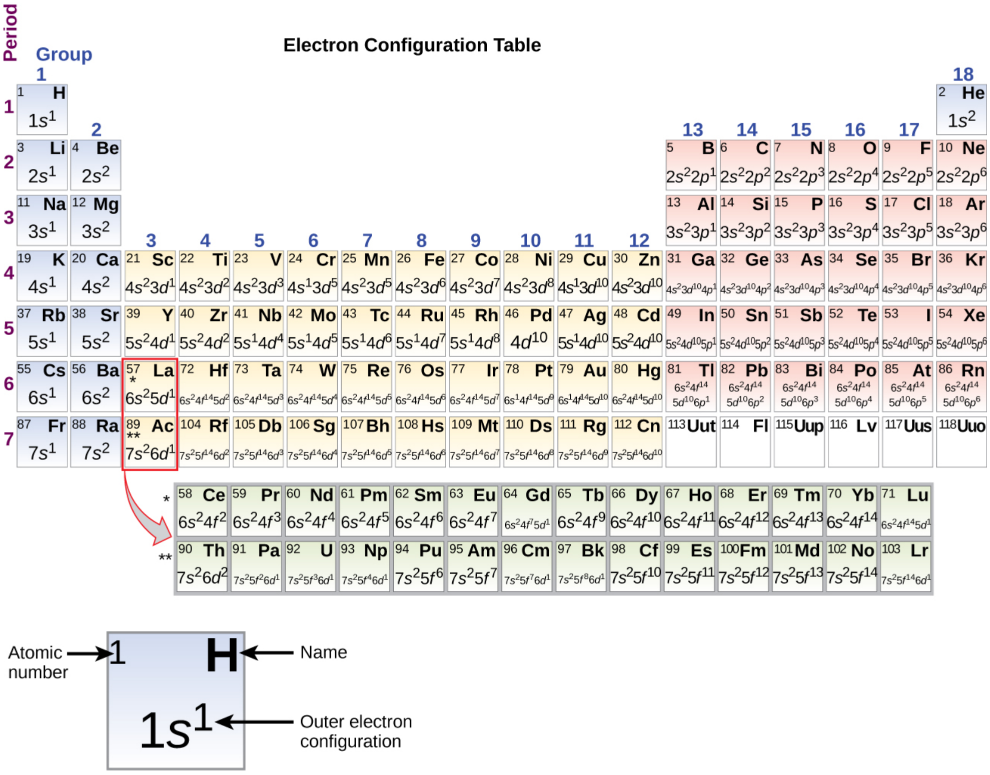
We see a few things from this plot. First of all, there is always a large gap between the p orbitals and the s orbitals in the shell above it. This gap is very important. It makes the elements with filled p orbitals and no extra electrons very stable. Such elements are called the noble gases and sit at the far right of the periodic table. Note that the noble gases do not have *all* the orbitals filled for a given n filled, just all the p orbitals filled for a given n.

The first row of the periodic table fills $n=1$. The second row fills $n=2$. The third row fills $n=3$, $\ell\leqslant 1$, but the $3d$ orbitals do not get filled until the 4th row. After argon, which is [Ne]$3s^2 3p^6$ with [Ne] meaning all the closed shells of neon (Ne=$1s^2 2s^2 2p^6$), one might think that $3d^1$ is next. Indeed, from the hydrogen atom, we know that the $3d$ orbitals have lower energy than the $4s$ orbitals, since the energy only depends on n. However, we see from Fig. 2 that by the time $n=4$, the degeneracy associated with the principle quantum number is pretty badly broken: the $4s^1$ level has lower energy than the $3d^1$ level. The result is that the two 4s orbitals fill first (K and Ca), then the ten $3d$ orbitals get filled (transition metals Sc to Zn), then the 4p orbitals fill until krypton, Kr$=$[Ar]$4s^2 3d^{10} 4p^6$. In other words, to a reasonable approximation, the energy levels are grouped not quite by n but in sets

- $(1s)$, $(2s2p)$, $(3s3p)$, $(4s4p3d)$, $(5s5p4d)$, $(6s6p\,4f\,5d)$, $\cdots$

The filling of the levels in these groups gives the "periodicity" of the periodic table. Each group is a row.

Here is a periodic table showing the valence electron configurations (valence here means the electrons outside of the last nobel gas)"


Electron Configuration Table

Period	Group 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H $1s^1$																	2 He $1s^2$
2	3 Li $2s^1$	4 Be $2s^2$											5 B $2s^2 2p^1$	6 C $2s^2 2p^2$	7 N $2s^2 2p^3$	8 O $2s^2 2p^4$	9 F $2s^2 2p^5$	10 Ne $2s^2 2p^6$
3	11 Na $3s^1$	12 Mg $3s^2$											13 Al $3s^2 3p^1$	14 Si $3s^2 3p^2$	15 P $3s^2 3p^3$	16 S $3s^2 3p^4$	17 Cl $3s^2 3p^5$	18 Ar $3s^2 3p^6$
4	19 K $4s^1$	20 Ca $4s^2$	21 Sc $4s^2 3d^1$	22 Ti $4s^2 3d^2$	23 V $4s^2 3d^3$	24 Cr $4s^1 3d^5$	25 Mn $4s^2 3d^5$	26 Fe $4s^2 3d^6$	27 Co $4s^2 3d^7$	28 Ni $4s^2 3d^8$	29 Cu $4s^1 3d^{10}$	30 Zn $4s^2 3d^{10}$	31 Ga $4s^2 3d^{10} 4p^1$	32 Ge $4s^2 3d^{10} 4p^2$	33 As $4s^2 3d^{10} 4p^3$	34 Se $4s^2 3d^{10} 4p^4$	35 Br $4s^2 3d^{10} 4p^5$	36 Kr $4s^2 3d^{10} 4p^6$
5	37 Rb $5s^1$	38 Sr $5s^2$	39 Y $5s^2 4d^1$	40 Zr $5s^2 4d^2$	41 Nb $5s^1 4d^4$	42 Mo $5s^1 4d^5$	43 Tc $5s^1 4d^6$	44 Ru $5s^1 4d^7$	45 Rh $5s^1 4d^8$	46 Pd $4d^{10}$	47 Ag $5s^1 4d^{10}$	48 Cd $5s^2 4d^{10}$	49 In $5s^2 4d^{10} 5p^1$	50 Sn $5s^2 4d^{10} 5p^2$	51 Sb $5s^2 4d^{10} 5p^3$	52 Te $5s^2 4d^{10} 5p^4$	53 I $5s^2 4d^{10} 5p^5$	54 Xe $5s^2 4d^{10} 5p^6$
6	55 Cs $6s^1$	56 Ba $6s^2$	57 La * $6s^2 5d^1$	72 Hf $6s^2 4f^{14} 5d^2$	73 Ta $6s^2 4f^{14} 5d^3$	74 W $6s^2 4f^{14} 5d^4$	75 Re $6s^2 4f^{14} 5d^5$	76 Os $6s^2 4f^{14} 5d^6$	77 Ir $6s^2 4f^{14} 5d^7$	78 Pt $6s^1 4f^{14} 5d^9$	79 Au $6s^1 4f^{14} 5d^{10}$	80 Hg $6s^2 4f^{14} 5d^{10}$	81 Tl $6s^2 4f^{14}$ $5d^{10} 6p^1$	82 Pb $6s^2 4f^{14}$ $5d^{10} 6p^2$	83 Bi $6s^2 4f^{14}$ $5d^{10} 6p^3$	84 Po $6s^2 4f^{14}$ $5d^{10} 6p^4$	85 At $6s^2 4f^{14}$ $5d^{10} 6p^5$	86 Rn $6s^2 4f^{14}$ $5d^{10} 6p^6$
7	87 Fr $7s^1$	88 Ra $7s^2$	89 Ac ** $7s^2 6d^1$	104 Rf $7s^2 5f^{14} 6d^2$	105 Db $7s^2 5f^{14} 6d^3$	106 Sg $7s^2 5f^{14} 6d^4$	107 Bh $7s^2 5f^{14} 6d^5$	108 Hs $7s^2 5f^{14} 6d^6$	109 Mt $7s^2 5f^{14} 6d^7$	110 Ds $7s^2 5f^{14} 6d^8$	111 Rg $7s^2 5f^{14} 6d^9$	112 Cn $7s^2 5f^{14} 6d^{10}$	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo

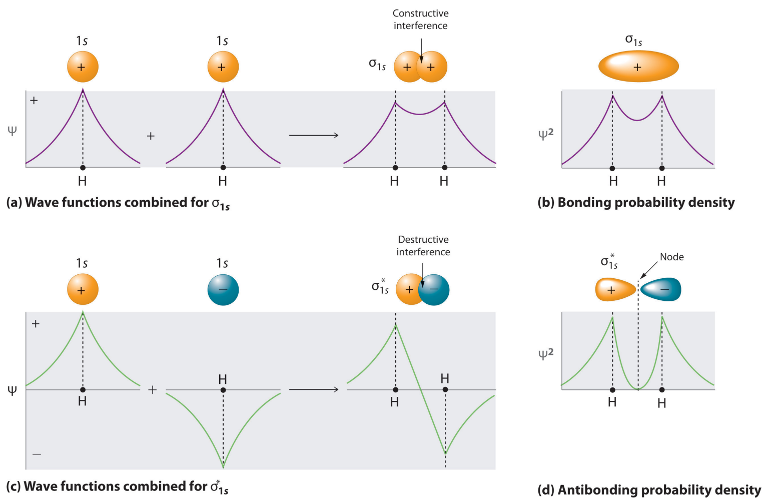
*	58 Ce $6s^2 4f^2$	59 Pr $6s^2 4f^3$	60 Nd $6s^2 4f^4$	61 Pm $6s^2 4f^5$	62 Sm $6s^2 4f^6$	63 Eu $6s^2 4f^7$	64 Gd $6s^2 4f^7 5d^1$	65 Tb $6s^2 4f^9$	66 Dy $6s^2 4f^{10}$	67 Ho $6s^2 4f^{11}$	68 Er $6s^2 4f^{12}$	69 Tm $6s^2 4f^{13}$	70 Yb $6s^2 4f^{14}$	71 Lu $6s^2 4f^{14} 5d^1$
**	90 Th $7s^2 6d^2$	91 Pa $7s^2 5f^2 6d^1$	92 U $7s^2 5f^3 6d^1$	93 Np $7s^2 5f^4 6d^1$	94 Pu $7s^2 5f^6$	95 Am $7s^2 5f^7$	96 Cm $7s^2 5f^7 6d^1$	97 Bk $7s^2 5f^8 6d^1$	98 Cf $7s^2 5f^{10}$	99 Es $7s^2 5f^{11}$	100 Fm $7s^2 5f^{12}$	101 Md $7s^2 5f^{13}$	102 No $7s^2 5f^{14}$	103 Lr $7s^2 5f^{14} 6d^1$

Figure 3. Periodic table showing valence electron configurations

2.3 Molecular orbital theory

Now that we understand the periodic table, based on orbitals of atoms in isolation, let's talk about bonds. Why do chemical bonds form?

Recall from quantum mechanics that if we have a potential well, there will be a bound state with some energy ε_0 and wavefunction $\psi(x)$. If we put two of these wells next to each other, then instead of two states with energy ε_0, one state will have energy $\varepsilon_-=\varepsilon_0-\Delta$ and the other energy $\varepsilon_+=\varepsilon_0+\Delta$. The state with lower energy is the symmetric combination $\psi_+=\frac{1}{\sqrt{2}}(\psi_1+\psi_2)$ and the state with high energy is the antisymmetric combination $\psi_-=\frac{1}{\sqrt{2}}(\psi_1-\psi_2)$. This is a very generic feature of quantum mechanical systems: when two systems are brought together, their energies will split, with some going lower and and some going higher.[1]



(a) Wave functions combined for σ_{1s}

(b) Bonding probability density

(c) Wave functions combined for σ^*_{1s}

(d) Antibonding probability density

Figure 4. When two 1D systems are brought together, the energy eigenstates are approximately the sum and difference of the two separate energy eigenstates.

The same thing happens when we bring two atoms together. Consider two hydrogen atoms. Each separately has two 1s orbitals (the two spin states) with energies ε_0. When we bring them together, two states get higher in energy and two get lower in energy, so their new energies are $\varepsilon_\pm = \varepsilon_0 \pm \Delta$ just like in 1D quantum mechanics. We call the lower energy states the **bonding orbitals** and the higher energy states the **antibonding orbitals**. This idea of constructing orbitals for the molecules by taking linear combinations of the atomic orbitals is called **molecular orbital theory**. It works pretty well and can explain a lot of chemistry.

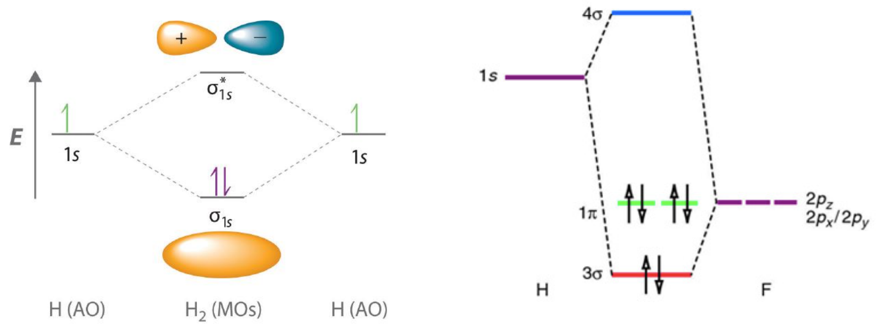


Figure 5. Left shows H_2. When the two H are apart, each has one electron (green arrows) in a 1s orbital. When they are close, the orbitals combine into a bonding (σ_{1s}) and antibonding ($\sigma^\star_{1s}$) orbital. The two elecrons go into the bonding orbital (purple arrows) since it has lower energy than $1s$. Right shows HF: when different elements bind, the molecular orbitals are asymmetric combinations of the two atomic orbitals.

So for two H atoms, there are 2 total electrons and the energy is lowered if both electrons are shared in the bonding orbitals. This is a **covalent bond**. On the other hand, if we bring two helium atoms together, with a total of 4 electrons, the two bonding orbitals would get filled with two electrons, but then the other two electrons must go into the antibonding orbitals. This is not energetically favorable, so it doesn't happen: He doesn't bond with itself.

[1] $\varepsilon_\pm = \langle\psi_\pm|H|\psi_\pm\rangle = \frac{1}{2}[\langle\psi_1|H|\psi_1\rangle + \langle\psi_2|H|\psi_2\rangle] \pm \frac{1}{2}[\langle\psi_1|H|\psi_2\rangle + \langle\psi_2|H|\psi_1\rangle] = \varepsilon_0 \mp \Delta$ with $\Delta = -\mathrm{Re}[\langle\psi_1|H|\psi_2\rangle]$.

For fluorine $F = 1s^2 2s^2 2p^5$, consider the three $2p$ orbitals with 5 electrons. If all three $2p$ orbitals are shared among an F_2 molecule, there would be 3 bonding and 3 antibonding orbitals to hold 10 total electrons. Thus 6 go into bonding and 4 into antibonding. Filling antibonding orbitals is not energetically favorable. So instead, each atom in F_2 leaves four of its five $2p$ electrons in their unshared orbitals and only shares one valence electron each. Then the problem is reduced to sharing a single orbital, say p_z, and one electron from each F can go into the bonding orbital. F_2 forms covalent bonds this way.

What about when two different atoms combine? The valence electrons (most weakly bound electrons) in each atom are the last ones to be added. The valence electrons of different elements have different binding energies, so the analog is a 1D asymmetric double-well quantum mechanics problem with wells of different depths, i.e. different binding energies E_1 and E_2. The depth of each well is the binding energy, or equivalently the

- **Ionization energy**: the amont of energy required to remove an electron from a neutral atom.

When different elements are brought together, the eigenstates are again linear combinations $\psi_+ \sim c_1\psi_1 + c_2\psi_2$ and $\psi_- = c_1\psi_1 - c_2\psi_2$ but the linear combinations are not symmetric ($c_1 \neq c_2$). The bonding orbital will be more like the lower-energy atomic orbital, and the antibonding orbital will be more like the higher-energy atomic orbital (see the right diagram in Fig. 5). Consequently the electrons will be localized close to the element with the lower-energy orbital.

The ionization energies of the elements can be measured. They look like this:

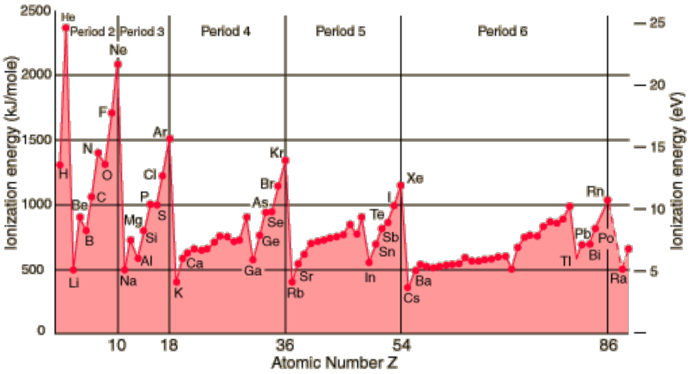


Figure 6. Ionization energies of the various elements

We see from this figure that the noble gases have the highest ionization energies so they are the most stable. Elements with one electron above a filled p orbital, namely those in the first column of the periodic table (H and the **Alkali metals**, Li, Na, K Rb) have low ionization energies: they want to give up an electron. The valence electrons are very weakly bound to these elements. On the other hand, elements that are one electron short of a filled p orbital (the **halogens**, F, Cl, Br, I) would become very stable if an electron were added. Their valence electrons are tightly bound and will not be given up easily. These have high ionization energies (but not as high as the noble gases which really really do not want to give up electrons).

So now if we bring an Alkali metal like sodium Na with binding energy $\varepsilon_1 \sim 5\text{eV}$ together with a halogen like Cl, with binding energy $\varepsilon_2 \sim 13\text{eV}$, the symmetric bonding molecular orbital will have energy $\varepsilon_- < \varepsilon_1$ and be localized near the Cl, while the antisymmetric anti-bonding molecular orbital will have energy $\varepsilon_+ > \varepsilon_2$ and be localized near Na. The two electrons will therefore both go into the bonding orbital and live close to the Cl. This type of bond where an electron is essentially donated from one atom to another is called an **ionic bond**. From molecular orbital theory, an ionic bond is rather an extreme form of covalent bond than something qualitatively different.

So molecular orbital theory explains some basic rules of thumb in chemistry such as the **octet rule**: atoms tend to prefer to have 8 electrons in the valence shell (ignoring the d and f orbitals). For example, Na gives up an electron to get a full shell and Cl receives one to get a full shell. You sometimes see this written as $\mathrm{Na}\cdot + \cdot\ddot{\underset{\cdot\cdot}{\mathrm{Cl}}}{:}\rightarrow \mathrm{Na}\quad :\ddot{\underset{\cdot\cdot}{\mathrm{Cl}}}{:}$. The octet rule is due to the trend that ionization energies increase until a p orbital is filled. Note that even in ionic bonds atoms never completely give up their electrons. Cl is neutral and does not form a bound state with an additional electron. Atoms towards the left side of the periodic table have fewer than half an octet filled and need to bond with something on the right side. Atoms toward the right side of the periodic table have more electrons to share, so it's easier for them to form covalent bonds and stable molecules.

Two concepts closely related and ionization energy are

- **electron affinity**: the energy change when a neutral atom attracts an electron to become a negative ion.

For example $\mathrm{Cl} + e^- \to \mathrm{Cl}^-$ releasing 3.6 eV of energy. Thus chlorine has an electron affinity of 3.6 eV. In other words while ionization energy is the change when losing an electron, electron affinity is the change when an electron is added to form a negative ion.

The other concept is.

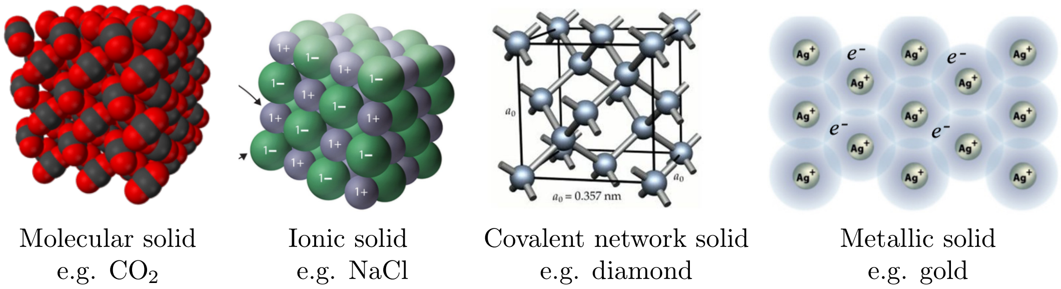
- **electronegativity**: how close an atom likes to pull bonding electrons towards itself

Electronegativity is vaguely defined and there are many competing definitions (such as the Pauling scale) that do not concern us. Qualitatively speaking, electronegativities are roughly proportional to ionization energies.

Elements on the far right (He, Ne, etc.) are noble gases. They have all filled shells and don't have any desire to lose or attract electrons. They have large ionization energies ($\sim$20eV for Ne) and no electron affinities at all since they cannot form anions. Noble gases don't bond, so electronegativity is not defined for them. Atoms next to them, the halogens (F, Cl, etc.), have high ionization energies ($\sim$12eV for Cl), high electron affinities ($\sim$3eV for Cl). They need one electron to fill a shell, so they desperately want it and have high electronegativities. Atoms on the far left (the alkali metals Li, Na etc.) will have filled orbitals if they lose one electron, so they have low ionization energies (5.3 eV for Li), no electron affinity, and low electronegativities. As you go from left to right in the table, the ionization energy, electron affinity and electronegativity increases.

2.4 Solids

We're almost back to statistical mechanics. The final concept from chemistry we need to understand is how to think about solids. Every element or compound forms a solid when cold enough. But these solids can have very different forms and properties depending on what the constituent atoms or molecules are. Almost all solids can be characterized as molecular, ionic, covalent-network or metallic:



Molecular solid
e.g. CO_2

Ionic solid
e.g. NaCl

Covalent network solid
e.g. diamond

Metallic solid
e.g. gold

Figure 7. Different types of solids

Molecules that are themselves quite stable, like CO_2 or H_2O or noble gases, have little interest in sharing electrons. Their solid forms are **molecular solids**: conglomerations of molecules with some regularity but no exactly periodic lattice structure. The solids are held together by various relatively weak attractive interactions, such as hydrogen bonds, dipole-dipole interactions, or dispersive forces. Ice is a molecular solid.

Elements that form ionic bonds, like table sald, NaCl, are happy sharing an electron pair between them. Their solid forms are **ionic solids**: the electronegative element (Cl for NaCl) draws the electron in close, leaving essentially pairs of Na^+ and Cl^- ions. These ions attract each other electrically holding the solid together.

Some elements, like carbon, like to form covalent bonds with their neighbors, leading to **covalent network solids**. The bonds in these solids are neighbor-to-neighbor and so the electrons are localized between the atoms, not distributed throughout the solid. A covalent network solid is in a sense one giant molecule. These solids are generally insulators, but some conduct. For example, diamond insulates, but graphite conducts. The difference is in the bonding structure. In diamond, each carbon is covalently bonded to 4 neighbors, tying up all its valence electrons in bonds and forming a rigid geometric lattice. This is what makes diamond so inflexible. In graphite, each carbon atom is bonded to only 3 neighbors, essentially in a plane, so 1 valence electron is mobile and can conduct electricity. The planes are only weakly attached to each other, which is why pencils work and graphene (2D graphite) exists.

The final solid form is the **metallic solid**. In a metallic solid, the electrons are loosely bound to the atoms and are easily shared among the atoms. It is these solids to which the free electron gas model applies.

Which elements form which type of solids? Let's go through elements from right to left on the periodic table. The noble gases are very stable and have no electron affinities, so they form molecular solids. Atoms which like to form diatomic molecules, such as the halogens (F_2, Cl_2, Br_2), oxygen O_2 and nitrogen N_2 will also form molecular solids, made out of these diatomic molecules. Carbon forms covalent networks, like diamond, graphite and graphene. Phosphorous is highly reactive and forms molecular solids with whatever it has reacted with. Sulfur forms octoatomic molecules S_8 that solidify. Selenium (Se) forms covalent networks. These are all the non-metallic elemental solids. Most of the rest of the periodic table have spare valence electrons and low electronegativities so they form elemental metallic solids. Elements on the boundary between metals and non-metals have intermediate properties. These are called **metalloids** and include the semiconductors:

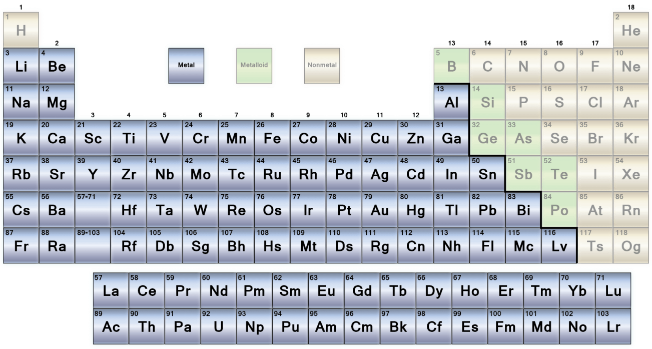

Metal
Metalloid
Nonmetal
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
1 H 2 He
3 Li 4 Be 5 B 6 C 7 N 8 O 9 F 10 Ne
11 Na 12 Mg 13 Al 14 Si 15 P 16 S 17 Cl 18 Ar
19 K 20 Ca 21 Sc 22 Ti 23 V 24 Cr 25 Mn 26 Fe 27 Co 28 Ni 29 Cu 30 Zn 31 Ga 32 Ge 33 As 34 Se 35 Br 36 Kr
37 Rb 38 Sr 39 Y 40 Zr 41 Nb 42 Mo 43 Tc 44 Ru 45 Rh 46 Pd 47 Ag 48 Cd 49 In 50 Sn 51 Sb 52 Te 53 I 54 Xe
55 Cs 56 Ba 57-71 72 Hf 73 Ta 74 W 75 Re 76 Os 77 Ir 78 Pt 79 Au 80 Hg 81 Tl 82 Pb 83 Bi 84 Po 85 At 86 Rn
87 Fr 88 Ra 89-103 104 Rf 105 Db 106 Sg 107 Bh 108 Hs 109 Mt 110 Ds 111 Rg 112 Cn 113 Nh 114 Fl 115 Mc 116 Lv 117 Ts 118 Og
57 La 58 Ce 59 Pr 60 Nd 61 Pm 62 Sm 63 Eu 64 Gd 65 Tb 66 Dy 67 Ho 68 Er 69 Tm 70 Yb 71 Lu
89 Ac 90 Th 91 Pa 92 U 93 Np 94 Pu 95 Am 96 Cm 97 Bk 98 Cf 99 Es 100 Fm 101 Md 102 No 103 Lr

Figure 8. Metals, non-metals and metalloids in the periodic table.

2.5 Tight-binding model

Just as molecules can be understood by combining the orbitals of separate atoms using molecular orbital theory, we can understand metallic solids by combining a regular array of atoms. This is known as the **tight-binding model**. The tight-binding model is basically molecular orbital theory applied to an array of atoms.

The analog quantum mechanics problem is a series of N particles-in-a-box or potential wells. When two wells with energies ε_0 are brought together, we get one eigenstate with lower energy and one with higher energy $\varepsilon_\pm = \varepsilon_0 \pm \Delta$. When three wells are brought together, the lowest energy lowers more, the highest energy raises more, the energies are roughly $\varepsilon_0 - 1.5\Delta, \varepsilon_0, \varepsilon_0 + 1.5\Delta$ and so on. Using the energy levels of the hydrogen atom, the basic picture looks like this

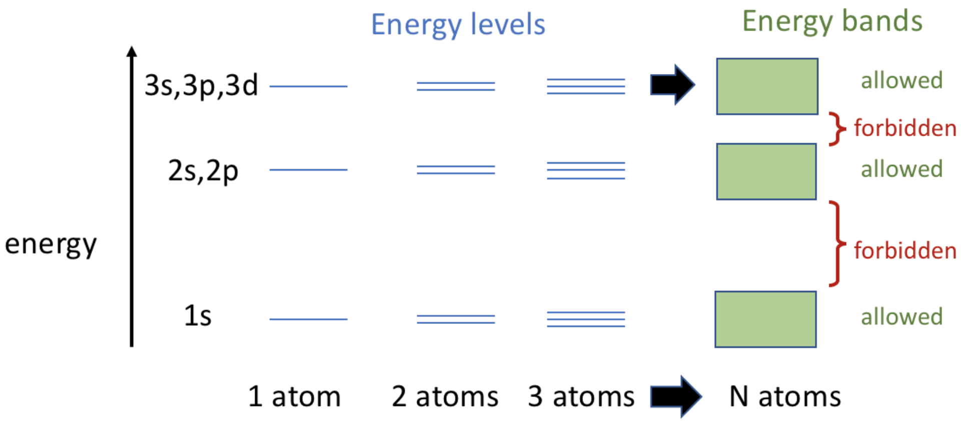


Figure 9. As orbitals mix, energy bands form

The discrete energy levels turn into **bands**. Half of each band comprises bonding orbitals with energies below the energy ε_0 of the isolated atom. The other half of each band comprises the antibonding orbitals with energies greater than ε_0. Every energy level is 2-fold degenerate because of the two electron spins. So if there are N atoms, there are $2N$ energy levels in each band. The number of bands is set by the number of original orbitals.

For elements above hydrogen, the $2p$ and $2s$ orbitals are not degnerate even for a single molecule (recall Fig. 2). Thus, when the orbitals combine, the center of the bands are offset but the bands may overlap. For example, in sodium: the bands look like

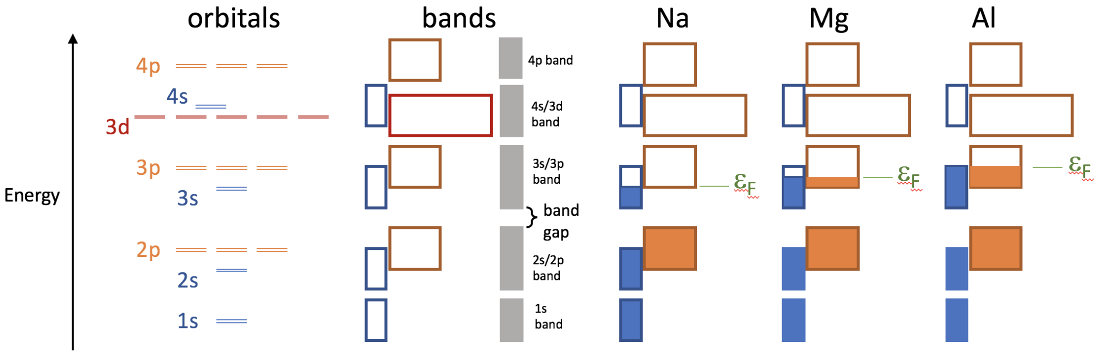


Figure 10. Bands of metals on the 3rd row of the periodic table. In Na the $3s$ / $3p$ band is $1/8$ filled. In Mg it is $1/4$ filled and in Al it is $3/8$ filled. In all three metals, the Fermi energy is within a band(not in a band gap).

Sodium is Na $= 1s^2 2s^2 2p^6 3s^1$, so each atom has one valence electron in the 3s orbital. When the bands form, the N valence electrons from the $3s^1$ orbitals get distributed over the N energy levels of the binding band. Thus the $3s$ band is half filled, as indicated.

Magnesium is $\mathrm{Mg} = 1s^2 2s^2 2p^6 3s^2$, you might think the $3s$ band would be completely filled, with the low energy and high energy states all populated, making Mg an insulator. However, we see from the figure that the $3p$ band overlaps the $3s$ band. So once the $3s$ band gets filled up to the level of the $3p$ band, the electrons start filling the $3p$ band. In this way the energy of magnesium is lower than the energy it would have if the $3s$ bands were completely filled. Thus there really is a hybrid $3s\,/\,3p$ band that is less than halfway filled for Mg.

Aluminium is $\mathrm{Al} = 1s^2 2s^2 2p^6 3s^2 3p^1$. It is similar to Mg, contributing 3 electrons per atom to the hybrid $3s\,/\,3p$ orbital that can hold 8 electrons total. Aluminum just has one more electron going into this hybrid band than magnesium.

You might think you can keep going this way. Silicon is $\mathrm{Si} = [\mathrm{Ne}]3s^2 3p^2$, so it could in principle fill the $3s\,/\,3p$ band up halfway. Halfway up isn't great though, because above halfway the anti-bonding orbitals would have to start being filled, and those have more energy than the free atoms, so it's not energetically favorable to fill them. Silicon has another option though. It has 4 electrons to share and needs 4 electrons to form a closed shell. So it can form covalent bonds with its 4 nearest neighbors. Thus instead of forming a metallic solid, Silicon forms a covalent network solid, like its upstairs neighbor carbon. Compared to carbon, silicon's electrons are farther out ($3s^2 3p^2$ rather than $2s^2 2p^2$), so they are relatively more weakly bound than in carbon. Silicon is conflicted: it is on the boundary between having metallic bonds and having covalent bonds. It is a metalloid.

To the right of silicon, the electrons would have to fill the anti-bonding part of the band. This is definitely unfavorable compared to forming covalent bonds. So the elements between Si and Ar, namely P, S and Cl are non-metallic. For example, the halogen at end of the row is chlorine $[\mathrm{Ne}]3s^2 3p^5$. If chlorine tried to share its 7 $s\,/\,p$ electrons in a metallic solid, the band would be half-filled after putting 4 electrons in, so the last 3 electrons would be in antibonding orbitals. On the other hand, if we just look at the singly-occupied $3p$ orbital, say $3p_z$ with a spin-up electron in it, this looks a lot like hydrogen with a single electron in a $1s$ orbital. It can form bonding orbitals using just this orbital, lowering the energy compared to separated chlorine atoms. Thus Cl forms diatomic molecules and molcular solids.

As you go down in the periodic table, the $3d$ and then $4d$ orbitals become relevant. These can hold a lot of electrons, so the bands are very wide and a lot of elements form metallic solids using these orbitals.

2.6 Summary

This section was a bit long, and it may not have been clear to you why all this chemistry was included in a physics class. The point was to understand why some elements form metals, some form insulators, and some form semiconductors (see next section). It is perhaps worth a quick summary of the main logical steps leading from QM to the classification of solids:

- The large degeneracy of energy levels of the hydrogen atom is broken in other elements. The closely spaced energies are in sets: $(1s)$, $(2s2p)$, $(3s3p)$, $(4s4p3d)$, $(5s5p4d)$, $(6s6p4f5d)$, $\cdots$
- When two atoms are brought together, the two valence orbitals from the atoms in isolation combine into a bonding orbital, with lower energy than both, and an anti-bonding orbital, with higher energy than both.
- When N orbitals combine, the many bonding and antibonding orbitals merge into bands, with the energy level of the isolated-atom orbitals falling right in the middle of the band.
- For elements on the left and in the center of the periodic table, it's energetically favorable to pool electrons, forming metals with the bonding-orbital part of the band filled.
- As you move to the right on the periodic table, electrons become more strongly bound (higher ionization energies), and it is no longer energetically favorable to form bands. Instead, single electrons are shared with nearest neighbors in covalent bonds.
- The column in the periodic table with carbon, silicon and germanium would have half-filled bands in metals, which is energetically neutral. They are on the fence between being metals and covalent network solids.

We will next apply this understanding to explain doped semiconductors and their use in computers.

3 Semiconductors

Now that we understand the origin of real bands in actual elements, we can start classifying materals. First some more terminology. If the Fermi level is between two bands we call the band above it the **conduction band** and the band below it the **valence band**. The **bandgap** is the energy difference between the valence and conduction band. In a metal, the Fermi level is within a band, so the valence and conduction bands overlap. In an **insulator** there is a big bandgap. A **semiconductor** is somewhere between an insulator and a conductor: it has a bandgap, typically of order 1 eV.

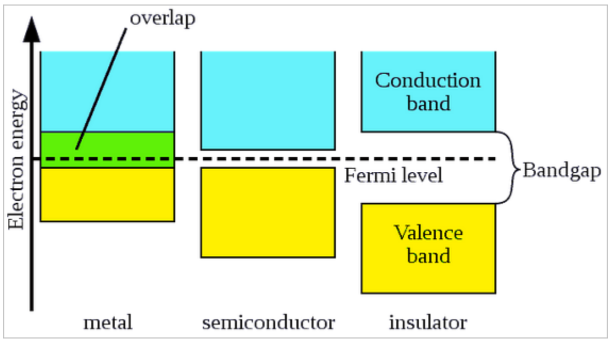


Figure 11. Metals have no bandgap, intrinsic semiconductors a small bandgap and insulators have a large bandgap.

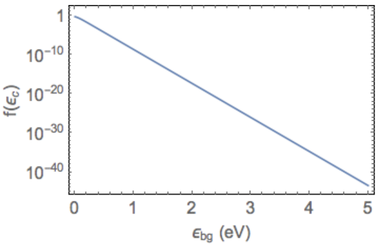
In intrinsic semiconductors, or insulators, the Fermi level is right in the center of the bandgap. This follows from charge conservation and the symmetry of the Fermi distribtion. Charge conservation implies the same total number of electrons and holes (since electron-hole pairs come from real electrons being excited from the valence to conduction band). Recall that the Fermi distribution

$$f(\varepsilon) = \frac{1}{e^{\beta(\varepsilon - \varepsilon_F)} + 1} \tag{1}$$

is symmetric around ε_F (that is, $f(\varepsilon_F + \Delta) = 1 - f(\varepsilon_F - \Delta)$). Also recall that the relevant energies have $\Delta \sim k_B T \ll \varepsilon_F$ so that $g(\varepsilon) \approx g(\varepsilon_F) \times \Theta(\varepsilon)$ where $\Theta(\varepsilon) = 0$ if $\varepsilon_C < \varepsilon < \varepsilon_V$, with ε_C the energy at the bottom of the conduction band and ε_V the energy at the top of the valence band, and $\Theta = 1$ outside of this region. Since the Fermi distribution is symmetric, it is only possible for $N_e = \int_{\varepsilon_F}^{\infty} d\varepsilon f(\varepsilon) g(\varepsilon)$ and $N_h = \int_0^{\varepsilon_F} d\varepsilon f(\varepsilon) g(\varepsilon)$ to be equal if the same number of electron and hole states are removed from the density of states by Θ, i.e. if $\varepsilon_C - \varepsilon_F = \varepsilon_F - \varepsilon_V$. Since $\varepsilon_C - \varepsilon_V = \varepsilon_{\text{bg}}$, this implies that $\varepsilon_F = \varepsilon_V + \frac{\varepsilon_{\text{bg}}}{2}$, which is right in the middle of the band. So the Fermi level is is the center of the bandgap.[2] If the system is modified so that the distribution of available levels is no longer symemtric between electrons and holds, the Fermi level can move. This happens in doped semiconductors (see Section 3.1).

Now, where does the bandgap size ∼1 eV typical for a semiconduction come from? At room temperature $k_B T = 0.025\,\text{eV}$. We can then compute the probability of finding an electron at the base of the conduction band, at energy $\varepsilon_C = \varepsilon_F + \frac{1}{2}\varepsilon_{\text{bg}}$ as

$$f\left(\varepsilon_F + \frac{1}{2}\varepsilon_{\text{bg}}\right) = \frac{1}{\exp\left(\frac{\varepsilon_{\text{bg}}}{2k_B T}\right) + 1} \approx \exp\left(-\frac{\varepsilon_{\text{bg}}}{0.05\text{eV}}\right) = \tag{2}$$



[2] Is is not at the *exact* center of the bandgap, because the density of states is not *exactly* symmetric around ε_F. However, corrections are of order $\left(\frac{\varepsilon_{\text{bg}}}{\varepsilon_F}\right)^2 \approx 0.01$ which is small enough to neglect.

We see that this function is exponentially falling. For $\varepsilon_{\text{bg}} = 1\text{eV}$, the base of the conduction band only has a 10^{-9} chance of being occupied. At $\varepsilon_{\text{bg}} = 2$ eV the probability is already down to 10^{-18}, and at $\varepsilon_{\text{bg}} = 3\text{eV}$ it's down to 10^{-27}. Considering there are of order $N_A \sim 10^{24}$ electrons around, we conclude that a bandgap of order $\varepsilon_{\text{bg}} \gtrsim 2.5$ eV a material will no longer conduct and become an insulator.

The bandgap in NaCl, which forms an ionic solid, is $\varepsilon_{\text{bg}} = 8.9\text{eV}$. NaCl is an insulator. Diamond (a solid form of carbon) at room temperature has $\varepsilon_{\text{bg}} = 5.4\text{eV}$. Diamond is also an insulator. Silicon, below diamond, has $\varepsilon_{\text{bg}} = 1.08\,\text{eV}$ making it on the boundary between conductor and insulator: silicon is a semiconductor. An important characteristic of a semiconductor is number density of conduction electrons (in the conduction band). To compute this, we can treat the electrons in the conduction band as having the usual non-relativistic dispersion relation: $\varepsilon = \varepsilon_F + \frac{1}{2}\varepsilon_{\text{bg}} + \frac{\hbar^2 \vec{k}^2}{2m_e}$ where $\varepsilon_C = \varepsilon_F + \frac{1}{2}\varepsilon_{\text{bg}}$ is the energy at the bottom of the conduction band. Then the density of states for conduction electrons is as in the free-electron gas, $g(\varepsilon) = \frac{V}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}\sqrt{\varepsilon - \varepsilon_F - \frac{1}{2}\varepsilon_{\text{bg}}}$.[3] Approximating $f(\varepsilon) \approx \exp\left(-\frac{\varepsilon - \varepsilon_F}{k_B T}\right)$ as in Eq. (2) we then get

$$n_e = \frac{1}{V}\int_{\varepsilon_F + \frac{1}{2}\varepsilon_{\text{bg}}}^{\infty} g(\varepsilon) f(\varepsilon) d\varepsilon = 2\left(\frac{m k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{\varepsilon_{\text{bg}}}{2k_B T}} \tag{3}$$

Similarly the density of holes is

$$n_h = \left(\frac{2m_e}{\hbar^2}\right)^{3/2} \int_{-\infty}^{\varepsilon_F - \frac{1}{2}\varepsilon_{\text{bg}}} \sqrt{\varepsilon_F - \frac{1}{2}\varepsilon_{\text{bg}} - \varepsilon}\, e^{-\frac{\varepsilon_F - \varepsilon}{k_B T}} d\varepsilon = 2\left(\frac{m k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{\varepsilon_{\text{bg}}}{2k_B T}} \tag{4}$$

So, in a pure intrinsic semiconductor the densities of electrons and holes are the same: $n_e = n_h = n_I$. n_I is called the **intrinsic carrier concentration**. Plugging in the numbers for silicon we get

$$n_I = 2\left(\frac{m k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\frac{\varepsilon_{\text{bg}}}{2k_B T}} = \left(2.4 \times 10^{19} \frac{1}{\text{cm}^3}\right) e^{-\frac{1.08\text{eV}}{0.05\text{eV}}} = 10^{10} \frac{1}{\text{cm}^3} \tag{5}$$

If the Fermi level were not in the middle of the band gap (as it won't be in doped semiconductors), the expressions $\varepsilon_F \pm \frac{1}{2}\varepsilon_{\text{bg}}$ in Eqs. (3) and (4) would change to $\varepsilon_F + a\varepsilon_{\text{bg}}$ and $\varepsilon_F - (1-a)\varepsilon_{\text{bg}}$, but we would still find

$$n_e n_h = n_I^2 = 4\left(\frac{m k_B T}{2\pi\hbar^2}\right)^3 e^{-\frac{\varepsilon_{\text{bg}}}{k_B T}} \tag{6}$$

This equation is sometimes called the law-of-mass-action for semiconductors, as it relates the equilibrium electron and hole concentrations much like the law of mass action does in chemistry.

Let's try to understand a little better why carbon is an insulator and silicon is a semiconductor. Carbon is $C = [\text{He}]2s^2 2p^2$. Its 4 valence electrons allow it to form 4 covalent bonds in a covalent network solid filling up it shells. This is more stable than forming metallic bonds. The atomic spacing in diamond is $a_C = 0.154\,\text{nm}$. Its electrical resistivity is $r_C = 10^{14}\Omega m$. Silicon is $\text{Si} = [\text{Ne}]3s^2 3p^2$. Silicon, like carbon, has a valence of 4, and forms covalent bonds, but the valence electrons of silicon are farther out, so the bonds are weaker. These weaker bonds are why the atomic spacing in silicon is about twice as large as diamond, $a_{\text{Si}} = 0.235\,\text{nm}$. Thus while silicon is sort of covalently bonded, it could equally well be thought of as having metallic bonds. The tight-binding model lets us interpolate between covalent and metallic. The resistivity of silicon is $r_{\text{Si}} = 0.001\,\Omega m$, many orders of magnitude smaller than diamond (but many orders of magnitude larger than metals, such as copper with a resistivity of $r_{\text{Cu}} = 1.7 \times 10^{-8}\Omega m$).

Below silicon on the periodic table is germanium: $\text{Ge} = [\text{Ar}]3d^{10}4s^2 4p^2$. Its electrons are even farther out, lowering its bandgap to $\varepsilon_{\text{bg}} = 0.67\,\text{eV}$, the lattice spacing is $a_{\text{Ge}} = 0.243\,\text{nm}$ and the resistivity $r_{\text{Ge}} = 0.0005\,\Omega m$, about half that of silicon. So silicon and germanium are semiconductors. Another important semiconductor is gallium arsenide GaAs, with $\varepsilon_{\text{bg}} = 1.43\text{eV}$.

[3] Technically, the mass here is an "effective mass" determined by the curvature of the dispersion relation at the base of the conduction band. For simplicity, we'll just take the effective mass to be the electron mass.

3.1 Doping

What makes semiconductors very important is that the valence and conduction bands and the bandgap are relativity easy to manipulate by adding impurities. This is called **doping**.

The properties of doped semiconductors are best understood in the language of electrons and holes. Recall from the last lecture that electrons and holes helped us understand the heat capacity of metals in the free electron model. The power of the picture comes from the symmetry of the Fermi function $f(\varepsilon)=\frac{1}{e^{\beta(\varepsilon-\varepsilon_F)}+1}$, namely that $f(\varepsilon_F+\Delta)=1-f(\varepsilon_F-\Delta)$. A gas of electrons has as many excited states above ε_F as there are holes below ε_F. Not only is the number of states the same, but the probability of finding an electron state at $\varepsilon_F+\Delta$ is the same as the probability of finding a hole at $\varepsilon_F-\Delta$. A hole contributes a positive amount to the energy of the electron gas, since the electron that should have been in the hole is missing. If a electron in the valence band at energy ε is excited into the conduction band at energy ε', of the $\varepsilon'-\varepsilon$ total energy, the part $\varepsilon_F-\varepsilon$ of it is attributed to the hole and the rest $\varepsilon'-\varepsilon_F$ is attributed to the electron.

Now we introduce another useful property of holes: they have positive charge. When an electron is excited out of an atom, it leaves a positive ion in its place. An analogy is a line of parked cars with a spot at the front. As a car moves out of its spot into the spot in front of it, it leaves a hole. Then another car behind it can fill the hole, leaving a different hole. In this way the hole moves backwards in the line, although it is really the cars that are moving. In the electron gas picture, the electrons are not associated with individual atoms, so the holes should not be associated with individual atoms either. Instead, we should think of the states near ε_F in a semiconductor as gas of negatively-charged electrons and positively-charged holes each of which has a minimum energy of excitation $\varepsilon\gtrsim\frac{\varepsilon_{\text{bg}}}{2}$.

Ok, now to doping. Let's take silicon as the semiconductor and consider adding a small amount of phosphorous. Phosphorous is the element to the right of silicon and therefore has one more valence electron (5 instead of 4): $P=[\text{Ne}]3s^2 3p^3$. The extra electron is weakly bound and, since this new electron is the 5th of the possible 8 that the 3s/3p orbital could hold, it would have to contribute to an antibonding orbital if P formed a metal. Therefore, its energy is close to the conduction band of Si, but slightly lower since it is energetically favorable *not* to fill antibonding orbitals. More quantitatively, the new levels in P-doped Si are centered at 45 meV below the conduction band. When we dope Si with P, the extra valence electrons from the phosphorous fill the new levels and are easily excited into the conduction band. Thus adding phosphorous increases the carrier concentration and makes silicon more conductive. P-doped Si is an example of an **n-type semiconductor** since the new particles added are **n**egatively charged, i.e. electrons.

We can also dope silicon by adding some of the element on its left, aluminum: $\text{Al}=[\text{Ne}]3s^2 3p^1$. Aluminum has one fewer electron than silicon. Thus it provides new places for the electrons in Si to go. The energies of these new states are slightly above the valence band, so we call these states "acceptor" sites. For aluminum, the new acceptor sites are centered 57 meV above the valence band. Thus, Al-doped Si is also more conductive than pure silicon because electrons can move from the valence band into these new acceptor sites. Note that Al does not contribute new electrons, so it is the valence electrons from Si that move into the new sites. Alternatively, we can say that Al contributes new holes. These new holes can move down into the valence band. The hole picture is nice because of the symmetry where we swap n-type ↔ p-type semiconductors and electron ↔ hole. Al-doped-Si is called a **p-type semiconductor** because holes are **p**ositively charged.

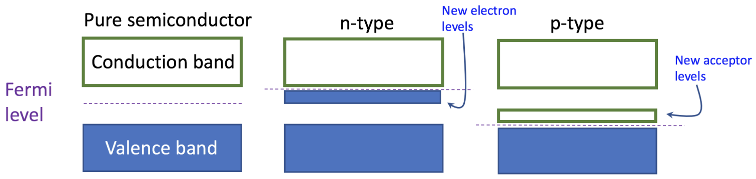


Figure 12. Doping of semiconductors changes the band structure.

FYI, typical doping fractions range from 1 part per thousand to 1 part per billion.

4 Diodes

Now it gets interesting. What happens if we put a p-type semiconductor next to an n-type one? The extra electrons from the P atoms will diffuse across the junction to find the Al atoms. Eventually, enough charge will be transferred that a substantial charge difference will build up, inhibiting further charge transfer. We call this the accumulated voltage the **junction potential** and denote it by V_{junc} The result looks like this:

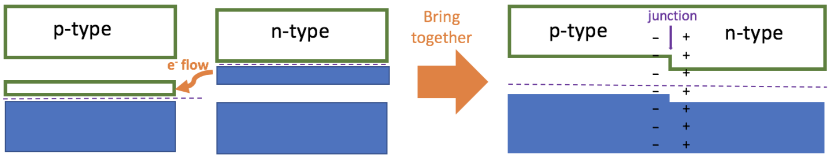


Figure 13. When p-type and n-type semiconductors are brought together, electrons move from the n side to the p side. Charge builds up in the interface forming a p-n junction.

To be more quantitative, consider first the pure n-type or pure p-type semiconductor. In these, the extra donor/acceptor sites significantly increases the carrier concentrations. On the n-type side, the donor (electron) concentrations are typically $n_e^{n\text{-type}}=1.0\times10^{16}\frac{1}{\text{cm}^3}$. Because of the law of mass action, Eq. (6), the concentration of holes on the n-type side is then $n_h^{n\text{-type}}=\frac{n_I^2}{n_e^{n\text{-type}}}=2.2\times10^4\frac{1}{\text{cm}^3}$ with $n_I=1.5\times10^{10}\frac{1}{\text{cm}^3}$ the intrinsic carrier concentration (set by the band gap) as in Eq. (5). On the p-type side, the acceptor (hole) concentrations are typically $n_h^{p\text{-type}}=1.0\times10^{15}\frac{1}{\text{cm}^3}$ and so $n_e^{p\text{-type}}=\frac{n_I^2}{n_p^{n\text{-type}}}=2.2\times10^5\frac{1}{\text{cm}^3}$. Now, when we put the two sides together the difference in electron/hole concentrations must be compensated for by the production of a junction potential. So we should have

$$n_h^{n\text{-type}}=n_h^{p\text{-type}}e^{-\frac{\varepsilon_{\text{junc}}}{k_BT}} \tag{7}$$

This implies

$$\varepsilon_{\text{junc}}=k_BT\ln\frac{n_h^{p\text{-type}}n_e^{n\text{-type}}}{n_I^2}=25\,\text{meV}\times\ln\frac{10^{15}10^{16}}{10^{20}}=0.61\,\text{eV} \tag{8}$$

The voltage is then $V_{\text{junc}}=\frac{\varepsilon_{\text{junc}}}{e}=0.61V$. Commercial silicon p-n junctions are typically engineered by varying the doping levels to have junction potentials of this size, around 0.6 V.

Now what happens when we apply an external voltage to the p-n junction? If the negative terminal is connected to the n side (**forward bias**), it supplies electrons that diffuse towards the junction. At the junction they help neutralize the charge barrier, lowering the junction potential. Alternatively, we can think of holes diffusing out of the positive terminal into the p side, to neutralize the junction potential. Once the voltage overcomes the junction potential ($\gtrsim 0.6V$), the natural diffusion of electrons from the n side to the p side (or holes from p to n) can resume. The electrons then annihilate the holes contributed from the battery and a steady-state current flows through the system.

If the applied voltage is in the other direction, however, so that the negative terminal connects to the p side, it will draw holes from the n side and send electrons into the p side. The electrons on the p side make the junction potential larger and no current passes through. The junction voltage simply increases. This is called **reverse bias**. If enough reverse-bias voltage is applied, above the **breakdown voltage** (typically $30V$-50V), then the electrons are forced across the junction and a current flows (in the opposite direction to the forward bias case).

Thus this p-n junction acts a **rectifier** or **diode**: it allows current to flow in only one direction: $V \gtrsim 0.6V$ is required in forward bias to conduct but $V \gtrsim 50V$ is required for reverse bias to conduct:

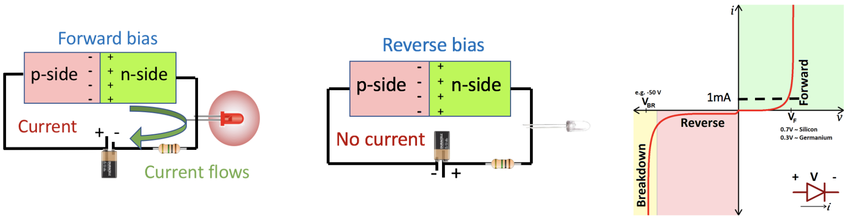
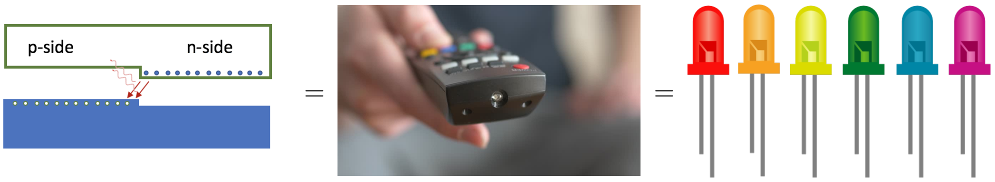


Figure 14. Current only flows through a p-n junction when a negative voltage is applied to the n side (forward bias). Flipping the battery (reverse bias) dues not induce current. Right shows the current induced for an applied voltage for a p-n junction.

4.1 Light-emitting diodes

In forward bias mode, with an applied voltage larger than the junction potential ($V_{\text{applied}} \gtrsim 0.6V$), the voltage continuously pulls holes out of the p-side and supplies electrons to the n side so that the current continues to flow. The flowing electrons enter on the n side into its conduction band. Similarly, the flowing holes enter the valence band on the p side. As the conduction electrons cross the junction from the n side to the p side they would need to pick up energy to stay in the conduction band (since the p-side conduction band is higher than the n-side one). As there is nowhere to get this energy from, the electrons instead fall down into the valence band on the p-side, annihilating holes. A falling electron loses roughly $\Delta\varepsilon \sim \varepsilon_{\text{bg}}$ of energy from this transition. Like any electronic transition, this energy leaves the system through a photon of wavelength $\lambda = \frac{hc}{\Delta\varepsilon}$. For silicon, the bandgap is around $\varepsilon_{\text{bg}} = 1.1$ eV which corresponds to a wavelength of $\lambda = 1130\,\text{nm}$, in the infrared. In summary, at voltages above around $0.6V$ of the junction potential, a p-n junction emits monochromatic infrared light. Such a device is called a **light-emitting diode** or **LED**.



(9)

The earliest LEDs were indeed in the infrared and provided the technology for the first remote controls – that little red dot on your old TV remote is an LED.

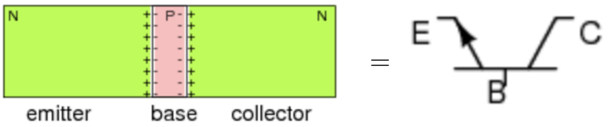
Since a 1.1 eV bandgap is in the infrared, in order to get to higher frequencies (visible)some fancy material-science engineering is required. As we noted before, you can't have a semiconductor with too large of a bandgap or it will just be an insulator. Moreover, the bigger the bandgap, the more unstable the diode becomes. An effective method for growing blue LEDs was finally found by Akasaki, Amano and Nakamura in the early 1990s. They received the Nobel Prize for their work in 2014. The reason that blue was so important is that blue is that there are many chemicals called phosphors that can be used to convert blue light to other colors. So the blue LED technology led to the ubiquitous extremely energy efficient LED lights found today. An LED lightbulb with the same luminescence as a 100 W incandescent lightbulb can use as little as 5 W: a 95% efficiency gain.

LEDs are extremely efficient since essentially all of the energy goes into light. This is in contrast to incandescent lights which use blackbody radiation: only a small fraction of the energy of an incandescent bulb is in the visible spectrum (you know how to calculate this!). Much of the energy of an incandescent bulb is in the infrared and dissipated as heat. That's why incandescent bulbs get very hot, but LEDs do not.

5 Transistors

It is hard to overestimate the importance of transistors in nearly every aspect of technology. Early computation used vacuum tubes or even mechanical relays as transistors. These worked, but were slow and clunky. It wasn't until Shockley, Bardeen and Brattain showed that transistors could be made out of semiconductors in 1947 (Nobel prize 1956) that modern computing took off. These solid-state transistors allowed for the miniaturization of computers and the efficiency improvements that we have seen over the last 70 years.

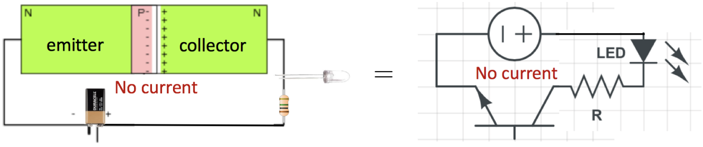
A **bipolar junction npn transistor** is made by combining two n-type semiconductors on either side of a very thin p-type semiconductor. We call the middle part the base, and the two sides the emitter and collector. The emitter is generally much more heavily doped than the collector. The transistor looks like this



(10)

The symbol on the right is how we represent an npn transistor in a circuit diagram.

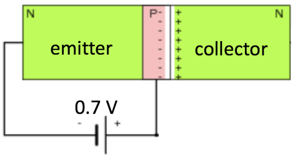
To a first approximation, this thing is just two diodes sandwiched together. So if we try to connect the emitter to the collector, with say a battery and an LED, nothing happens:



(11)

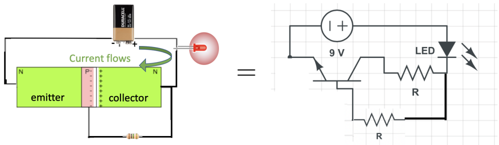
Turning the battery around doesn't help either: no current flows with either positive or negative voltage.

Current would flow if we connect the base to the emitter or collector. If we apply a voltage forward bias between the emitter and base we can eliminate their junction potential:



(12)

The forward bias on the emitter NP junction allows the emitter to conduct freely, in either direction. Thus, now if we run current between the emitter and collector, with say an LED on one side, we would find we find a light:



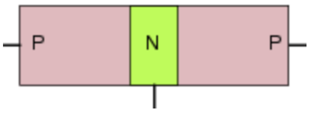
(13)

The point is that adding applying a 0.7 V voltage between the base and the emitter controls whether current can flow or not into the circuit.[4]

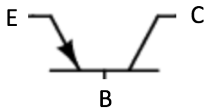
[4] Real transistors are slightly asymmetric between collector and emitter. A typical voltage drop between base and emitter is $V_{be} \approx 0.6 - 0.7V$, and between base and collector is $V_{bc} \approx 0.5 - 0.6V$, so that the collector is at slightly lower potential than the emitter ($V_{ce} \approx 0.1V$).

Note that for this to work the base has to be thin. The thinner it is, the more likely electrons are to be swept across the base and deposited into the collector rather than exiting out the bottom into the forward-biasing circuit. Typically more than 99% of the current flows into the collector.

The other kind of transistor is a pnp transistor which has two p-type semiconductors sandwiching an n type semiconductor in the middle. These are drawn as



=



(14)

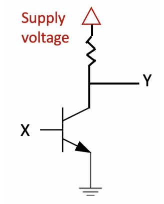
The arrow on the emitter shows which way current can flow when the transistor is turned on. Remember the convention in electronics is that current arrows leave the positive terminal and flow into the negative terminal. So in npn transistors, current flows from collector to emitter when a positive voltage is applied to the base; in pnp transistor, current flows from emitter to collector when the base is connected to negative voltage (ground).

5.1 From transistors to logic gates

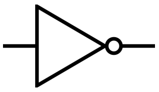
We have shown that a transistor is a kind of gate. If you put a positive voltage to the base of a npn transistor, it opens the gate, allowing current to flow from collector to emitter. If you put a negative voltage to the base of a pnp transistor, it opens the gate, allowing current to flow from emitter to collector. What can we do with these things? The next step to getting from transistors to computers is building the basic set of boolean logic gates. These are things like NOT, AND, NAND, OR, NOR or XOR.

Let's start with NOT. A NOT gate, also called an **inverter**, takes 0 and turns it into 1 and takes 1 and turns it into 0. It can be simply built from a npn transistor by connecting X to the base and Y to the collector. We also hook a supply voltage up to the collector and the emitter to the ground. So it looks like this

NOT =



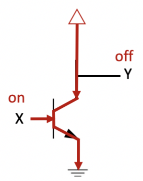
=



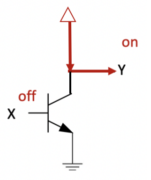
(15)

So if X is 0 (off) then the current will not be able to pass through the transistor and is diverted to Y. Thus $X = 0$ implies $Y = 1$. If X is 1 (on) then the gate is open and the current can pass right through the transistor to ground, giving Y nothing. Thus $X = 1$ implies $Y = 0$. This is a not gate:

$1 \to 0$:



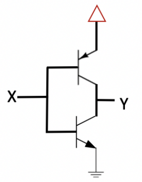
$0 \to 1$:



(16)

By the way, this gate also shows how transistors act like amplifiers: a small voltage (0.7V) applied at X can be turned into an arbitrarily large voltage at Y (that of the supply voltage). The efficiency of amplification led to transistor radios in the 1950s: a weak radio signal comes in that we can idealize as a binary pattern 010010101 at X. Adding two NOT gates in series with a large supply voltage produces the same pattern but amplified at Y. Previous radios used vacuum tubes for amplification and were heavy, delicate pieces of furniture. The method of amplification using solid state transistors allowed radios to be small and portable, ushering in the consumer electronics revolution.

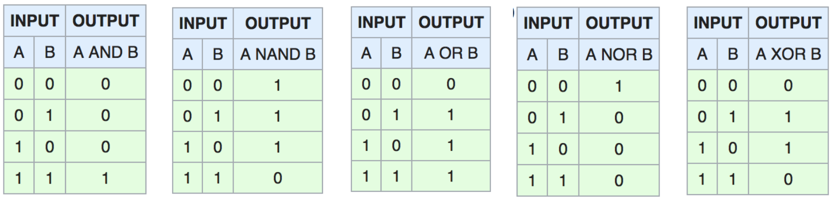
This style of logic gate construction is called NMOS (n-type metal-oxide semiconductor). It uses only npn transistors.[5] An alternative circuit design with identical function are the CMOS (complimentary metal-oxide semiconductor) circuits that combine npn and pnp transistors. For example, a CMOS inverter gate looks like



(17)

Note that the CMOS does not need a resistor to prevent shorting: in the off position, no current is drawn. Thus CMOS has very little static power consumption which is one of its main advantages. CMOS circuits replaced NMOS as the standard sometime in the 1990s. The integrated circuits in your computer and your phone undoubtedly use CMOS.[6]

The other standard logic gates are all 2 → 1 gates, so they take two inputs and give one out. AND for example, gives $A \wedge B$ meaning that if A and B are both 1, then it gives 1 otherwise it gives 0. The definitions of these various operations are



INPUT		OUTPUT
A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

AND

INPUT		OUTPUT
A	B	A NAND B
0	0	1
0	1	1
1	0	1
1	1	0

NAND

INPUT		OUTPUT
A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

OR

INPUT		OUTPUT
A	B	A NOR B
0	0	1
0	1	0
1	0	0
1	1	0

NOR

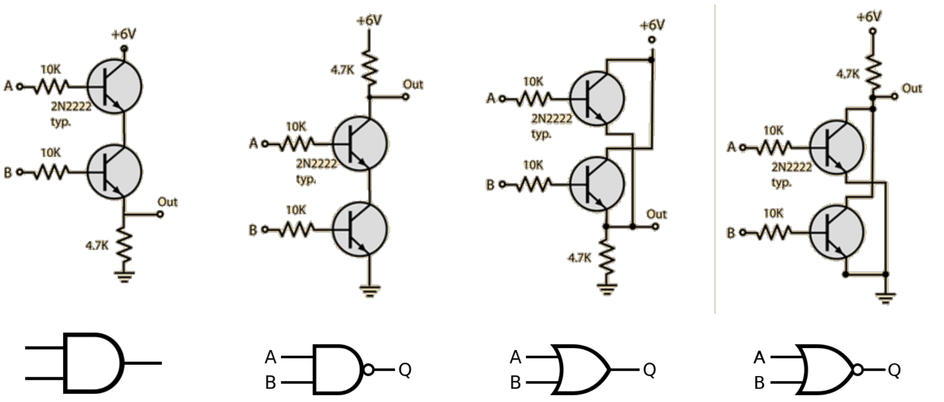
INPUT		OUTPUT
A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

XOR

(18)

Example NMOS circuit diagrams to build some of these with transistors are:

AND NAND OR NOR



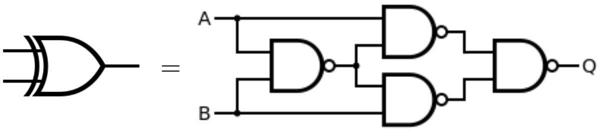
(19)

Let's look at the NAND gate. The inputs A and B both have to be on for current to flow to ground. If either is off, then the current flows to "Out". This is consistent with the NAND logic table. The AND gate is the same, but the output is after the second emitter, so both have to be on for current to flow out. For OR, if the based of either transistor gets current, then the current will pass through to "Out", consistent with OR logic.

[5] npn transistors are faster than pnp transistors, since electrons have larger mobility than holes.

[6] Another technical point is that CMOS circuits generally use field-effect transistors (FETs) rather than the bipolar junction transistors (BJTs) that we have been discussing. FETs, like BJTs, are solid-state devices made form p-type and n-type semiconductors. Their design allows them to control the current flow through the transistor using voltage across the collector-emitter channel, so that, in contrast to a BJT, no current flows through the base.

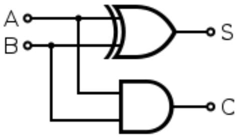
The XOR gate is an “exclusive OR”, meaning it gives 1 if *and only if* 1 of the inputs is 1 and other other is zero. We can construct an XOR gate out of NAND gates:



(20)

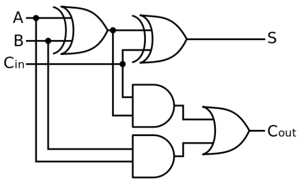
Can you figure out why this combination of NAND gates acts as an XOR gate?

Note from the logic table in Eq. (18) for XOR is the same as for the addition of 1 bit binary numbers. Although the single XOR gate only adds two 1 bit numbers modulo 2, which is maybe not so interesting, XOR gates can be combined to make more complicated adders. To add more than 2 bits, we need to carry over the second binary digit. We do this with a carry block in what’s called a **half adder**



(21)

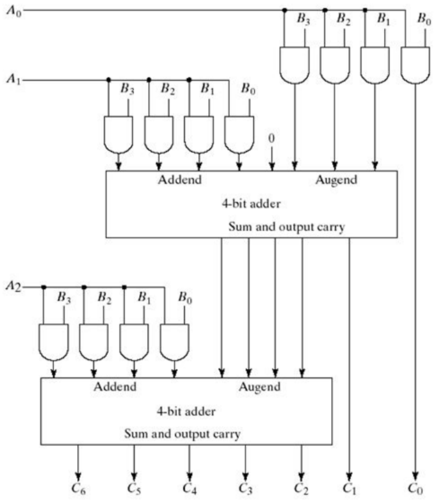
In this diagram, A and B are added with an XOR into S (for sum). The S bit is 0 if $A = B = 0$, but also if $A = B = 1$. If $A = B = 1$ the answer should be 2, and we need a bit in the second “digit”. The “carry” bit C carries this information: the AND sets bit C only if $A = B = 1$. Now we just hook this up with another bit to add 3 bits in a **full adder**:



(22)

Can you figure out why this adds the 3 bits? You can check your undestanding at http://www.falstad.com/circuit/e-fulladd.html.

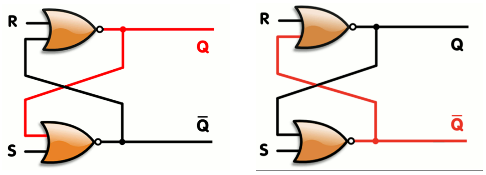
From here, you can add more and more bits together by sewing in successive adders. You can make more complicated integrated circuits to multiply numbers and do other operations with them. For example, here’s a circuit that multiplies 3 bits $A = (A_2, A_1.A_0)$ by 4 bits $B = (B_0, B_1, B_2, B_3)$ to give 7 bits in C



(23)

So to multiply 7 × 13 you would set $A = 111$ and $B = 1101$. The output should be the 7 bits in $C = 1011011$.

There's one final element needed to make a computer: memory. There are many ways to make memory, but the basic form is a **latch** or **flip-flop**. For example, we can make such a thing using cross-coupled NOR gates:



(24)

This object called a or **SR latch** (for Set-Reset) has two stable states, shown in red. Either $Q=1$ and $\bar{Q}=0$ (left) or $Q=0$ and $\bar{Q}=1$ (right). To change from the left state to the right, we can put a current into the R input. Since R feeds into a NOR gate, it only gives current out if both inputs are 0 so when we change R, the current stops. Once the current stops, it no longer flows into the bottom NOR gate. Since $S=0$ this then makes the current go out of that gate and link up with the top NOR. At this point the current can be removed from R and the circuit stays in the right state. To go from right to left, we can pulse S. In this way, we can store a bit, and toggle it back and forth through the Set and Reset switches: Set makes it 1 (left), Reset makes it 0 (right).

Now that we have memory and can compute things, it's just a matter of putting them together in ever more complex combinations. Note that NAND has 2 transistors, XOR has 8, AND and OR have 3, so the full adder has 25 transistors. So a 4×4 bit adder should have roughly 400 transistors.

Adding 4 bit numbers doesn't seem like much, but you can actually reuse circuit elements fairly easily, storing the output then feeding it back in. In early computers, the transistors were not the solid-state devices we discussed here but rather vacuum tubes. The Mark 1 computer that you've walked by a thousand times in the Science Center but never looked at has around 1000 mechanical relay transistors. It operated on 24 bit numbers and took about $0.3s$ to add two such numbers. An iPhone X has 4 billion transistors and can add two 24 bit numbers in around $\frac{1}{\text{GHz}}=10^{-9}s$.

6 Summary

The first half of this lecture introduces molecular orbital theory and explained how to think about the periodic table from the point of view of the types of solids formed. A summary of the main results of the first have was given in Section 2.6. There's a lot of physical chemistry in this first half. So if you never plan on taking a physical chemistry or inorganic chemistry class, you are strongly encouraged to study this first half in depth.

The second half of the lecture focused on semiconductors and why they are so important for technology. Semiconductors are like insulators, in that their Fermi level is in a band gap, but the band gap is small, typically of order 1 eV, so it's pretty easy for electrons in the band below the Fermi level (the valence band) to get excited into the band above the Fermi level (the conduction band). Indeed, at room temperature 10^{-9} of the electrons are excited. This is in constrast to an insulator like diamond where 10^{-45} of the electrons are excited. By applying external voltage was can easily manipulate the electrical properties of semiconductors.

The key use of semiconductors is in making diodes, which are materials in which current can only flow in one direction. Diodes are made by doping semiconductors by adding atoms with extra valence electrons, $(n-\text{type})$ or fewer valence electrons $(p-\text{type})$. Light-emitting-diodes make very energy efficient lightbulbs.

Diodes can be sandwiched together in the npn or pnp order to form transistors. Their important property is that the conduct only when a voltage is applied to the base (middle part). So that they acts as if statements: *if* a voltage is applied, let current flow. Combining transistors in various ways one can make logic gates, and then computers. A cell phone typically has billions of transistors.