Semiconductors

Jiinyuan Wu

September 12, 2023

In the next several sections, we discuss the following topics:

- The behavior of the semiconductor when doping is low/relatively high/so high that an "impurity band" forms.
- The role of temperature.
- Junctions.

1 Two-band model

At $T = 0$ there is strictly no such thing as a semiconductor: there are just band metals or band insulators. When $T > 0$, however, the following two mechanisms provide a band insulator with non-zero carrier concentration:

- The first is that electrons jump from the valence band to the conduction band because of thermal fluctuation, or more specifically, because of Fermi-Dirac distribution.
- The second is dopoing adds energy levels near the highest point of the valence band (E_v) or the lowest point of the conduction band (E_c), and again, thermally excited electrons will jump to them. From another point of view we may say doped atoms eat electrons or give electrons, so some electrons are missing if we only look at Bloch states in the spectrum.

In both mechanisms, we have electrons in the conduction band and holes in the valence band, which are **carriers** of electric current.

In the discussion below, we assume that the band gap is *much larger* than $k_\mathrm{B}T$, and this means when $E \geq E_\mathrm{c}$,

$$n_\mathrm{electron} = \mathrm{e}^{-(E-\mu)/k_\mathrm{B}T}, \tag{1}$$

and when $E \leq E_\mathrm{v}$

$$n_\mathrm{hole} = \mathrm{e}^{-(\mu-E)/k_\mathrm{B}T}. \tag{2}$$

The conditions are usually satisfied in most semiconductors; in more accurate numerical predictions, of course, we don't really need the approximations. In semiconductor physics we usually use the following abbreviations:

- p (positive), v (valence), h (hole) mean holes in the valence band, and
- n (negative), c (conduction), e (electron) mean electrons in the conduction band.

Thus the density of electrons and holes are given by

$$n = \int_{E_\mathrm{c}}^{\infty} D_\mathrm{c}(E)\,\mathrm{d}E\,\mathrm{e}^{-(E-\mu)/k_\mathrm{B}T}, \tag{3}$$

and

$$p = \int_{-\infty}^{E_\mathrm{v}} D_\mathrm{v}(E)\,\mathrm{d}E\,\mathrm{e}^{-(\mu-E)/k_\mathrm{B}T}. \tag{4}$$

Consider the simplest case, where we have a band gap (possibly indirect)

$$E_\mathrm{g} = E_\mathrm{c} - E_\mathrm{v} \tag{5}$$

between the highest valence band the the lowest conduction band. The $\Delta E \gg k_\mathrm{B}T$ condition indicates that only the top of the highest valence band and the bottom of the lowest conduction band matters, and therefore the hyperbolic approximation works, and the conduction band is

completely characterized by its effective mass and the momentum with lowest energy, and so is the case for the valence band. We have

$$D(E) = \frac{1}{2\pi^2}\left(\frac{2m^*}{\hbar^2}\right)\sqrt{E - E_{\max/\min}} \tag{6}$$

for hyperbolic bands, and therefore we have (note that in the valence band, the effective mass of the electron is negative, so the effective mass of the hole is positive)

$$n = 2\left(\frac{m_{\mathrm{e}}^* k_{\mathrm{B}} T}{2\pi\hbar^2}\right)^{3/2} \mathrm{e}^{-(E_{\mathrm{c}}-\mu)/k_{\mathrm{B}}T}, \tag{7}$$

$$p = 2\left(\frac{m_{\mathrm{h}}^* k_{\mathrm{B}} T}{2\pi\hbar^2}\right)^{3/2} \mathrm{e}^{-(\mu-E_{\mathrm{v}})/k_{\mathrm{B}}T}. \tag{8}$$

Note that if there is doping, electrons donated or attracted by the doped atoms are not included in the above equations. Thus, we find

$$n \cdot p = 4\left(\frac{k_{\mathrm{B}} T}{2\pi\hbar^2}\right)^3 (m_{\mathrm{e}}^* m_{\mathrm{h}}^*)^{3/2} \mathrm{e}^{-E_{\mathrm{g}}/k_{\mathrm{B}}T}. \tag{9}$$

Note that this result doesn't contain μ; specifically, its derivation involves no information about doping: whatever the doping is, the product of the number of holes in the valence band and the number of electrons in the conduction band is always decided by the shape of the two bands, the band gap, and the temperature.

2 The intrinsic semiconductor limit

In the **intrinsic semiconductor limit**, there is no doping, and only the first mechanism – thermal excitation in the conduction band and the valence band – works. This limit is useful when T is very high so the effect of doping can be ignored. So in the opposite, *when T is low enough, even materials that are very clean can't be described well by the intrinsic semiconductor limit*, because in this case, if the material has any semiconductor property, then it has to come from doping.

2.1 Isotropic two-band model

Since there is no doping, we have $n = p$ because of charge neutrality, and we have $n = \sqrt{np}$, and thus from (9), we find

$$\mu = E_{\mathrm{v}} + \frac{1}{2}E_{\mathrm{g}} + \frac{3}{4}k_{\mathrm{B}}T\ln\left(\frac{m_{\mathrm{h}}^*}{m_{\mathrm{e}}^*}\right). \tag{10}$$

Here we can see that when $T = 0$, actually $\mu \neq E_{\mathrm{F}} = E_{\mathrm{v}}$, but this doesn't matter: when T is zero and we are working with an insulator, putting μ anywhere between E_{c} and E_{v} is acceptable.

Here Drude model works, because the density of electrons and holes is small (TODO: why no strong correlation, like Wigner crystal?), and therefore the average distance between electrons is much larger than the thermal de Broglie wave length, and the quantum nature of electrons isn't apparent. (TODO: this is what Sohrab said; but Drude model still works for conductivity even in the quantum case so I don't know what he meant here) So we have

$$\sigma_{\mathrm{e}} = n\mu_{\mathrm{e}}, \quad \sigma_{\mathrm{H}} = p\mu_{\mathrm{H}}, \tag{11}$$

and the mobilities are

$$\mu_{\mathrm{e}} = -\frac{\tau e}{m_{\mathrm{e}}^*}, \quad \mu_{\mathrm{h}} = \frac{\tau e}{m_{\mathrm{h}}^*}. \tag{12}$$

Usually the mobility of holes isn't as good as electrons because in general $m_{\mathrm{h}}^* > m_{\mathrm{e}}^*$, and for large effective mass we expect low mobility. This can be found by looking at (12), and (12) comes from the semiclassical EOM of electrons and holes, so we can also explain the fact from a quantum perspective. Suppose a hole is stuck in the lattice (it may slightly distort the lattice, so what is stuck is actually a hole and some phonons, or a polaron). It may still tunnel away because x is localized and p is uncertain, which may give it a kinetic energy large enough to escape, but if m_{h}^* is large enough, the kinetic energy fluctuation can be ignored, so the hole is safely localized.

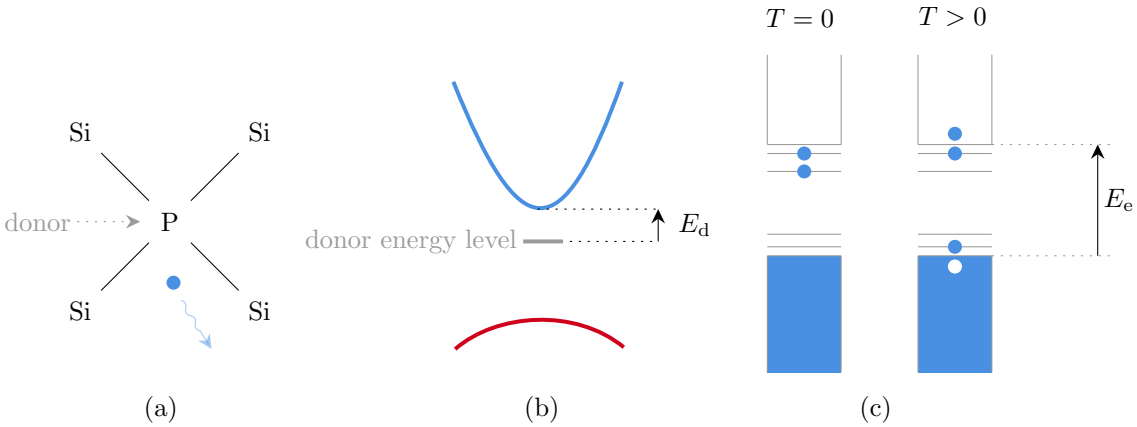

Figure 1: Doping: an additional electron is introduced into the material after an electron donor is doped (a), which is bound around the donor at the ground state and therefore a donor energy level is formed (b); the electron may be thermally excited into the conduction band (c).

2.2 The anisotropic two-band model

When the anisotropy of the lattice is strong, we have

$$n = 2\left(\frac{k_\mathrm{B}T}{2\pi\hbar^2}\right)^{3/2}(\tilde{m}_\mathrm{e}^*\tilde{m}_\mathrm{h}^*)^{3/4}\mathrm{e}^{-E_\mathrm{g}/k_\mathrm{B}T}, \tag{13}$$

where $\tilde{m}$ means $(\det \boldsymbol{M})^{1/3} = \sqrt[3]{m_1^* m_2^* m_3^*}$, with the three m^*'s being the eigenvalues of the mass matrix, which is defined as

$$E = E_0 \pm \frac{1}{2}\boldsymbol{k}\cdot\boldsymbol{M}^{-1}\cdot\boldsymbol{k}, \tag{14}$$

where E_0 is the bottom/top of the band.

2.3 More valleys

It's possible that we have two valleys in the conduction bands or in the valence bands; the latter happens more frequently, for some reasons. In this case a analytic solution is hard to get.

3 Doping

The overall effect of doping is to introduce a difference between the number of electrons/holes in the conduction band/valence band:

$$\Delta n = n - p \neq 0. \tag{15}$$

Suppose n_i and p_i are number of electrons and holes in the intrinsic semiconductor, given by (7) and (10). Recall that (9) is obtained without assuming that the system is intrinsic, and that $n_\mathrm{i}, p_\mathrm{i}$ also satisfy the equation, and we find

$$n\cdot p = n_\mathrm{i}^2 = p_\mathrm{i}^2, \tag{16}$$

which in turns reads

$$n = \frac{\Delta n + \sqrt{4n_\mathrm{i}^2 + \Delta n^2}}{2}, \quad p = \frac{-\Delta n + \sqrt{4n_\mathrm{i}^2 + \Delta n^2}}{2}. \tag{17}$$

The **extrinsic limit** is when $\Delta n \gg n_\mathrm{i}$. Now the problem is how to find Δn (and therefore to find how we can achieve the extrinsic limit). This can be done by calculating the particle number expectation of each impurity, which, in turn, reduces to the problem of calculating the electronic energy levels introduced by the impurities.

3.1 What does doping do?

Let's consider the example in Figure 1, where a P atom is doped into a silicon single crystal. P has one more electron than Si, so if the electron is able to go around in the crystal, it should appear in the conduction band. But of course it's still possible that the electron is bound around the P atom. The Hamiltonian of an electron in the conduction band is therefore roughly

$$H = -\frac{\hbar^2 \boldsymbol{\nabla}^2}{2m_{\mathrm{c}}^*} - \frac{e^2}{\epsilon |\boldsymbol{r}|}, \tag{18}$$

where ϵ comes from the screening effects of other electrons, which may be estimated by, say, GW approximation. The Hamiltonian has a bound state with negative energy as the ground state, and therefore we see a donor energy level formed in the spectrum of the system, illustrated in Figure 1. Similarly, when an electron attractor is introduced in the system, in the electron representation it forms an attractor energy level above the valence band (and therefore in the hole representation, an energy level below the hole band which is the inverse of the valence band). We use E_{d} and E_{a} to refer to the gap between the donor energy level and the bottom of the conduction band and the attractor energy level and the top of the valence band, respectively.

When $T = 0$, electrons stay at donor energy levels, and attractor energy levels are empty, or in other words are filled by holes (Figure 1(c)); when the temperature rises, electrons go to the conduction band and the attractor energy levels, the latter also known as "holes go to the valence band".

3.2 Doped electron/hole concentration

Usually, we assume that there is at most one electron at each impurity energy level, or otherwise strong Hubbard expulsion occurs. The Hubbard U as estimated as

$$U \sim \frac{e^2}{a_0^* \epsilon} \sim 100\,\mathrm{meV} \gg k_{\mathrm{B}} T, \tag{19}$$

where a_0^* is the Bohr radius in which m is replaced by m^*. Thus, we find the double occupation configurations are highly unlikely in the temperature range that most condensed matter experiments are carried out. The partition function of a donor energy level is therefore

$$Z_{\mathrm{d}} \approx 1 + 2\mathrm{e}^{-(E_{\mathrm{d}} - \mu)/k_{\mathrm{B}} T}, \tag{20}$$

and the electron occupation expectation at the donor energy level is

$$\langle n_{\mathrm{d,\ single}} \rangle = \frac{1}{Z}(0 \cdot 1 + 1 \cdot 2\mathrm{e}^{-(E_{\mathrm{d}} - \mu)/k_{\mathrm{B}} T}). \tag{21}$$

If there is only one kind of donor doped into the material, and we assume that the concentration of donors N_{d} is small enough so that there is no interaction between donors, then the total number of electrons around donors is

$$n_{\mathrm{d}} = N_{\mathrm{d}} \cdot \langle n_{\mathrm{d,\ single}} \rangle = \frac{1}{1 + \frac{1}{2}\mathrm{e}^{(E_{\mathrm{d}} - \mu)/k_{\mathrm{B}} T}}. \tag{22}$$

3.3 Charge conservation

Since the expression of n_{d} and p_{a} are already known, we can invoke the charge conservation equation

$$n + n_{\mathrm{d}} - p - p_{\mathrm{a}} = N_{\mathrm{d}} - N_{\mathrm{a}}, \tag{23}$$

and

4 Spatial inhomogeneity

4.1 The theoretical framework

Suppose we are going to study a heterostructure containing two concatenated bulk systems. Below we take the usual assumptions required by Boltzmann equation and assume that the

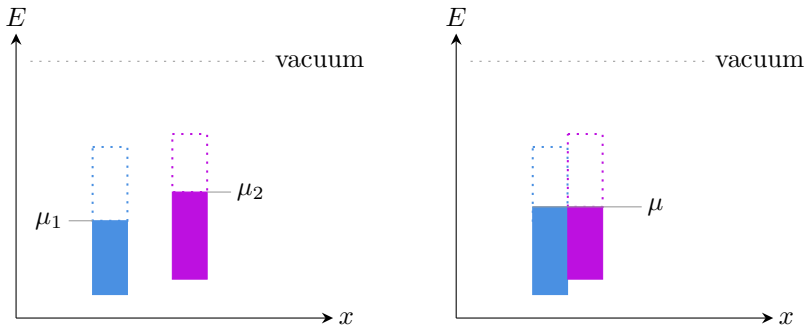

Figure 2: Two metals are placed together; since there is only one chemical potential in a system, the charge distribution changes.

spatial changes of the material properties are smooth enough, etc. so we can have well-defined $\boldsymbol{r}$ and $\boldsymbol{k}$ at the same time and don't need to worry about, say, interference.

The charge density is

$$\rho(\boldsymbol{r}) = -en(\boldsymbol{r}) + ep(\boldsymbol{r}), \tag{24}$$

and we have

$$\nabla^2\phi = -\frac{1}{\epsilon_0\epsilon_\mathrm{r}}\rho(\boldsymbol{r}). \tag{25}$$

Since we have picked up the Boltzmann assumptions, $n(\boldsymbol{r})$ and $p(\boldsymbol{r})$ can be decided from the bulk band structure ε_c and ε_v, μ and $\phi(\boldsymbol{r})$, and thus we obtain a set of self-consistent equilibrium equations, which are stationary cases of the whole Boltzmann transportation problem.

Actually, in order to make the Boltzmann assumptions correct, we need to assume a smooth transition region between the two types of material where the electronic structure is somehow a mixture of ε_c and ε_v, and indeed such a region exists, with a length scale of $\simeq 10\,\text{Å}$; but in analytic analysis, this region is usually ignored, and we still work with a stepwise change of electronic structure.

4.2 Example: junction between two metals

The case of a metal-metal junction is shown in Figure 2. Here we define the energy zero point to be the vacuum energy (i.e. the energy of the first electronic state going out of the material), and therefore the work function – the minimal energy needed to go out of the material – is just $|\mu|$.

After charge redistribution, we see that there are more electrons on the left side, and more holes on the right side. Thus, as a very crude model, we may assume that

$$\rho(x) = \begin{cases} -n_1, & -d_1 < x < 0, \\ n_2, & 0 < x < d_2, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

The potential ϕ can be immediately solved from $\rho(x)$: noticing that $\phi(\pm\infty)$ are not infinite, and hence when $x < -d_1$ and $x < d_2$, ϕ is a constant and can't have linear dependence on x, we have

$$\phi(x) = \begin{cases} \phi(-\infty), & x < -d_1, \\ \phi(\infty), & x > d_2, \\ \phi(-\infty) + \frac{n_1 e}{2\epsilon_1}(x+d_1)^2, & -d_1 < x < 0, \\ \phi(\infty) - \frac{n_2 e}{2\epsilon_2}(x-d_2)^2, & 0 < x < d_2. \end{cases} \tag{27}$$

By continuity condition and that $n_1 d_1 = n_2 d_2$, we are able to find n_1, n_2; as an estimation of order of magnitude, when $\epsilon_1 = \epsilon_2 = \epsilon$, we get

$$d_1 = \sqrt{\frac{2\epsilon\Delta}{e}\frac{n_2/n_1}{n_1+n_2}}, \tag{28}$$

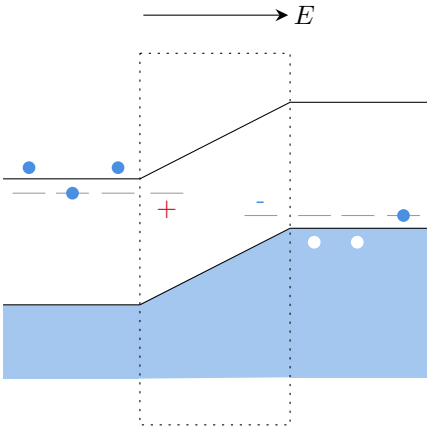

Figure 3: A p-n junction; the dotted line box is the depletion area. Note that since positive charges are accumulated on the left side in the depletion area, ϕ is higher on the left side, and thus $-e\phi$ is lower, and hence the shape of the bands.

and

$$d_2 = \sqrt{\frac{2\epsilon\Delta}{e}\frac{n_1/n_2}{n_1+n_2}}. \tag{29}$$

From usual properties of metals, usually $n_1 n_2 \sim 1\times10^{22}\,\mathrm{cm}^{-3}$, while $d_1 \sim 1\,\text{Å} - 3\,\text{Å}$. Note that here we deliberately don't specify that $\phi(\infty) = \phi(-\infty)$: the values obtained above work even when a voltage is applied to the junction. Note that d is within the aforementioned 10 Å mixed electronic structure region, which means the Boltzmann theory may fail for metal-metal junction; for semiconductor junctions however usually the $d \sim 100\,\text{Å}$, and there is no problem with that.

What?

Since the distribution of n and p always look like something times $\mathrm{e}^{-(E-e\phi-\mu)/k_\mathrm{B}T}$, we can attract $-e\phi$ into the band structure, and get $\varepsilon_{\boldsymbol{k}}(\boldsymbol{r})$; alternatively we can attract it into the chemical potential. In an equilibrium junction, the perspective of space-dependent chemical potential and the perspective of space-dependent band structure are *equivalent*: once one of them is chosen, the other has to be given up.

On the other hand, when an external electric field is applied, we should allow spatial dependence of *both* μ and $\varphi_{n\boldsymbol{k}}$. The point here is in when a diode is made, the charge density distribution deviates from the distribution in homogeneous p- or n-type materials, while when an external electric field is applied, we *don't* expect any change in charge distribution after a stable current is formed. We change the band distortion in the depletion area to model the decrease/increase of energy barrier, but then the chemical potential also has to have spatial dependence or otherwise the fact that the electron distributions in the p-part and n-part are the same can't be reflected.

What was I talking about??? This paragraph is weird.

4.3 Semiconductor-semiconductor p-n junction

Now consider a p-n junction, i.e. the junction between a p-semiconductor and an n-semiconductor. Similar to the case of metals, around the interface of the two, electrons and holes recombine, and when equilibrium is established, we have positive charge distribution on one side and negative charge distribution on the other side. The main difference between the p-n junction and the metal-metal junction is the former is always gapped; therefore, in the charge redistribution region, although we have non-zero charge distributions and an electric field is present, they are localized ones (like donor nucluei or attractor ions): we *no longer* have carriers. Thus the region is rightfully called the **depletion region**. Inside the depletion region we see band bending due to the electrostatic field established (Figure 3).

4.4 The p-n junction in circuit theory

It should be noted that the way of thinking is different in classical circuit analysis from the way in condensed matter physics: after the (usually stationary) relation between the quantities of a

system is solved, it is used as a *constraint* in circuit equations. In condensed matter physics, we talk about response functions, but in electronics the cause-effect relation is not emphasized. This is comparable to the way of thinking in scattering theory, where we focus on the scattering stationary states, and indeed, in quantum optics we also talk about scattering matrices, which can be derived from scattering stationary states and relates a_{in}'s and a_{out}'s. But the formalism in electronics is more generalized: non-unitary processes can also be modeled as constraints, the most famous example being the Ohm's law.