Contents

Chapter 3

Semiconductors and Insulators

3.1 Introduction

The Bloch energy band structure of noninteracting electrons in a periodic potential leads us to a broad classification of crystalline solids: (i) *metals*, in which the density of states $g(\varepsilon_{\rm F})$ at the Fermi level is nonzero, and (ii) *insulators*, where $g(\varepsilon_{\rm F}) = 0$. In an insulator, each Bloch band is either completely filled, corresponding to two electrons per unit cell, or completely empty. Thus, band insulators necessarily have an even number of electrons per unit cell.

3.1.1 Band gaps and transport

In the presence of a uniform electric field $\boldsymbol{E}$ at $T = 0$, a metal respond by generating a current density $\boldsymbol{j} = \sigma \boldsymbol{E}$, where σ is the conductivity matrix. In isotropic systems[1], $\sigma = ne^2\tau/m^*$, as we have seen. In an insulator, there is an energy gap, which is typically taken to mean $\sigma(T = 0, E) = 0$. In fact, this is not quite right, because there can be quantum tunneling between valence and conduction bands at finite E, a process known as *Zener tunneling*. For a direct gap between isotropic valence and conduction bands, the tunneling rate $T = 0$, *i.e.* the number density of electrons tunneling from valence to conduction band per unit time, is found to be[2]

$$\gamma(T = 0, E) = \frac{e^2 E^2 m_{\rm r}^{1/2}}{18\pi\hbar^2 \Delta^{1/2}} \exp\left(-\frac{\pi m_{\rm r}^{1/2} \Delta^{3/2}}{2\hbar e E}\right) \tag{3.1}$$

where $m_r = m_{\rm c}^* m_{\rm v}^*/(m_{\rm c}^* + m_{\rm v}^*)$ is the reduced mass of the valence holes and conduction electrons. Note that the current response is highly nonlinear, and furthermore that the exponential factor $\exp(-E_0/E)$, with $E_0 \equiv \pi m_{\rm r}^{1/2} \Delta^{3/2}/2e\hbar$, overwhelms the E^2 prefactor in the $E \to 0$ limit. Here,

$$E_0 = 5.7 \times 10^7 \, \frac{\rm V}{\rm cm} \times \big(\Delta[{\rm eV}]\big)^{3/2} \sqrt{\frac{m_{\rm r}}{m_{\rm e}}} \quad . \tag{3.2}$$

[1] With regard to tensors of rank two like $\sigma_{\alpha\beta}$, cubic symmetry is sufficient in order that the conductivity tensor be a multiple of the unit matrix.

[2] See E. O. Kane, *J. Phys. Chem. Solids* **12**, 181 (1959).

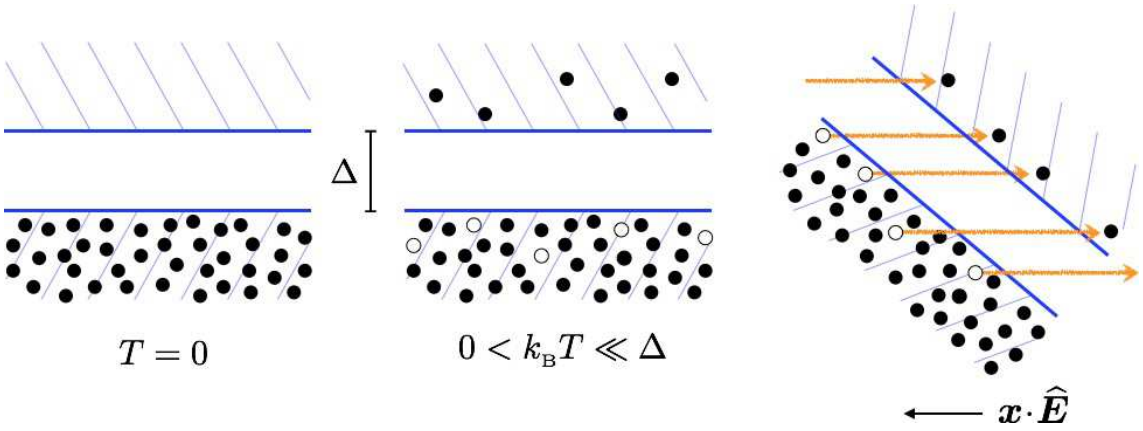


Figure 3.1: Schematics of electron occupation in semiconductor valence and conduction bands. Left: $T = 0$. Right: $0 < k_{\rm B}T \ll \Delta$. Right: Exaggerated schematic of Zener tunneling ($T = 0$).

The current density is then $j = e\gamma d$, where d is the thickness over which the field extends. At the level of *linear response*, though, the $T = 0$ conductivity of all insulators is zero.

At finite temperature, due to thermal fluctuations there is a finite electron density $n_{\rm c}$ in the conduction band, and a finite hole density $p_{\rm v}$ in the valence band, with, as we shall see, $n_{\rm c}(T) = p_{\rm v}(T) \propto \exp(-\Delta/2k_{\rm B}T)$. The conductivity is

$$\sigma(T) = \frac{n_{\rm c}(T)\,e^2\tau_{\rm c}}{m_{\rm c}^*} + \frac{p_{\rm v}(T)\,e^2\tau_{\rm v}}{m_{\rm v}^*} \propto e^{-\Delta/2k_{\rm B}T} \quad . \tag{3.3}$$

At $T = 300\,{\rm K}$, we have $k_{\rm B}T = 0.0258\,{\rm eV}$, so for an insulator like carbon diamond, for which $\Delta_{\rm Si} = 5.47\,{\rm eV}$ (indirect gap), $\Delta_{\rm Si}/2k_{\rm B}T = 106$, and the room temperature conductivity is essentially zero. Germanium, however, has a gap of $\Delta_{\rm Ge} = 0.66\,{\rm eV}$ (also indirect), hence $\Delta_{\rm Ge}/2k_{\rm B}T = 12.8$, and the Boltzmann weight is not nearly as small. You should know that the energy gap varies with temperature, mostly because anharmonic lattice vibrations cause the lattice to expand and the atomic positions to fluctuate. Typically one has $\Delta(T) \simeq \Delta(0) - ak_{\rm B}T$ with $a \approx 5$. Band gaps are also pressure-dependent, with $\Delta(p) \simeq \Delta(0) + bp$, with $b \approx 7 \times 10^{-9}\,{\rm eV/cm^2\,g}$. While there is no sharp distinction between semiconductors and insulators, at room temperature one typically classifies solids according to their conductivity, *viz.*

$$\begin{aligned}
\text{metals}: &\quad \rho(300\,{\rm K}) \lesssim 10^{-6}\,\Omega\cdot{\rm cm} \\
\text{semiconductors}: &\quad \rho(300\,{\rm K}) \in \left[10^{-3}\,\Omega\cdot{\rm cm}\,,\,10^{9}\,\Omega\cdot{\rm cm}\right] \\
\text{insulators}: &\quad \rho(300\,{\rm K}) \gtrsim 10^{12}\,\Omega\cdot{\rm cm} \quad .
\end{aligned}$$

The resistivity of metals and semiconductors depends on the scattering mechanisms which are responsible for momentum relaxation among the charge carriers.

Most semiconductors are covalently bonded crystals coming from column IV of the periodic table (*e.g.*, elemental semiconductors Si, Ge, and grey Sn), or compounds such as III-V materials (GaAs, GaP, InS, InP, GaSb, AlSb, *etc.*) and II-VI materials (PbS, PbSe, SnTe, *etc.*).

group	material	Δ (eV)	gap	m_c^*/m_e	m_v^*/m_e	ϵ	lattice constant (Å)	type
IV	C	5.47	I	0.2	0.25	5.7	3.567	D
IV	Si	1.12	I	1.64 (l) , 0.082 (t)	0.16 (l) , 0.49 (t)	11.9	5.431	D
IV	Ge	0.66	I	0.98 (l) , 0.19 (t)	0.04 (l) , 0.28 (t)	16.0	5.646	D
IV−IV	SiC	3.00	I	0.60	1.00	9.66	$a=3.086$, $c=15.117$	W
III−V	AlAs	2.36	I	0.11	0.22	10.1	5.661	Z
III−V	AlP	2.42	I	0.212	0.145	9.8	5.464	Z
III−V	AlSb	1.58	I	0.12	0.98	14.4	6.136	Z
III−V	GaAs	1.42	D	0.063	0.076 (lh) , 0.5 (hh)	12.9	5.653	Z
III−V	GaN	3.44	D	0.27	0.8	10.4	$a=3.189$, $c=10.4$	W
III−V	GaP	2.26	I	0.82	0.60	11.1	5.451	Z
III−V	GaSb	0.72	D	0.042	0.40	15.7	6.096	Z
III−V	InAs	0.36	D	0.023	0.40	15.1	6.058	Z
III−V	InP	1.35	D	0.077	0.64	12.6	5.869	Z
III−V	InSb	0.17	D	0.0145	0.40	16.8	6.479	Z
II−VI	CdS	2.5	D	0.14	0.51	5.4	5.825	Z
II−VI	CdS	2.49	D	0.20	0.7	9.1	$a=4.136$, $c=7.714$	W
II−VI	CdSe	1.70	D	0.13	0.45	10.0	6.050	Z
II−VI	ZnS	3.66	D	0.39	0.23	8.4	5.410	Z
II−VI	ZnS	3.78	D	0.287	0.49	9.6	$a=3.822$, $c=6.26$	W
IV−VI	PbS	041	I	0.25	0.25	17.0	5.936	R
IV−VI	PbTe	0.31	I	0.17	0.20	30.0	6.462	R

Table 3.1: Common semiconductors and their properties at $T = 3000\,\mathrm{K}$. Gap types: D (direct) and I (indirect). Structure: D (Diamond), W (Wurtzite), Z (Zincblende), R (Rocksalt). Hole masses: hh (heavy hole), lh (light hole). Source: S. M. Sze, *Physics of Semiconductors*.

3.1.2 Hall effect

High field Hall effect measurements, which give $\sigma_{xy} = (p_v - n_c)\, ec/B$, may be used to obtain an independent measurement of the carrier concentration without requiring knowledge of the scattering times τ_v and τ_c, which appear in the diagonal conductivity σ_{xx}. Of course, for a pure (*i.e. intrinsic*) semiconductor, $n_c = p_v$, but in the *extrinsic* case, impurities (*i.e. dopants*) lead to the condition $n_c \neq p_v$, as we shall see. Such independent measurements of carrier concentration confirm that the rapid changes in $\sigma_{xx}(T)$ are predominantly due to variations in carrier concentration.

3.1.3 Optical absorption

The energy gap Δ in a semiconductor may be measured by the temperature dependence of $\ln \sigma_{xx}(T)$, but more directly by optical absorption. When a photon of energy $\hbar\omega > \Delta$ is absorbed by a semiconductor,

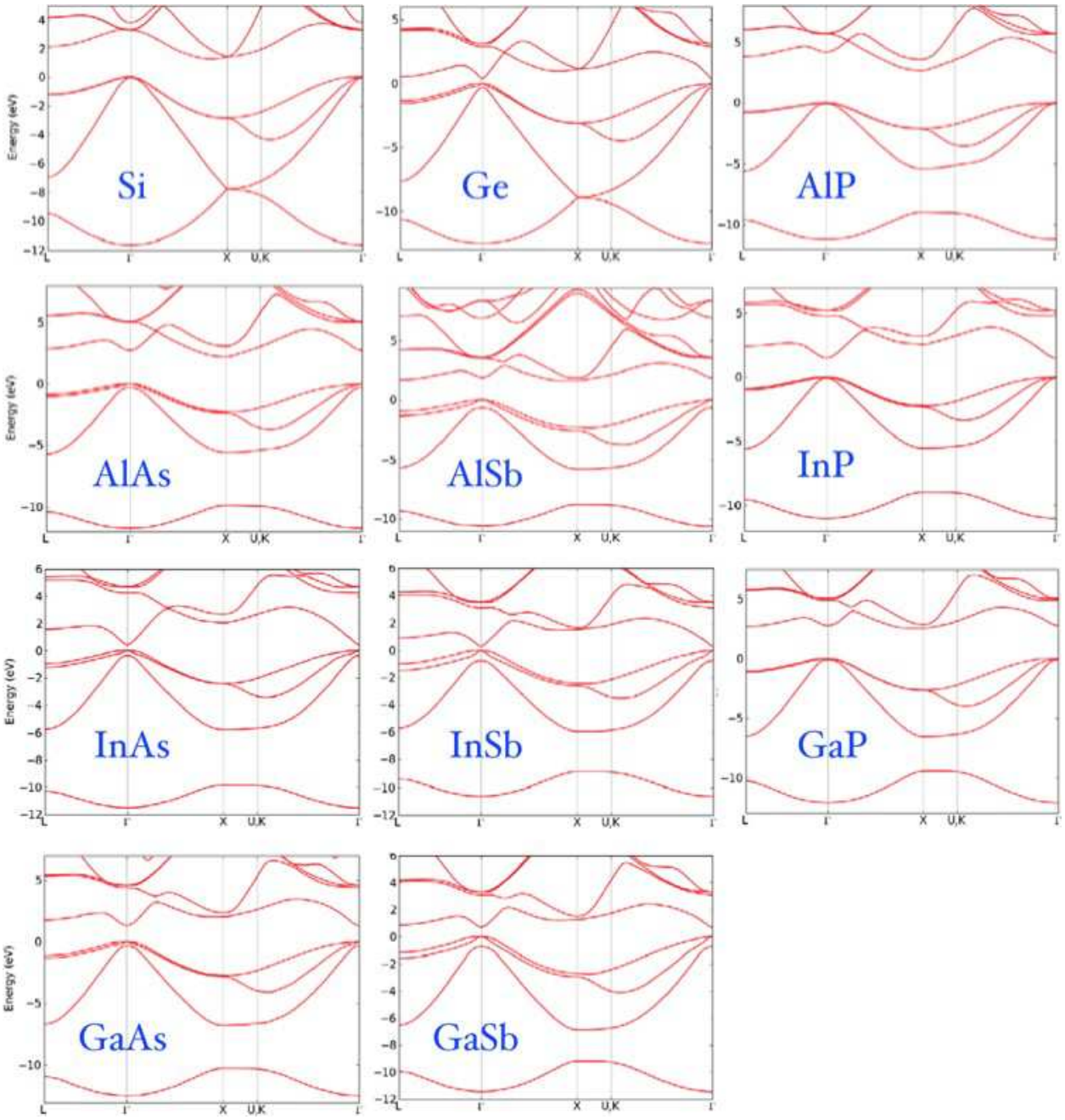
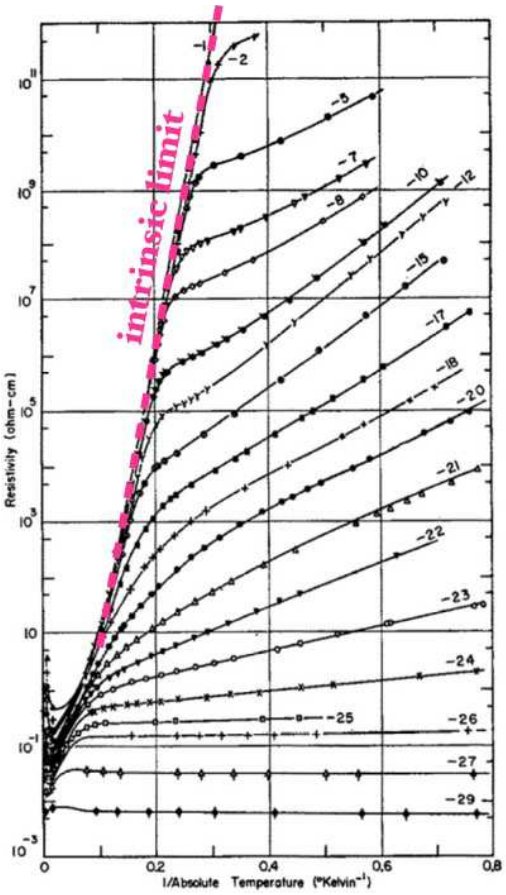


Figure 3.2: Pseudopotential calculation of band structures of diamond and zincblende semiconductors, with spin-orbit effects included. From B. D. Malone and M. L. Cohen, *J. Phys. Condens. Matter* **25**, 105503 (2013).

it creates an conduction electron - valence hole pair, as depicted in Fig. 3.4. At the simplest level of description, the absorption edge coincides with the band gap. At a greater level of refinement, the Coulomb interaction between conduction electron and valence hole must be accounted for, and results



Sample code number	N_D (cm^{-3})
−1	$5{\cdot}3\times10^{14}$
−2	$9{\cdot}3\times10^{14}$
−5	$1{\cdot}6\times10^{15}$
−7	$2{\cdot}3\times10^{15}$
−8	$3{\cdot}0\times10^{15}$
−10	$5{\cdot}2\times10^{15}$
−12	$8{\cdot}5\times10^{15}$
−15	$1{\cdot}3\times10^{16}$
−17	$2{\cdot}4\times10^{16}$
−18	$3{\cdot}5\times10^{16}$
−20	$4{\cdot}5\times10^{16}$
−21	$5{\cdot}5\times10^{16}$
−22	$6{\cdot}4\times10^{16}$
−23	$7{\cdot}4\times10^{16}$
−24	$8{\cdot}4\times10^{16}$
−25	$1{\cdot}2\times10^{17}$
−26	$1{\cdot}3\times10^{17}$
−27	$2{\cdot}7\times10^{17}$
−29	$9{\cdot}5\times10^{17}$

Figure 3.3: Resistivity of antimony-doped germanium as a function of $1/T$ for varying impurity concentrations. Table at right correlates donor density N_D with sample code number. The magenta "intrinsic limit" line corresponds to the limit $N_D \to 0$. Data from H. J. Fritzsche, *J. Phys. Chem. Solids* **6**, 69 (1958).

in structure to the absorption curve below the Δ threshold. Since $\hbar c = 1973\,\mathrm{eV}{\cdot}\text{Å}$, the wavelength of light at the band gap energy is

$$\lambda = \frac{2\pi\hbar c}{\Delta} = \frac{12400\,\text{Å}}{\Delta[\mathrm{eV}]} \quad . \tag{3.4}$$

Since both energy and momentum must be conserved[3], we have separately $\boldsymbol{k}_\gamma = \boldsymbol{k}_\mathrm{e} - \boldsymbol{k}_\mathrm{h}$ as well as $\hbar c k_\gamma = \Delta$. Thus, $|\boldsymbol{k}_\mathrm{e} - \boldsymbol{k}_\mathrm{h}| = \Delta/\hbar c = \Delta[\mathrm{eV}]\,/\,1973\,\mathrm{eV}\cdot\text{Å}$. Typically Δ is on the order of eV, hence the difference in electron and hole wavevectors is on the order of a milli-Ångstrom, which is insignificant on the scale of the Brillouin zone. Thus, *optical transitions are vertical*, meaning they involve no change in wavevector for the electron as it is excited from valence to conduction band. The reason is that the speed of light is very big.

[3]In fact, only *crystal momentum* must be conserved, meaning the wavevector $\boldsymbol{k}$ is conserved modulo any reciprocal lattice vector.

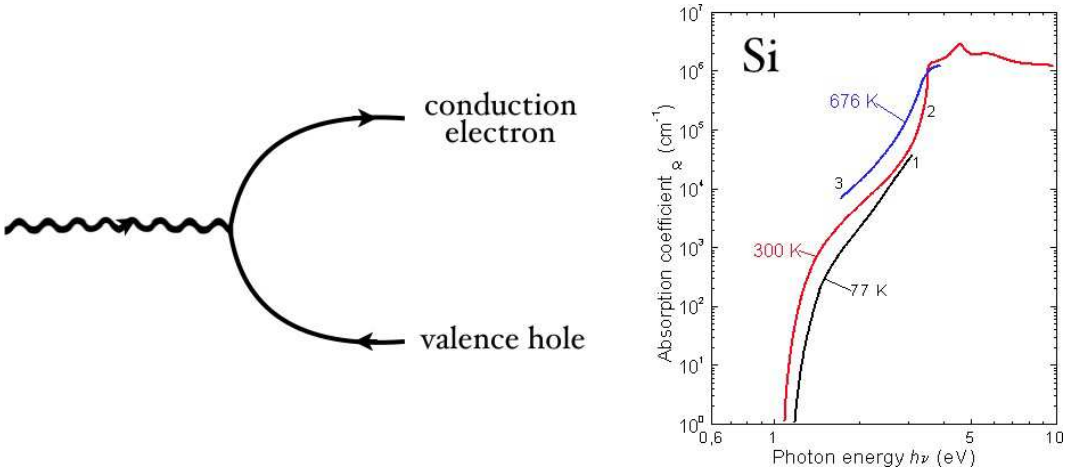


Figure 3.4: Left: A photon with $\hbar\omega > \Delta$ creating an electron-hole pair. Right: Optical conductivity of Si ($\Delta(0\,\mathrm{K}) = 1.17\,\mathrm{eV}$).

Under a constant flux of light, the carrier density $n_{\rm c} = p_{\rm v} \equiv n$ obeys

$$\frac{dn}{dt} = \alpha - \beta n^2 \quad . \tag{3.5}$$

The first term accounts for the creation of $\mathrm{e}-\mathrm{h}$ pairs due to photoexcitation. The second term accounts for the *recombination* of photoexcited $\mathrm{e}-\mathrm{h}$ pairs. In equilibrium, $n = (\alpha/\beta)^{1/2}$.

3.1.4 Direct *versus* indirect gaps

Fig. 3.5 shows the cases of *direct* and *indirect* gap semiconductors. In a direct gap material, the conduction band minimum and the valence band maximum occur at the same point (or points) in the Brillouin zone. In an indirect gap material, there is a difference $\boldsymbol{Q} = \boldsymbol{k}_{\rm c}^{\rm min} - \boldsymbol{k}_{\rm v}^{\rm max}$ (modulo $\boldsymbol{G}$). Examples of direct gap materials include $\alpha-$Sn, Se, Te, GaAs, and ZnS. Examples of indirect gap materials include Si, Ge, AlSb, GaP, and PbTe. In an indirect gap material, something other then the photon must supply the missing momentum $\hbar\boldsymbol{Q}$ in order for the material to absorb light at the band gap, and that something is usually a phonon (*i.e.* a lattice vibration). Since phonon frequencies are on the order of meV (Debye temperatures hundreds of Kelvins, $k_{\rm B} = 86.2\,\mu\mathrm{eV/K}$), the additional phonon energy is small compared with Δ, except perhaps in narrow gap materials. Since the number of phonons vanishes as T^3 at low temperatures, the optical absorption at $\hbar\omega = \Delta$ will be temperature dependent.

3.1.5 Mobility

Mobility μ is defined by the relation of the diffusional drift velocity to the applied electric field strength:

$$\boldsymbol{v}_{\rm drift} = \mp\mu\boldsymbol{E} \quad , \tag{3.6}$$

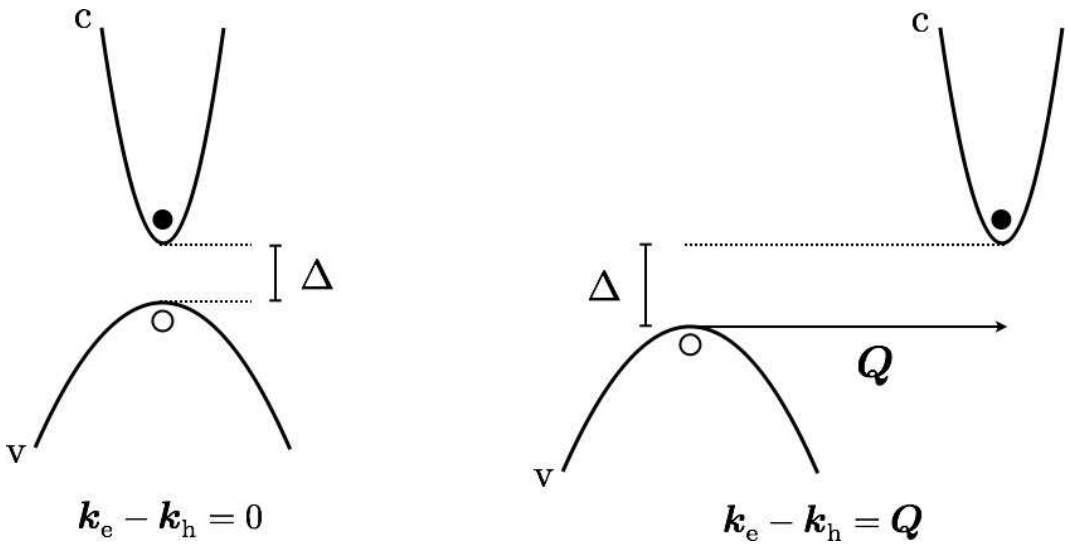


Figure 3.5: Left: A direct gap semiconductor with a low-energy particle-hole pair. The net crystal momentum is $\boldsymbol{K} = \boldsymbol{k}_{\rm e} - \boldsymbol{k}_{\rm h} = 0$. Right: An indirect gap semiconductor with a low-energy particle-hole pair. The net crystal momentum is $\boldsymbol{K} = \boldsymbol{k}_{\rm e} - \boldsymbol{k}_{\rm h} = \boldsymbol{Q}$, where $\boldsymbol{Q}$ is the wavevector difference between valence band maximum and conduction band minimum.

with the upper sign holding for electrons and the lower sign for holes. The current density is $\boldsymbol{j} = nq\boldsymbol{v}_{\rm drift} = n|q|\mu\boldsymbol{E}$, where q is the charge. Hence we have $\sigma = n|q|\mu$. When both signs of carriers are present,

$$\sigma = n_{\rm c}\, e\, \mu_{\rm c} + p_{\rm v}\, e\, \mu_{\rm v} \quad . \tag{3.7}$$

3.1.6 Effective mass

Generally speaking, the dispersion in the vicinity of a local extremum in the conduction (maximum) or valence (minimum) band dispersions obeys

$$\begin{aligned} E_{\rm c}(\boldsymbol{k}) &= E_{\rm c}^0 + \tfrac{1}{2}\hbar^2\,(m_{\rm c}^*)^{-1}_{\alpha\beta}\,(k^\alpha - Q_{\rm c}^\alpha)(k^\beta - Q_{\rm c}^\beta) + \ldots \\ E_{\rm v}(\boldsymbol{k}) &= E_{\rm v}^0 - \tfrac{1}{2}\hbar^2\,(m_{\rm v}^*)^{-1}_{\alpha\beta}\,(k^\alpha - Q_{\rm v}^\alpha)(k^\beta - Q_{\rm v}^\beta) + \ldots \quad , \end{aligned} \tag{3.8}$$

where $\boldsymbol{Q}_{\rm c,v}$ is the wavevector of the extremum, and $m^*_{\rm c,v}$ is the effective mass tensor. The band gap is given by $\Delta = E_{\rm c}^0 - E_{\rm v}^0$.

Since the effective mass tensors are each symmetric, they may be diagonalized along their principal axes, in which case we may write

$$E(\boldsymbol{k}) = E^0 \pm \left(\frac{\hbar^2(\Delta k_1)^2}{2m_1^*} + \frac{\hbar^2(\Delta k_2)^2}{2m_2^*} + \frac{\hbar^2(\Delta k_3)^2}{2m_3^*} \right) + \mathcal{O}\big((\Delta k)^4\big) \quad , \tag{3.9}$$

where $\Delta\boldsymbol{k} = \boldsymbol{k} - \boldsymbol{Q}$. *Nota bene* : the tensors $m_{\rm c}^*$ and $m_{\rm v}^*$ may not commute, in which case their principal axes do not coincide, and they may not both be rendered diagonal in the same basis. Furthermore,

since the extrema of $E_{\rm c,v}(\boldsymbol{k})$ may not lie at the Γ point, the level sets of $E_{\rm c,v}(\boldsymbol{k})$ may not be connected, *i.e.* they may consist of several disjoint components. Indeed this is the case in silicon, where the six equivalent conduction band minima lie along the ΓX directions ($\langle 100 \rangle$). In germanium, the conduction band minima occur at the fourfold degenerate L point. Oftentimes, as in the cases of Si, CdTe, InSb, and several other materials, there are more than one electron or hole bands in play.

3.2 Number of Carriers in Thermal Equilibrium

We define

$$\begin{aligned} n_{\rm c}(T,\mu) &= \text{number density of electrons in the conduction band} \\ p_{\rm v}(T,\mu) &= \text{number density of holes in the valence band} \quad . \end{aligned}$$

Quantum thermodynamics in the grand canonical ensemble then says

$$n_{\rm c}(T,\mu) = \int\limits_{E_{\rm c}^0}^{\infty} d\varepsilon \; g_{\rm c}(\varepsilon)\, f(\varepsilon - \mu) \tag{3.10}$$

and

$$p_{\rm v}(T,\mu) \int\limits_{-\infty}^{E_{\rm v}^0} d\varepsilon \; g_{\rm v}(\varepsilon) \Big\{ 1 - f(\varepsilon - \mu) \Big\} = \int\limits_{-\infty}^{E_{\rm v}^0} d\varepsilon \; g_{\rm v}(\varepsilon)\, f(\mu - \varepsilon) \quad , \tag{3.11}$$

where $f(u) = \big[\exp(u/k_{\rm B}T) + 1 \big]^{-1}$ is the Fermi distribution. For quadratic extrema, the conduction and valence band densities of states behave as $g_{\rm c}(\varepsilon) \propto (\varepsilon - E_{\rm c}^0)^{1/2}$ and $g_{\rm v}(\varepsilon) \propto (E_{\rm v}^0 - \varepsilon)^{1/2}$. We can define the function $\bar{f}(u) \equiv f(-u)$ to be the Fermi distribution function for holes.

The dependence of $n_{\rm c}(T,\mu)$ and $p_{\rm v}(T\,\mu)$ on the chemical potential in Eqns. 3.10 and 3.11 is complicated. In the Maxwell-Boltzmann limit, however, where $|E_{\rm c,v}^0 - \mu| \gg k_{\rm B}T$, the Fermi functions become $f(\varepsilon - \mu) \simeq e^{-(\varepsilon-\mu)/k_{\rm B}T}$ for $\varepsilon > E_{\rm c}^0$ and $f(\mu - \varepsilon) \simeq e^{-(\mu-\varepsilon)/k_{\rm B}T}$ for $\varepsilon < E_{\rm v}^0$. Thus, we may write

$$\begin{aligned} n_{\rm c}(T,\mu) &= N_{\rm c}(T)\, e^{-(E_{\rm c}^0 - \mu)/k_{\rm B}T} \\ p_{\rm v}(T,\mu) &= P_{\rm v}(T)\, e^{-(\mu - E_{\rm v}^0)/k_{\rm B}T} \quad , \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} N_{\rm c}(T) &= \int\limits_{E_{\rm c}^0}^{\infty} d\varepsilon \; g_{\rm c}(\varepsilon)\, e^{-(\varepsilon - E_{\rm c}^0)/k_{\rm B}T} \\ P_{\rm v}(T) &= \int\limits_{-\infty}^{E_{\rm v}^0} d\varepsilon \; g_{\rm v}(\varepsilon)\, e^{-(E_{\rm v}^0 - \varepsilon)/k_{\rm B}T} \quad . \end{aligned} \tag{3.13}$$

Now for ellipsoidal bands as in Eqn. 3.9, the density of states, including spin degeneracy, is given by

$$g(\varepsilon) = \frac{\sqrt{2}}{\pi^2\hbar^3}\,\sqrt{m_1^*\,m_2^*\,m_3^*}\;\varepsilon^{1/2}\,\Theta(\varepsilon) \quad , \tag{3.14}$$

in which case

$$N_{\rm c}(T) = 2\,\lambda_{T,{\rm c}}^{-3} \qquad , \qquad P_{\rm v}(T) = 2\,\lambda_{T,{\rm v}}^{-3} \quad , \tag{3.15}$$

where the thermal wavelengths are given by $\lambda_{T,{\rm c/v}} = \left(2\pi\hbar^2/m^*_{\rm c/v}k_{\rm B}T\right)^{1/2}$, with the DOS mass $m^*_{\rm c/v}$ given by $m^*_{\rm c/v} = \left(m^*_{1,{\rm c/v}}\,m^*_{2,{\rm c/v}}\,m^*_{3,{\rm c/v}}\right)^{1/3}$. It is convenient to express the quantities $N_{\rm c}(T)$ and $P_{\rm v}(T)$ as

$$\begin{aligned} N_{\rm c}(T) &= 2.51\times10^{19}\,{\rm cm}^{-3}\left(\frac{m_{\rm c}^*}{m_{\rm e}}\right)^{3/2}\left(\frac{T}{300\,{\rm K}}\right)^{3/2} \\ P_{\rm v}(T) &= 2.51\times10^{19}\,{\rm cm}^{-3}\left(\frac{m_{\rm v}^*}{m_{\rm e}}\right)^{3/2}\left(\frac{T}{300\,{\rm K}}\right)^{3/2} \quad , \end{aligned} \tag{3.16}$$

which, along with Eqn. 3.12, tells us that $10^{19}\,{\rm cm}^{-3}$ is an approximate upper limit to the carrier concentration in a nondegenerate semiconductor[4].

3.2.1 Intrinsic semiconductors

How do we find the chemical potential $\mu(T)$? In an *intrinsic* semiconductor, the number of conduction electrons must be equal to the number of valence holes, hence

$$n_{\rm c}(T,\mu) = p_{\rm v}(T,\mu) \quad , \tag{3.17}$$

which is one equation in the two unknowns (T,μ). The solution set is thus the desired function $\mu(T)$. The above equation is difficult to solve owing to the complicated dependence of the integrals in Eqns. 3.10 and 3.11 on μ, but if we are in the Maxwell-Boltzmann limit, a solution is readily available. Writing

$$n_{\rm c}(T,\mu) = N_{\rm c}(T)\,e^{-(E_{\rm c}^0-\mu)/k_{\rm B}T} = P_{\rm v}(T)\,e^{-(\mu-E_{\rm c}^0)/k_{\rm B}T} = p_{\rm v}(T,\mu) \quad , \tag{3.18}$$

we have

$$e^{(2\mu-E_{\rm c}^0-E_{\rm v}^0)/k_{\rm B}T} = \frac{P_{\rm v}(T)}{N_{\rm c}(T)} = \left(\frac{m_{\rm v}^*}{m_{\rm c}^*}\right)^{3/2} \quad , \tag{3.19}$$

and we find

$$\mu(T) = \tfrac{1}{2}\big(E_{\rm c}^0 + E_{\rm v}^0\big) + \tfrac{3}{4}\,k_{\rm B}T\,\ln\!\left(\frac{m_{\rm v}^*}{m_{\rm c}^*}\right) \quad . \tag{3.20}$$

As $T\to0$, the chemical potential tends to the average $\mu(0) = \frac{1}{2}\big(E_{\rm c}^0 + E_{\rm v}^0\big)$. For finite temperature T, $\mu(T)$ increases with temperature if $m_{\rm v}^* > m_{\rm c}^*$ and decreases with T if $m_{\rm v}^* < m_{\rm c}^*$. Since the ratio $m_{\rm v}^*/m_{\rm c}^*$ is of order unity, we have $|\mu - E_{\rm c,v}^0| = \frac{1}{2}\,\Delta + \mathcal{O}(1)\cdot k_{\rm B}T$, and provided $k_{\rm B}T \ll \Delta$, the degeneracy condition applies. In most semiconductors, $\Delta \gg k_{\rm B}T_{\rm room}$, so intrinsic semiconductors are almost always degenerate at and below room temperature.

[4] "Nondegenerate" means $|E_{\rm c,v}^0 - \mu| \gg k_{\rm B}T$.

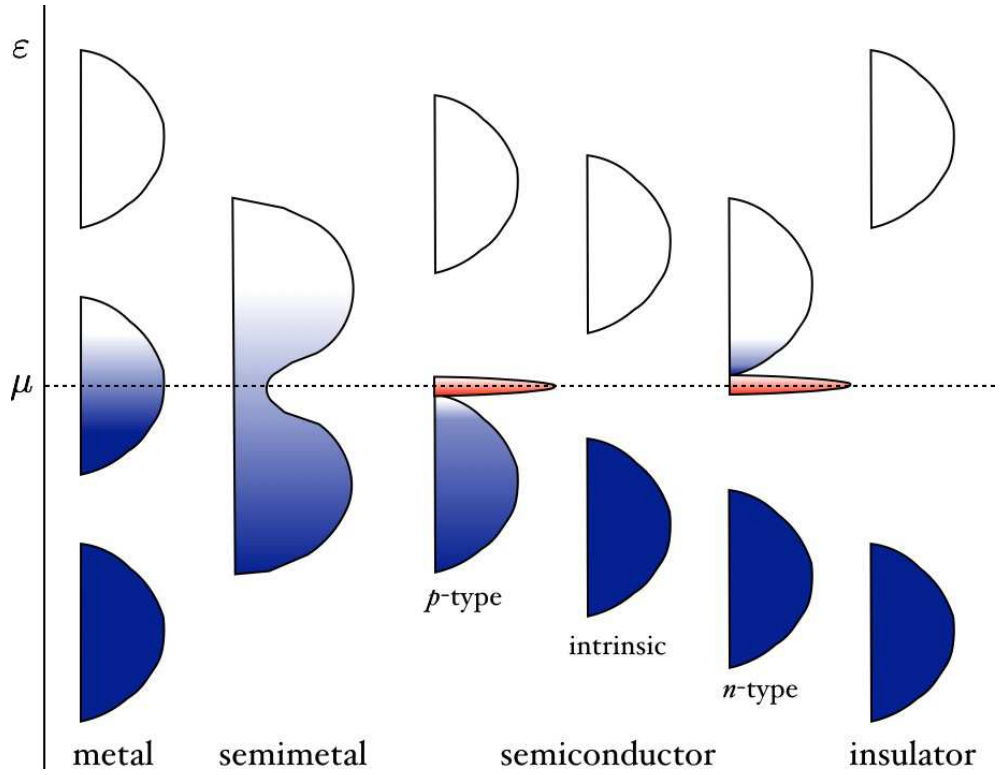


Figure 3.6: Bands and their fillings in metals, semiconductors, and insulators. Narrow red-shaded regions denote impurity bands of acceptors (p-type) and donors (n-type).

3.2.2 Extrinsic semiconductors

In an extrinsic semiconductor, *dopant* ions with energy levels just above the valence band (acceptors) or just below the conduction band (donors) contribute valence holes or conduction electrons, respectively, and there is an imbalance $n_{\rm c} - p_{\rm v} = \Delta n$. When acceptors are present, the chemical potential lies close to the valence band maximum, and the system is said to be p-type. The charge carriers are then valence holes. When donors are present, the chemical potential lies close to the conduction band minimum, and the system is said to be n-type. The charge carriers are then conduction electrons. The densities $n_{\rm c}$ and $p_{\rm v}$ denote only conduction electrons and valence holes, and do not include contributions from impurity states. Regardless of the shift in μ due to extrinsic effects, in the Maxwell-Boltzmann limit the product $n_{\rm c}(T,\mu)\,p_{\rm v}(T,\mu)$ is independent of μ, and given by

$$n_{\rm c}(T,\mu)\,p_{\rm v}(T,\mu) = N_{\rm c}(T)\,P_{\rm v}(T)\,e^{-(E_{\rm c}^0 - E_{\rm v}^0)/k_{\rm B}T} \equiv n_{\rm i}^2(T) \quad , \tag{3.21}$$

where

$$\begin{aligned} n_{\rm i}(T) &\equiv 2\,\bar{\lambda}_T^{-3/2}\,e^{-\Delta/2k_{\rm B}T} \\ &= 2.5\times 10^{19}\,{\rm cm}^{-3}\left(\frac{\sqrt{m_{\rm v}^*\,m_{\rm c}^*}}{m_{\rm e}}\right)^{3/2}\left(\frac{T}{300\,{\rm K}}\right)^{3/2}e^{-\Delta/2k_{\rm B}T} \quad , \end{aligned} \tag{3.22}$$

II	III	IV	V	VI
	B [He] $2s^2\,2p^1$	C [He] $2s^2\,2p^2$	N [He] $2s^2\,2p^3$	O [He] $2s^2\,2p^4$
	Al [Ne] $3s^2\,3p^1$	Si [Ne] $3s^2\,3p^2$	P [Ne] $3s^2\,3p^3$	S [Ne] $3s^2\,3p^4$
Zn [Ar] $4s^2\,3d^{10}$	Ga [Ar] $4s^2\,3d^{10}\,4p^1$	Ge [Ar] $4s^2\,3d^{10}\,4p^2$	As [Ar] $4s^2\,3d^{10}\,4p^3$	Se [Ar] $4s^2\,3d^{10}\,4p^4$
Cd [Kr] $5s^2\,4d^{10}$	In [Kr] $5s^2\,4d^{10}\,5p^1$	Sn [Kr] $5s^2\,4d^{10}\,5p^2$	Sb [Kr] $5s^2\,4d^{10}\,5p^3$	Te [Kr] $5s^2\,4d^{10}\,5p^4$
Hg [Xe] $6s^2\,5d^{10}$	Tl [Xe] $6s^2\,5d^{10}\,6p^1$	Pb [Xe] $6s^2\,5d^{10}\,6p^2$	Bi [Xe] $6s^2\,5d^{10}\,6p^3$	Po [Xe] $6s^2\,5d^{10}\,6p^4$

Figure 3.7: Relevant group II through group VI elements and their electronic configurations.

with $\bar{\lambda}_T \equiv \sqrt{\lambda_{T,\rm c}\,\lambda_{T,\rm v}}$.

In the extrinsic case, then, $n_{\rm c} - p_{\rm v} = \Delta n$ and $n_{\rm c}\,p_{\rm v} = n_{\rm i}^2$ are two equations in two unknowns, with the solution

$$\begin{aligned} n_{\rm c} &= \sqrt{n_{\rm i}^2 + \tfrac{1}{4}(\Delta n)^2} + \tfrac{1}{2}\,\Delta n \\ p_{\rm v} &= \sqrt{n_{\rm i}^2 + \tfrac{1}{4}(\Delta n)^2} - \tfrac{1}{2}\,\Delta n \quad . \end{aligned} \tag{3.23}$$

If we furthermore define the quantity $\mu_{\rm i}$ such that

$$n_{\rm c}(T,\mu) \equiv n_{\rm i}(T)\,e^{(\mu-\mu_{\rm i})/k_{\rm B}T} \qquad , \qquad p_{\rm v}(T,\mu) \equiv n_{\rm i}(T)\,e^{(\mu_{\rm i}-\mu)/k_{\rm B}T} \quad , \tag{3.24}$$

then the quantities $(\Delta n\,,\,\mu_{\rm i})$ are related by

$$\Delta n = 2\,n_{\rm i}(T)\,\sinh\!\left(\frac{\mu-\mu_{\rm i}}{k_{\rm B}T}\right) \quad . \tag{3.25}$$

Now if $\Delta n/n_{\rm i}$ is small, then $|\mu-\mu_{\rm i}| \ll k_{\rm B}T$, and if the degeneracy condition applies, this means that μ is far from $E^0_{\rm c,v}$, and both $n_{\rm c} \approx p_{\rm v} \approx n_{\rm i}$. This remains the case so long as $|\Delta n| \ll n_{\rm i}$.

In the opposite limit, when $|\Delta n| \gg n_{\rm i}$, we have

$$\sqrt{n_{\rm i}^2 + \tfrac{1}{4}(\Delta n)^2} = \tfrac{1}{2}|\Delta n| + \frac{n_{\rm i}^2}{|\Delta n|} + \ldots \quad , \tag{3.26}$$

and therefore

$$\begin{aligned} n\text{-type} \quad &: \quad \frac{\Delta n}{n_{\rm i}} \gg +1 \qquad \Rightarrow \qquad n_{\rm c} = \Delta n \quad , \quad p_{\rm v} = \frac{n_{\rm i}^2}{\Delta n} \\ p\text{-type} \quad &: \quad \frac{\Delta n}{n_{\rm i}} \ll -1 \qquad \Rightarrow \qquad n_{\rm c} = \frac{n_{\rm i}^2}{|\Delta n|} \quad , \quad p_{\rm v} = |\Delta n| \quad . \end{aligned} \tag{3.27}$$

3.3 Donors and Acceptors

3.3.1 Impurity charges in a semiconductor

Silicon is a group IV element in the periodic table. To its left sits aluminum and to its right sits phosphorus. Consider a Si crystal in which one of the Si atoms has been replaced by a P atom. Compared to silicon, phosphorus has one extra nuclear charge and one additional electron. In free space, this last P electron has a binding energy of $10.5\,$eV, the first ionization potential of P. In a crystal, this binding energy is drastically reduced, due to two effects:

- The static dielectric constant of the semiconductor crystal is typically large ($\epsilon_{\rm Si} = 11.9$, $\epsilon_{\rm InSb} = 16.8$, $\epsilon_{\rm PbTe} = 30$). Small gaps lead to large dielectric constants[5]. Later on we shall derive the expression

$$\epsilon \lesssim 1 + \left(\frac{\hbar\omega_{\rm pv}}{\Delta}\right)^2 \quad , \tag{3.28}$$

  where $\omega_{\rm pv} = (4\pi n_{\rm v} e^2/m_{\rm e})^{1/2}$, with $n_{\rm v}$ the number density of valence electrons. So the attraction between the phosphorus core and the added electron is reduced by a factor of ϵ, provided the radius of the electronic orbit is on the order of several lattice spacings.

- The effective mass of the electrons in the conduction band is $m_{\rm c}^*$, which is generally about a tenth of the electron mass $m_{\rm e}$.

The radius of the orbit is thus not $a_{\rm B} = \hbar^2/me^2$, but rather

$$r_0 = \frac{\epsilon\,\hbar^2}{m_{\rm c}^* e^2} = \frac{m_{\rm e}}{m_{\rm c}^*}\,\epsilon\,a_{\rm B} \quad . \tag{3.29}$$

If $\epsilon \approx 10$ and $m_{\rm c}^* \approx 0.1\,m_{\rm e}$, we have $r_0 \approx 100\,a_{\rm B}$. The binding energy W of the lowest hydrogenic state is then given by

$$W_{\rm d} = \frac{e^2}{2\epsilon\,r_0} = \overbrace{\frac{e^2}{2a_{\rm B}}}^{13.6\,{\rm eV}} \cdot \overbrace{\frac{m_{\rm c}^*}{m_{\rm e}}\frac{1}{\epsilon^2}}^{\approx 10^{-3}} \tag{3.30}$$

[5] In a metal, where ther is no gap, $\epsilon = \infty$.

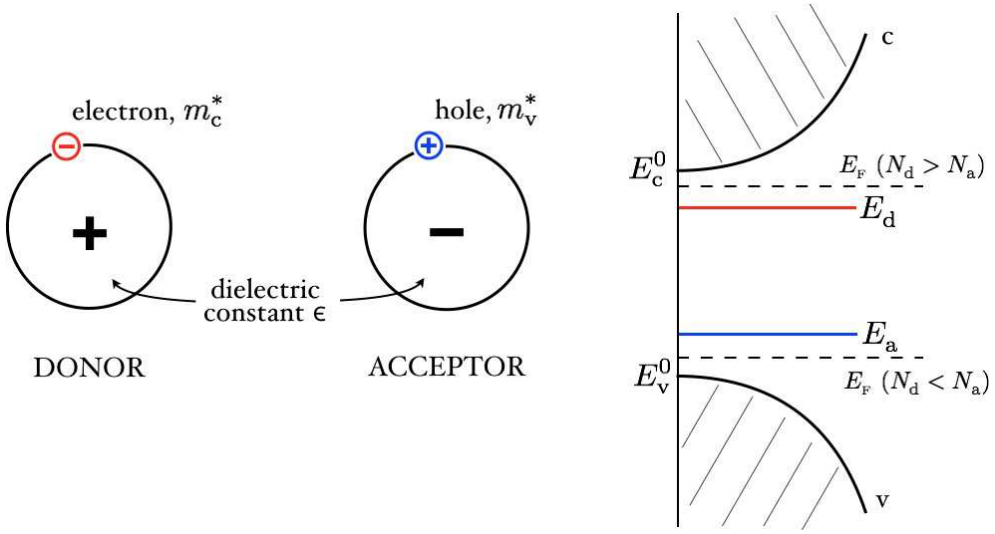


Figure 3.8: Donors and acceptors.

Thus, the donor binding energy is $W_{\rm d} \approx 10^{-3}\,$Ry, which is usually very small in comparison with the band gap Δ. This means that the donor levels lie just below the conduction band minimum $E_{\rm c}^0$, as depicted in Fig. 3.8. The calculation for hole binding by acceptors is identical, aside from the replacement of $m_{\rm c}^*$ by $m_{\rm v}^*$. Thus,

$$\begin{aligned} E_{\rm d} &= E_{\rm c}^0 - \frac{e^2}{2a_{\rm B}} \cdot \frac{m_{\rm c}^*}{m_{\rm e}} \frac{1}{\epsilon^2} \\ E_{\rm a} &= E_{\rm v}^0 + \frac{e^2}{2a_{\rm B}} \cdot \frac{m_{\rm v}^*}{m_{\rm e}} \frac{1}{\epsilon^2} \quad . \end{aligned} \tag{3.31}$$

		group III : acceptors						group V : donors					
	$\Delta(300\,{\rm K})$	$W_{\rm a}$	B	Al	Ga	In	Tl	$W_{\rm d}$	N	P	As	Sb	Bi
Si	1.12	48	45.0	68.5	71	155	245	113	140	45.3	53.7	42.7	70.6
Ge	0.67	15	10.8	11.1	11.3	12.0	13.5	28	–	12.9	14.2	10.3	12.8

Table 3.2: Donor and acceptor binding energies in Si and Ge (in meV).

3.3.2 Donor and acceptor quantum statistics

In the presence of donor and acceptor ions, the net change in the background ionic charge density is given by $\Delta\rho_{\rm ion} = e\,(N_{\rm d} - N_{\rm a})$. Since the net system is charge neutral, this must be balanced by the net electronic charge density, $\Delta\rho_{\rm elec} = -e\,(n_{\rm c} + n_{\rm d} - p_{\rm v} - p_{\rm a})$ where $n_{\rm d}$ and $p_{\rm a}$ are the number densities of donor electrons and acceptor holes in equilibrium, respectively. The charge neutrality condition is then

$\vert\,\Psi\,\rangle$	E	$\hat{n}$
$\vert\,0\,\rangle$	0	0
$\vert\uparrow\rangle\,,\vert\downarrow\rangle$	$E_{\rm d}$	1
$\vert\uparrow\downarrow\rangle$	$2E_{\rm d}+U$	2

Table 3.3: Donor states and their energies.

$\Delta\rho_{\rm ion}+\Delta\rho_{\rm elec}=0$, which requires

$$n_{\rm c}-p_{\rm v}+n_{\rm d}-p_{\rm a}=N_{\rm d}-N_{\rm a}\quad . \tag{3.32}$$

The question now arises of how to compute $n_{\rm d}$ and $p_{\rm a}$. The simplest assumption is to assume the donor and acceptor levels of each spin are independently occupied according to a Fermi distribution, just like the conduction and valence band levels. Under this assumption,

$$n_{\rm d}=2N_{\rm d}\,f(E_{\rm d}-\mu)\qquad,\qquad p_{\rm a}=2N_{\rm a}\,f(\mu-E_{\rm a})\quad, \tag{3.33}$$

where the factor of 2 comes from spin degeneracy. However, donor and acceptor state wavefunctions are *localized* in space, and so donor states with two electrons and acceptor states with two holes are energetically disfavored. Consider the case of a donor level. When one electron of either spin polarization is present $\big(\,\vert\uparrow\rangle$ or $\vert\downarrow\rangle\,\big)$, the energy is $E=E_{\rm d}$. When two electrons are present in the state $\vert\uparrow\downarrow\rangle$, the energy is $E=2E_{\rm d}+U$, where U is an extra Coulomb repulsion energy between the two electrons. Thus, in thermal equilibrium at temperature T, the average donor occupancy is

$$\langle\hat{n}\rangle=\frac{2\,e^{-\beta(E_{\rm d}-\mu)}+2\,e^{-\beta(2E_{\rm d}-2\mu+U)}}{1+2\,e^{-\beta(E_{\rm d}-\mu)}+e^{-\beta(2E_{\rm d}-2\mu+U)}}\quad . \tag{3.34}$$

When $U=0$, we obtain $\langle n\rangle=2f(E_{\rm d}-\mu)$. When $e^{-\beta U}\ll 1$, we have

$$\langle\hat{n}\rangle=\frac{2}{e^{\beta(E_{\rm d}-\mu)}+2}\quad . \tag{3.35}$$

More generally, for a donor with a $g_{\rm d}$-fold degeneracy of the $\hat{n}=1$, and for acceptor states with a $g_{\rm a}$-fold degeneracy of the $\hat{p}=1$ level (*i.e.* one hole), we have that the average occupancy is

$$\langle\hat{n}\rangle=\frac{g_{\rm d}}{e^{\beta(E_{\rm d}-\mu)}+g_{\rm d}}\qquad,\qquad\langle\hat{p}\rangle=\frac{g_{\rm a}}{e^{\beta(\mu-E_{\rm a})}+g_{\rm a}}\quad . \tag{3.36}$$

This means that the donor electron density and acceptor hole density are given by

$$n_{\rm d}(T,\mu)=\frac{g_{\rm d}N_{\rm d}}{e^{\beta(E_{\rm d}-\mu)}+g_{\rm d}}\qquad,\qquad p_{\rm a}(T,\mu)=\frac{g_{\rm a}N_{\rm a}}{e^{\beta(\mu-E_{\rm a})}+g_{\rm a}}\quad . \tag{3.37}$$

Typically $g_{\rm d}=2$, but in many cases there is an extra degeneracy of the acceptor states. We are left with the following equation to be solved for $\mu(T)$:

$$N_{\rm c}(T)\,e^{-(E_{\rm c}^0-\mu)/k_{\rm B}T}-P_{\rm v}(T)\,e^{-(\mu-E_{\rm v}^0)/k_{\rm B}T}=\frac{N_{\rm d}}{g_{\rm d}\,e^{\beta(\mu-E_{\rm d})}+1}-\frac{N_{\rm a}}{g_{\rm a}\,e^{\beta(E_{\rm a}-\mu)}+1}\quad . \tag{3.38}$$

Consider now the case where $E_{\rm d} - \mu \gg k_{\rm B}T$ and $\mu - E_{\rm a} \gg k_{\rm B}T$. In this case, Eqn. 3.38 becomes

$$N_{\rm c}(T)\, e^{-(E_{\rm c}^0-\mu)/k_{\rm B}T} - P_{\rm v}(T)\, e^{-(\mu-E_{\rm v}^0)/k_{\rm B}T} = N_{\rm d} - N_{\rm a} \quad . \tag{3.39}$$

From Eqn. 3.23, we have the solution

$$\left\{ \begin{matrix} n_{\rm c} \\ p_{\rm v} \end{matrix} \right\} = \tfrac{1}{2}\sqrt{(N_{\rm d}-N_{\rm a})^2 + 4n_{\rm i}^2} \;\pm\; \tfrac{1}{2}(N_{\rm d}-N_{\rm a}) \quad . \tag{3.40}$$

with $n_{\rm i}(T) = 2\,\bar{\lambda}_T^{-3/2}\, e^{-\Delta/2k_{\rm B}T}$. For light doping, where $|N_{\rm d} - N_{\rm a}| \ll n_{\rm i}$, we have

$$\left\{ \begin{matrix} n_{\rm c} \\ p_{\rm v} \end{matrix} \right\} \approx n_{\rm i} \pm \tfrac{1}{2}(N_{\rm d}-N_{\rm a}) \quad . \tag{3.41}$$

In the opposite limit, where $|N_{\rm d} - N_{\rm a}| \gg n_{\rm i}$, we have

$$N_{\rm d} > N_{\rm a} \quad : \quad n_{\rm c} \approx N_{\rm d} - N_{\rm a} \qquad , \qquad p_{\rm v} \approx \frac{n_{\rm i}^2}{N_{\rm d}-N_{\rm a}} \tag{3.42}$$

and

$$N_{\rm d} < N_{\rm a} \quad : \quad n_{\rm c} \approx \frac{n_{\rm i}^2}{N_{\rm a}-N_{\rm d}} \qquad , \qquad p_{\rm v} \approx N_{\rm a} - N_{\rm d} \quad . \tag{3.43}$$

3.3.3 Chemical potential *versus* temperature in doped semiconductors

How does the chemical potential $\mu(T)$ behave as a function of temperature in a doped semiconductor? In n-doped materials, charge neutrality requires that all the acceptor levels are singly occupied at $T = 0$, to compensate for the extra background charge. Thus, $\varepsilon_{\rm F} = \mu(T=0)$ must lie between the donor energy $E_{\rm d}$ and the conduction band minimum $E_{\rm c}^0$. For p-doped materials, $\varepsilon_{\rm F} = \mu(T=0)$ lies between the acceptor energy $E_{\rm a}$ and the valence band maximum $E_{\rm v}^0$.

What happens for large T? The answer depends on what we mean by "large", *i.e.* large compared to what? Assuming $\Delta \gg |E_{\rm d,a} - \mu|$, we have, from Eqn. 3.20, $\mu(T) = \frac{1}{2}\big(E_{\rm c}^0 + E_{\rm v}^0\big) + \frac{3}{4}\,k_{\rm B}T\,\ln\big(m_{\rm v}^*/m_{\rm c}^*\big)$. At still higher temperatures, if we make the dubious assumption that there is a further separation of energy scales which allows us to consider only the valence and conduction bands, we may write

$$2 = \int\limits_{E_{\rm v}^-}^{E_{\rm c}^+} d\varepsilon\; \tilde{g}(\varepsilon)\, f(\varepsilon-\mu) = \int\limits_{E_{\rm v}^-}^{E_{\rm c}^+} d\varepsilon\; \frac{\tilde{g}(\varepsilon)}{2+\beta(\varepsilon-\mu)+\ldots} = 2 - \beta\mu + \tfrac{1}{4}\beta \int\limits_{E_{\rm v}^-}^{E_{\rm c}^+} d\varepsilon\; \tilde{g}(\varepsilon)\,\varepsilon + \mathcal{O}(\beta^2) \quad , \tag{3.44}$$

where $\tilde{g}(\varepsilon) = \tilde{g}_{\rm v}(\varepsilon) + \tilde{g}_{\rm c}(\varepsilon)$ is the total density of states per unit cell for both bands, and where $E_{\rm v}^-$ and $E_{\rm c}^+$ denote the *lowest* energy of the valence band and *highest* energy of the conduction band, respectively. Recall that $\int\limits_{-\infty}^{\infty} d\varepsilon\; \tilde{g}_{\rm v,c}(\varepsilon) = 2$, *i.e.* each band accommodates a maximum of two electrons per cell. Thus, we conclude

$$\mu(T \gg \Delta) = \tfrac{1}{2}\langle\varepsilon\rangle_{\rm v} + \tfrac{1}{2}\langle\varepsilon\rangle_{\rm c} \equiv \bar{E}_{\rm cv} \quad , \tag{3.45}$$

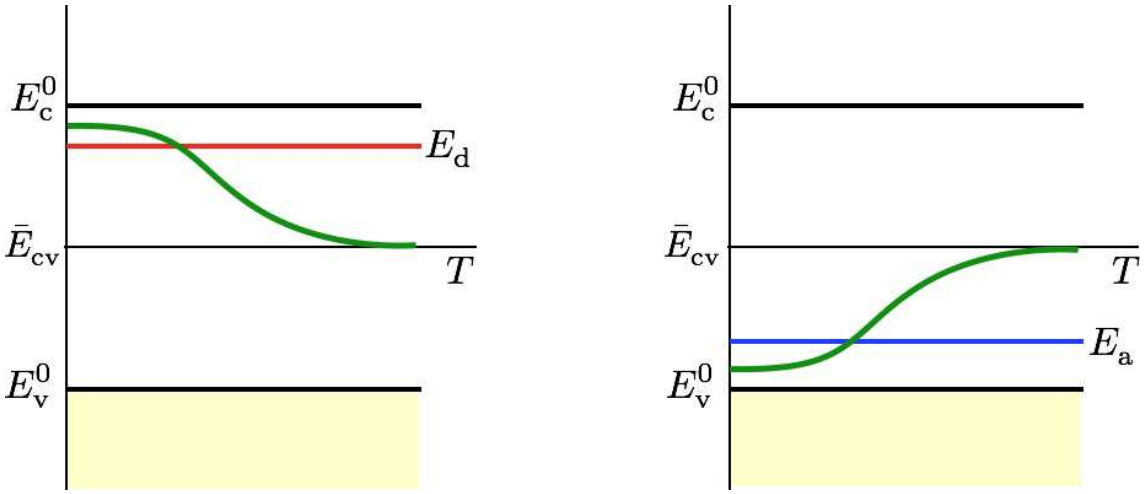


Figure 3.9: Evolution of the chemical potential with temperature in extrinsic semiconductors. The high temperature limit still assumes that only the valence and conduction bands need be considered.

where

$$\langle\varepsilon\rangle_{\rm v} = \tfrac{1}{2}\int\limits_{E_{\rm v}^-}^{E_{\rm v}^0} d\varepsilon\ \tilde{g}_{\rm v}(\varepsilon)\,\varepsilon \qquad , \qquad \langle\varepsilon\rangle_{\rm c} = \tfrac{1}{2}\int\limits_{E_{\rm c}^0}^{E_{\rm c}^+} d\varepsilon\ \tilde{g}_{\rm c}(\varepsilon)\,\varepsilon \tag{3.46}$$

are the *average* band energies. The situation is depicted in Fig. 3.9.

Of course, eventually other bands will enter the picture. In the infinite temperature limit, even the crystalline potential is irrelevant, and we can appeal to the classical result $n_\sigma = \lambda_T^{-3}\,\exp(\mu/k_{\rm B}T)$, for each spin polarization σ, where $\lambda_T = (2\pi\hbar^2/m_{\rm e}k_{\rm B}T)^{1/2}$. One then has $\mu \sim -\frac{3}{2}T\ln T$ as $T\to\infty$.

3.4 Inhomogeneous Semiconductors

Most of the technological uses of semiconductors are associated with materials which have inhomogeneous doping profiles: p-n junctions, MOSFETs, heterojunctions, *etc.* The parade example is the p-n

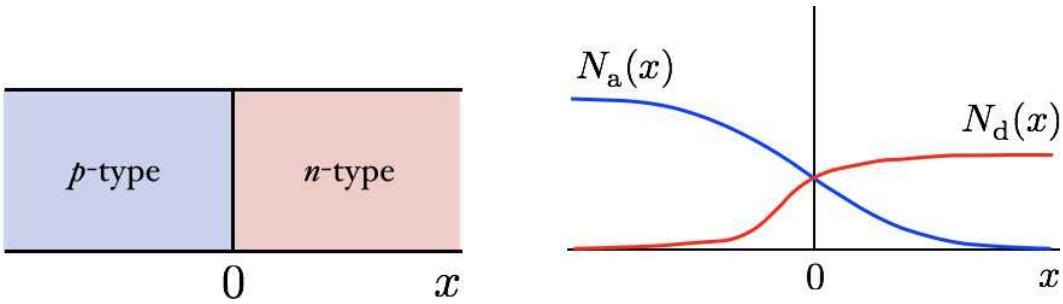


Figure 3.10: The p-n junction. Left: idealized case. Right: typical doping profile.

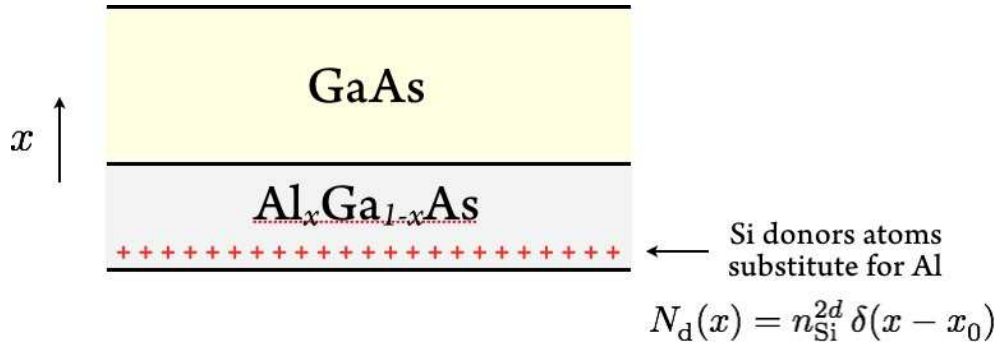


Figure 3.11: A GaAs - Al_xGa_xAs heterostructure formed by δ-doping.

junction, depicted in Fig. 3.10. We imagine a doping gradient such that the system is p-doped for $x < 0$ and n-doped for $x > 0$. Typically this is accomplished by varying the impurity concentration in a melt from which the solid is formed. Advances in growth techniques such as MBE (molecular beam epitaxy) now allow for layer-by-layer growth and doping profiles with nearly atomic precision. An important example of this is the δ-doped GaAs - Al_xGa_xAs heterostructure sketched in Fig. 3.11. As we shall discuss further below, Al_xGa_x has a larger band gap than GaAs. Substituting Si for an Al atom results in a donor, but, as we shall see, the valence electrons in the Al_xGa_x migrate over to the GaAs side of the heterostructure. By placing the Si dopants far from the interface, one thereby creates a two-dimensional electron gas (2DEG) with extremely high mobility. This feature was crucial to the 1982 discovery of the fractional quantum Hall effect in by Tsui, Störmer, and Gossard.

3.4.1 Modeling the p - n junction

Here we follow the rather clear discussion in chapter 29 of Ashcroft and Mermin, *Solid State Physics*.

We assume that the acceptor and donor densities are spatially varying according to

$$N_{\rm d}(x) = N_{\rm d}\,\Theta(x) \qquad , \qquad N_{\rm a}(x) = N_{\rm a}\,\Theta(-x) \quad . \tag{3.47}$$

In general, inhomogeneous doping along the x-direction in space will lead to a spatially varying electrostatic potential $\phi(x)$. Semiclassically, Bloch electrons in such a spatially varying potential are described by the Hamiltonian

$$H_n = E_n\left(\frac{\boldsymbol{p}}{\hbar} + \frac{e}{\hbar c}\,\boldsymbol{A}(\boldsymbol{r})\right) - e\,\phi(x) \quad . \tag{3.48}$$

We will consider the case with $\boldsymbol{B} = 0$, in which case we may choose a gauge in which $\boldsymbol{A} = 0$. Notice that the crystalline potential is not present explicitly, but rather is embodied in the Bloch dispersion $E_n(\boldsymbol{k})$. Such a description is valid provided the potential $\phi(x)$ varies slowly on atomic scales., *i.e.* $|\boldsymbol{\nabla}\phi| \ll \Delta/ae$. If we further assume that the nondegeneracy condition $|E^0_{\rm c,v} - \mu| \gg k_{\rm B}T$ holds, then we have

$$\begin{aligned} n_{\rm c}(x) &= 2\,\lambda_{T,{\rm c}}^{-3}\,\exp\left(-\frac{E^0_{\rm c} - e\,\phi(x) - \mu}{k_{\rm B}T}\right) \\ p_{\rm v}(x) &= 2\,\lambda_{T,{\rm v}}^{-3}\,\exp\left(-\frac{\mu + e\,\phi(x) - E^0_{\rm v}}{k_{\rm B}T}\right) \quad . \end{aligned} \tag{3.49}$$

Thus, at the ends of the junction, we have

$$\begin{aligned} n_{\rm c}(x=+\infty) &= 2\,\lambda_{T,{\rm c}}^{-3}\,\exp\left(-\frac{E_{\rm c}^0 - e\,\phi(+\infty) - \mu}{k_{\rm B}T}\right) = N_{\rm d} \\ p_{\rm v}(x=-\infty) &= 2\,\lambda_{T,{\rm v}}^{-3}\,\exp\left(-\frac{\mu + e\,\phi(-\infty) - E_{\rm v}^0}{k_{\rm B}T}\right) = N_{\rm a} \quad , \end{aligned} \tag{3.50}$$

where the second equality in each line follows from the analysis of §3.3.2, assuming that the condition $|N_{\rm d} - N_{\rm a}| \gg 2\,(\lambda_{T,{\rm v}}\,\lambda_{T,{\rm c}})^{-3/4}\,e^{-\Delta/2k_{\rm B}T} \equiv n_{\rm i}(T)$ holds. Multiplying these two equations yields the result

$$e\,\Delta\phi = \Delta + k_{\rm B}T\ln\!\left(\tfrac{1}{2}N_{\rm a}\lambda_{T,{\rm v}}^3\right) + k_{\rm B}T\ln\!\left(\tfrac{1}{2}N_{\rm d}\lambda_{T,{\rm c}}^3\right) \quad , \tag{3.51}$$

where $\Delta\phi \equiv \phi(+\infty) - \phi(-\infty)$ is the potential drop across the sample. We may now write

$$\begin{aligned} n_{\rm c}(x) &= N_{\rm d}\,e^{-e[\phi(+\infty)-\phi(x)]/k_{\rm B}T} \\ p_{\rm v}(x) &= N_{\rm a}\,e^{-e[\phi(x)-\phi(-\infty)]/k_{\rm B}T} \quad . \end{aligned} \tag{3.52}$$

Next we would like to determine the potential function $\phi(x)$ throughout the sample. To do this, we invoke Poisson's equation,

$$\frac{d^2\phi}{dx^2} = -\frac{4\pi\rho}{\epsilon} = \frac{4\pi e}{\epsilon}\left\{N_{\rm a}(x) + n_{\rm c}(x) - N_{\rm d}(x) - p_{\rm v}(x)\right\} \quad , \tag{3.53}$$

where we are further assuming $n_{\rm d}(x) \ll N_{\rm d}(x)$ and $p_{\rm a}(x) \ll N_{\rm a}(x)$, which are valid provided

$$\left|E_{\rm d,a} - \mu - e\,\phi(x)\right| \gg k_{\rm B}T \quad . \tag{3.54}$$

Since $n_{\rm c}(x)$ and $p_{\rm v}(x)$ depend on $\phi(x)$ through Eqn. 3.49, Eqn. 3.53 may be regarded as a nonlinear second order ODE for $\phi(x)$, rendered inhomogeneous through the appearance of the source terms $N_{\rm a}(x)$ and $N_{\rm d}(x)$. The self-consistent nature of Poisson's equation calls to mind the Debye-Hückel theory of screening in classical plasmas, except here we are not permitted to linearize in $\beta = (k_{\rm B}T)^{-1}$. To render our problem analytically tractable, we'll assume that $x > d_n$ that $n_{\rm c}(x) = n_{\rm c}(+\infty) = N_{\rm d}$ and thus

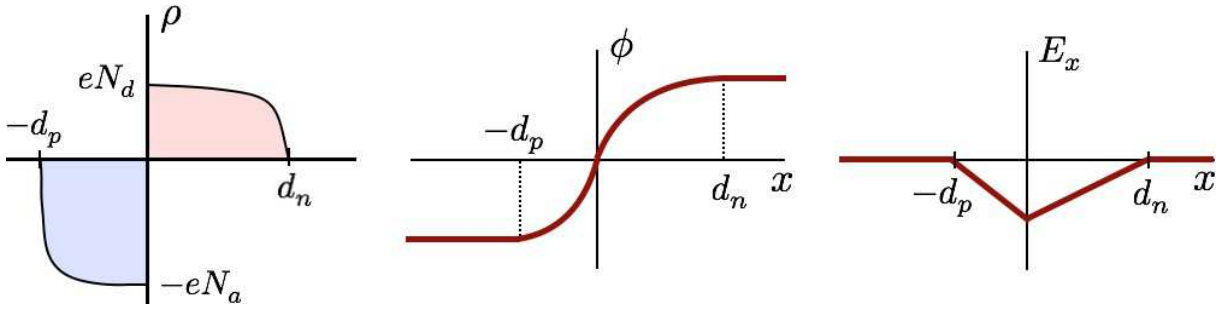


Figure 3.12: The p-n junction in equilibrium. Left: charge density $\rho(x)$. Middle: electrical potential $\phi(x)$. Right: electric field $E_x(x)$.

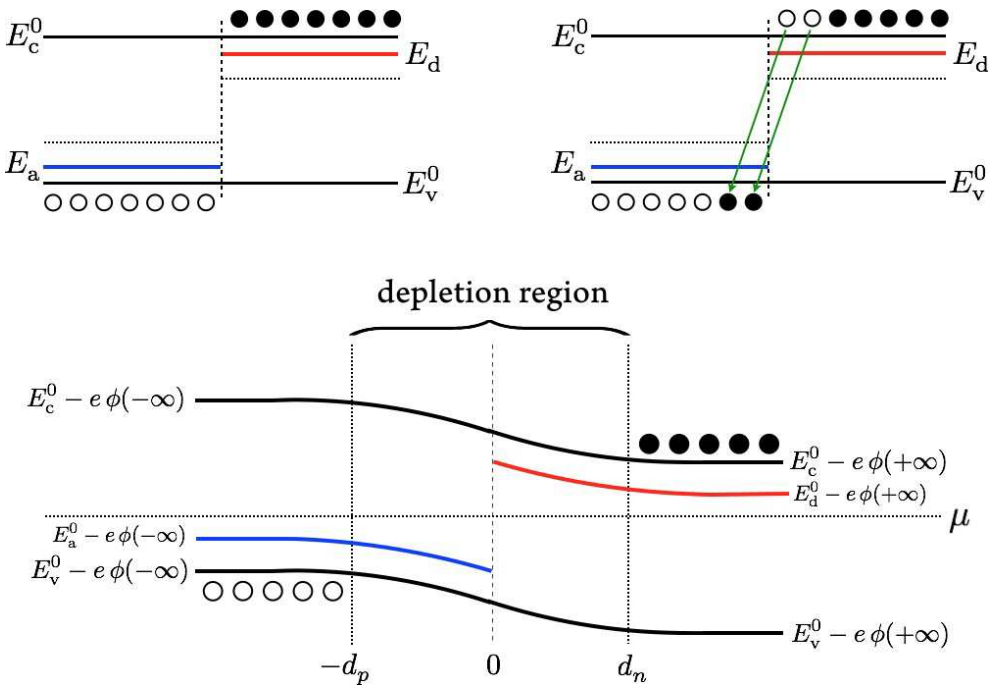


Figure 3.13: The p-n junction. Top left: p-type (left) and n-type (right) doped semiconductor at finite T with mismatched chemical potentials $\mu_{\rm n}(x)$ (dotted horizontal lines). Top right: Conduction bandelectrons fill valence band holes, thereby lowering the total energy and creating a depletion region devoid of mobile carriers, where the local imbalance of donor *versus* acceptor ions produces a local charge density $\rho(x)$ and electric potential $\phi(x)$. Bottom: The p-n junction in equilibrium. The electrochemical potential $\mu = \mu_{\rm n}(x) - e\phi(x)$ is constant throughout space. The energy bands bend with the local potential $\phi(x)$.

$\rho(x > d_n) = 0$; similarly we assume that for $x < -d_p$ that $p_{\rm v}(x) = p_{\rm v}(-\infty) = N_{\rm a}$ and $\rho(x < -d_p) = 0$. In the course of our calculations, we shall determine the unknown distances $d_{n,p}$.

Outside of the region $x \in [-d_p\,,\,d_n]$, called the *depletion region* or the *space charge later*, the charge density is zero. Within the space charge layer, we take

$$-\frac{\epsilon}{4\pi}\,\frac{d^2\phi}{dx^2} = \rho(x) \approx \begin{cases} 0 & \text{if} \quad x \le -d_p \\ -eN_{\rm a} & \text{if} \quad -d_p < x \le 0 \\ +eN_{\rm d} & \text{if} \quad 0 < x \le d_n \\ 0 & \text{if} \quad x > d_n \quad . \end{cases} \tag{3.55}$$

Integrating, we have

$$\phi(x) = \begin{cases} \phi(-\infty) & \text{if} \quad x \le -d_p \\ \phi(-\infty) + 2\pi\epsilon^{-1}eN_{\rm a}\,(x+d_p)^2 & \text{if} \quad -d_p < x \le 0 \\ \phi(+\infty) - 2\pi\epsilon^{-1}eN_{\rm d}\,(x-d_n)^2 & \text{if} \quad 0 < x \le d_n \\ \phi(+\infty) & \text{if} \quad x > d_n \quad . \end{cases} \tag{3.56}$$

We now match the potential $\phi(x)$ and its derivative $\phi'(x)$ at $x = 0$, obtaining

$$\phi(0^-) = \phi(0^+) \qquad \Rightarrow \qquad \Delta\phi = 2\pi\epsilon^{-1}e\big(N_{\rm d}d_n^2 + N_{\rm a}d_p^2\big) \tag{3.57}$$

and

$$\phi;(0^-) = \phi;(0^+) \qquad \Rightarrow \qquad d_n N_{\rm d} = d_p N_{\rm a} \quad . \tag{3.58}$$

Solving these two equations for the unknowns $d_{n,p}$, we have

$$d_p = \left[\frac{N_{\rm d}/N_{\rm a}}{N_{\rm d}+N_{\rm a}}\cdot\frac{\epsilon\,\Delta\phi}{2\pi e}\right]^{1/2} \quad , \qquad d_n = \left[\frac{N_{\rm a}/N_{\rm d}}{N_{\rm d}+N_{\rm a}}\cdot\frac{\epsilon\,\Delta\phi}{2\pi e}\right]^{1/2} \quad , \tag{3.59}$$

where $\Delta\phi$ is given in Eqn. 3.51. Typically $d_{n,p}$ are on the order of 100 Å to 1000 Å. The charge density ρ, electrical potential ϕ, and electric field $E_x = -d\phi/dx$ are all depicted in Fig. 3.12.

Let's reflect on the physics of why all this happens, following the sketches in Fig. 3.13. Conduction electrons on the n-type ($x > 0$) side of the junction can lower their energy by recombining with valence holes on the p-type side ($x < 0$). This leads to a departure from local charge neutrality, which thereby discourages further charge separation. Finally, a built-in potential $\phi(x)$ is established.

3.4.2 Rectification

Under equilibrium conditions, the electrochemical potential μ is constant across the junction. What happens if a bias voltage V is imposed? We'll define $V > 0$ (*forward bias*) as the condition where the external voltage source raises the electrical potential on the p side, and $V < 0$ (*reverse bias*) as the condition where the external voltage source raises the electrical potential on the n side. Let $\phi_0(x)$ be the $V = 0$ solution we have just derived, and let $\phi(x)$ be the solution in the presence of the external voltage source. Then

$$\phi(+\infty) - \phi(-\infty) = \phi_0(+\infty) - \phi_0(-\infty) - V \quad . \tag{3.60}$$

Most of this potential drop still occurs in the depletion region. The reason is the depletion refion is depleted of charge carriers (duh!) and therefore has a higher electrical resistance than the bulk p-type and n-type regions. As you know, in a circuit comprised of several resistors in series, the greatest potential drop occurs across the largest resistor. Therefore we have from Eqn. 3.59,

$$d_{n,p}(V) = d_{n,p}(0)\cdot\left(1 - \frac{V}{\Delta\phi_0}\right)^{1/2} \quad . \tag{3.61}$$

As shown in the sketch in Fig. 3.13, there are no conduction electrons on the p-side of the junction, nor valence holes on the n-side. Strictly speaking, this is incorrect, since thermal fluctuations will produce particle-hole excitations across the gap. This is a small but crucial effect. In a homogeneous semiconductor, these excitations would simply recombine without significant consequence, however in a p-n junction, there is an internal electric field pushing positive charges to the left and negative charges to the right. So valence holes move to the p side and conduction electrons to the n side, even at $V = 0$. This leads to a *generation current* $j_{\rm gen} = e\,(J_{\rm h} - J_{\rm e})$, where $J_{\rm h} > 0$ and $J_{\rm e} < 0$ are the number current densities

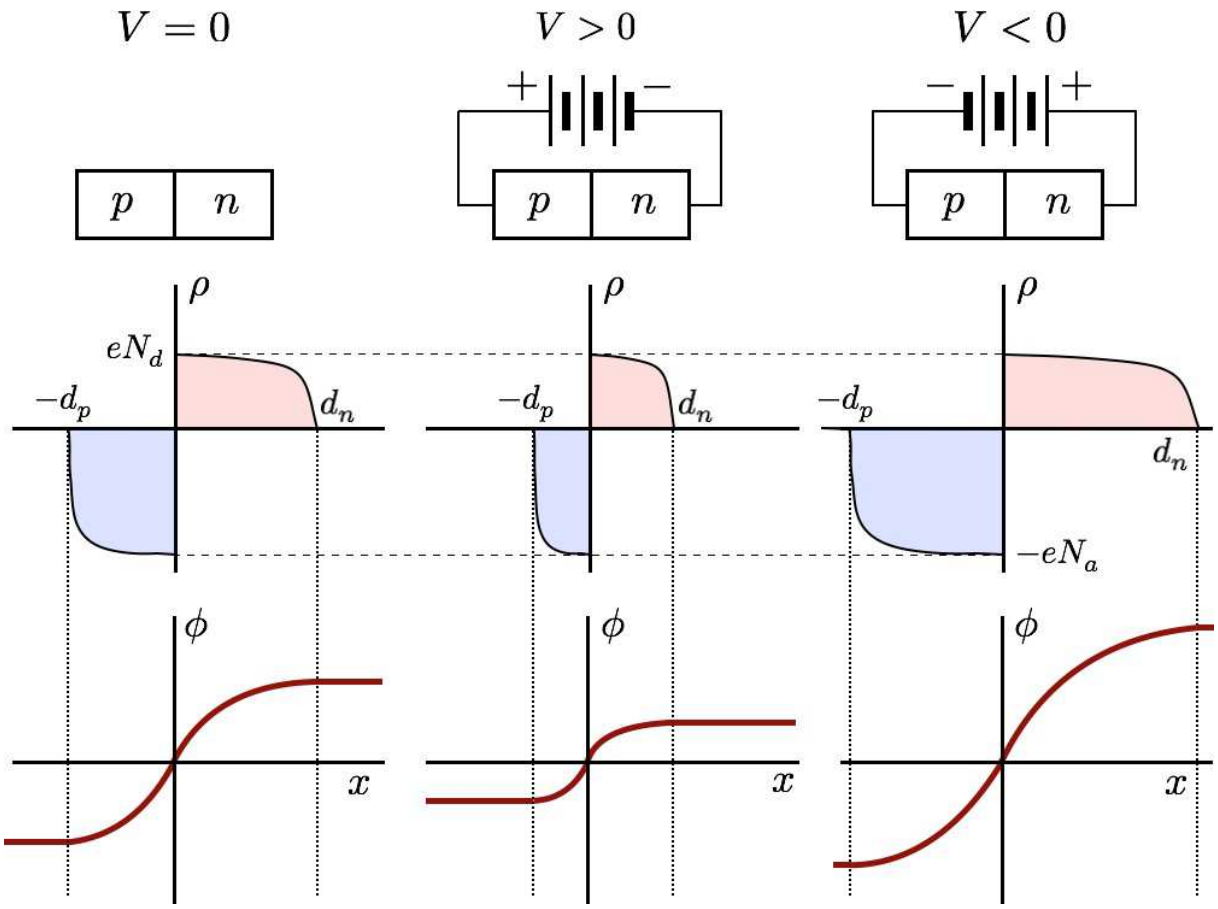


Figure 3.14: The biased p-n junction. Left: zero bias. Center: forward bias. Right: reverse bias.

of holes and electrons, respectively. Both these currents are proportional to $\exp(-\Delta/k_{\rm B}T)$ and are fairly insensitive to any bias V.

In equilibrium, there can be no net hole or electron current. Therefore there must be a counter-current of holes flowing from p to n, and of electrons flowing from n to p. These currents, which are akin to salmon swimming upstream, since they must flow in opposition to the electric field and overcome the built-in potential barrier $\Delta\phi = \Delta\phi_0 - V$, are called *recombination currents*. When $V = 0$, there is precise cancellation of the generation and recombination currents. Since $J^{\rm rec} \propto \exp\big[-e(\Delta\phi_0 - V)\big]$ and $J^{\rm rec}(V=0) = -J^{\rm gen}$ (for each species), we conclude

$$J^{\rm rec}(V) = -J^{\rm gen}\, e^{eV/k_{\rm B}T} \quad . \tag{3.62}$$

The electrical current is then

$$\begin{aligned} j(V) &= e\,J_{\rm h}^{\rm rec} + e\,J_{\rm h}^{\rm gen} - e\,J_{\rm e}^{\rm rec} - e\,J_{\rm e}^{\rm gen} \\ &= e\left(J_{\rm h}^{\rm gen} - J_{\rm e}^{\rm gen}\right)\left(1 - e^{eV/k_{\rm B}T}\right) = e\left(|J_{\rm h}^{\rm gen}| + |J_{\rm e}^{\rm gen}|\right)\left(e^{eV/k_{\rm B}T} - 1\right) \quad . \end{aligned} \tag{3.63}$$

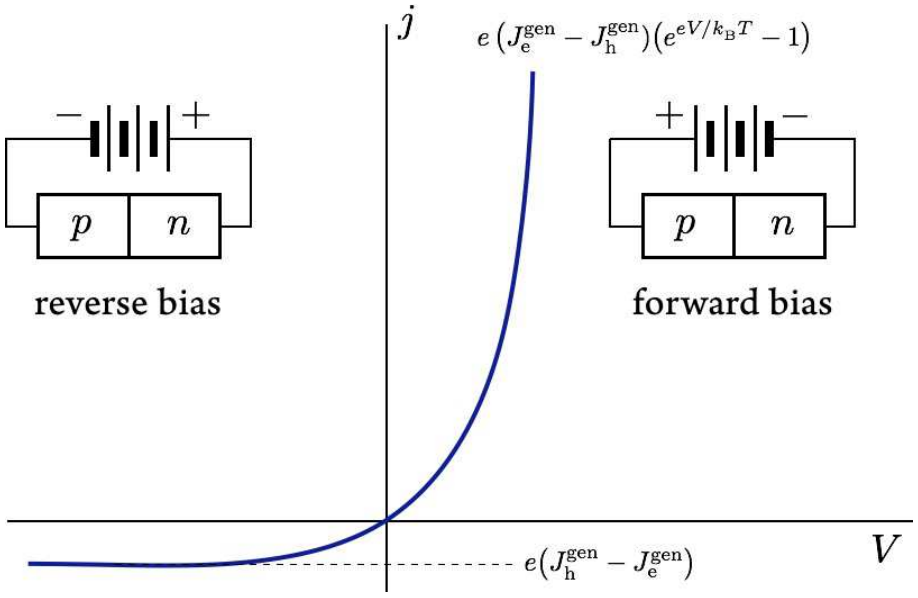


Figure 3.15: $j(V)$ for a biased p-n junction.

Note that $|J_{\rm h}^{\rm gen}|$ and $|J_{\rm e}^{\rm gen}|$ are each proportional to $\exp(-\Delta/k_{\rm B}T)$. The current-voltage relationship is sketched in Fig. 3.15. A p-n junction is thus a *current rectifier*. This is how a diode works: passing alternating current through such a junction yields a direct current.

3.4.3 MOSFETs and heterojunctions

In a metal, internal electric fields are efficiently screened and excess charge migrates rapidly to the surface, with charge density fluctuations attenuated exponentially as one enters the bulk. The Thomas-Fermi screening length, $\lambda_{\rm TF} = \left(4\pi e^2 g(\varepsilon_{\rm F})\right)^{-1/2}$, is short (a few Ångstroms) due to the large density of states at the Fermi level. In semiconductors, the Fermi level lies somewhere in the gap between valence and conduction bands, and the density of states at $\varepsilon_{\rm F}$ is quite low. Screening is due to thermally excited charge carriers, and since the carrier density is small in comparison to that in metals, the screening length is many lattice spacings.

Consider now a junction between a semiconductor and a metal, with an intervening insulating layer. This is called MIS, or metal-insulator-semiconductor, junction. If the metal is unbiased relative to the semiconductor, their chemical potentials will align. The situation for a p-type semiconductor - metal junction is depicted in the left panel of Fig. 3.16. Next consider the case in which the metal is biased negatively with respect to the semiconductor, *i.e.* the metal is placed at a negative voltage $-V$. There is then an electric field $\boldsymbol{E} = -\boldsymbol{\nabla}\phi$ pointing *out* of the semiconductor. Electric fields point in the direction positive charges want to move, hence in this case valence holes are attracted to the interface, creating an *accumulation layer* of holes, as depicted in the right panel of Fig. 3.16. On the metallic side, electrons migrate to the interface for the same reason. *No charges move across the insulating barrier.* Thus, a dipole layer is created across the barrier, with the dipole moment pointing into the semiconductor. This creates

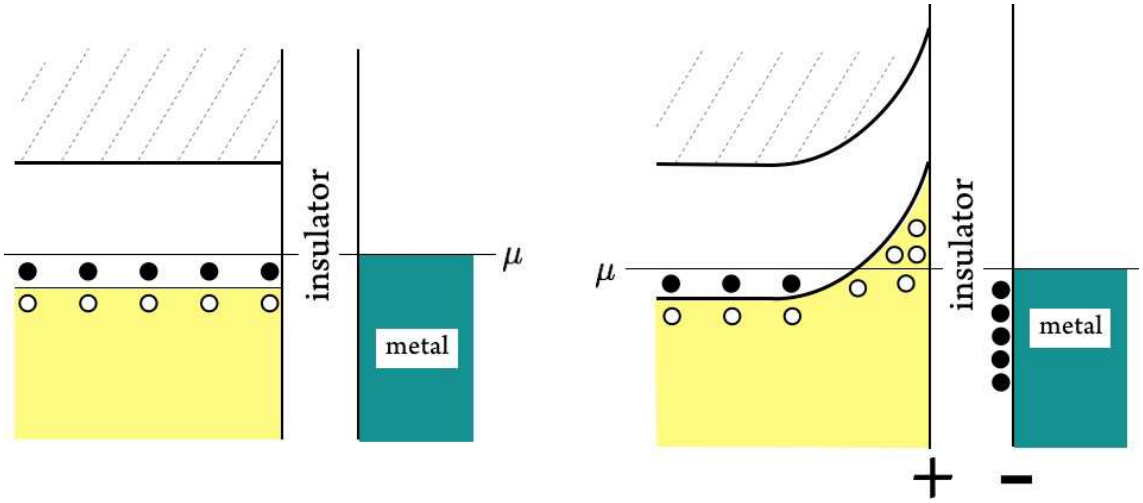


Figure 3.16: Junction between a p-type semiconductor and a metal. Left: Zero bias. Right: Metal biased negative with respect to semiconductor, creating an accumulation layer of holes and a net dipole moment at the interface.

an internal potential whose net difference $\phi_{\text{metal}} - \phi_{\text{semiconductor}}(-\infty) = V$ exactly cancels the applied bias. This condition in fact is what determines the width of the accumulation layer.

What happens when the metal is biased positively? In this case, the electric field points into the semiconductor, and valence holes are repelled from the semiconductor surface, which is then negatively charged. This, in turn, repels electrons from the nearby metallic surface. The result is a space charge *depletion layer* in the semiconductor, which is devoid of charge carriers (*i.e.* valence holes). This situation

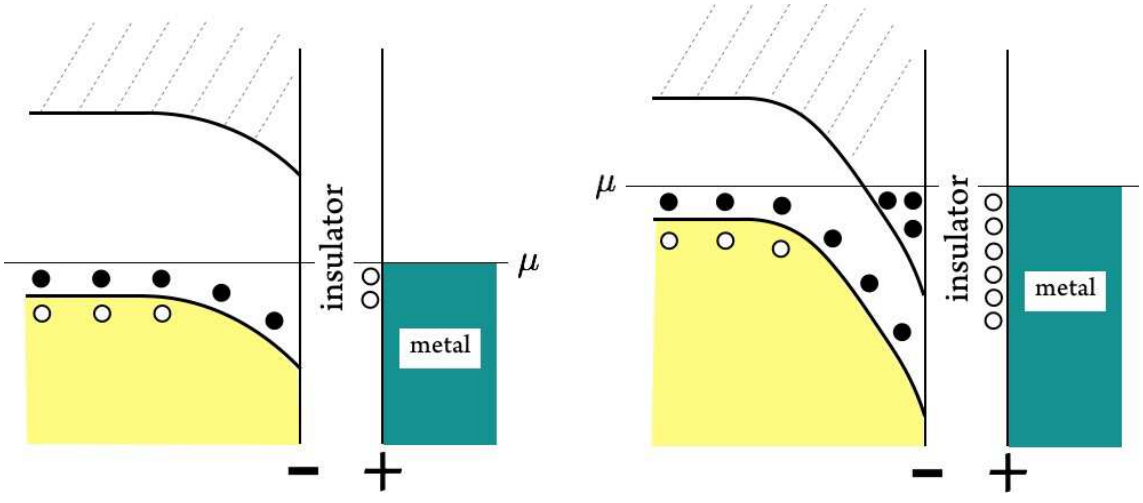


Figure 3.17: Junction between a p-type semiconductor and a metal. Left: Metal biased positive with respect to semiconductor, creating a space charge depletion layer. Right: Strong positive bias creates an inversion layer of n-type carriers in the p-type material.

is sketched in the left panel of Fig. 3.17.

Finally, what happens if the bias voltage on the metal exceeds the band gap? In this case, the field is so strong that not only are valence holes expelled from the surface, but conduction electrons are present, as shown in the right panel of Fig. 3.17. The presence of n-type carriers in a p-type semiconductor is known as *n-inversion*.

Remember this:

- *Accumulation* : presence of additional n-carriers in an n-type material, or additional p-carriers in a p-type material.
- *Depletion* : absence of n-carriers in an n-type material, or p-carriers in a p-type material.
- *Inversion* : presence of n-carriers in a p-type material, or p-carriers in an n-type material.

Inversion occurs when the presence of a depletion layer does not suffice to align the chemical potentials of the two sides of the junction.

The MOSFET

A MOSFET (Metal-Oxide-Semiconductor-Field-Effect-Transistor) consists of two back-to-back p-n junctions, and, transverse to this, a gate-bulk-oxide capacitor. The situation is depicted in Fig. 3.18. If there is no gate voltage ($V_{\rm g} = 0$), then current will not flow at any bias voltage V because necessarily one of the p-n junction will be reverse-biased. The situation changes drastically if the gate is held at a high positive potential $V_{\rm g}$, for then an n-type accumulation layer forms at the bulk-gate interface, thereby connecting

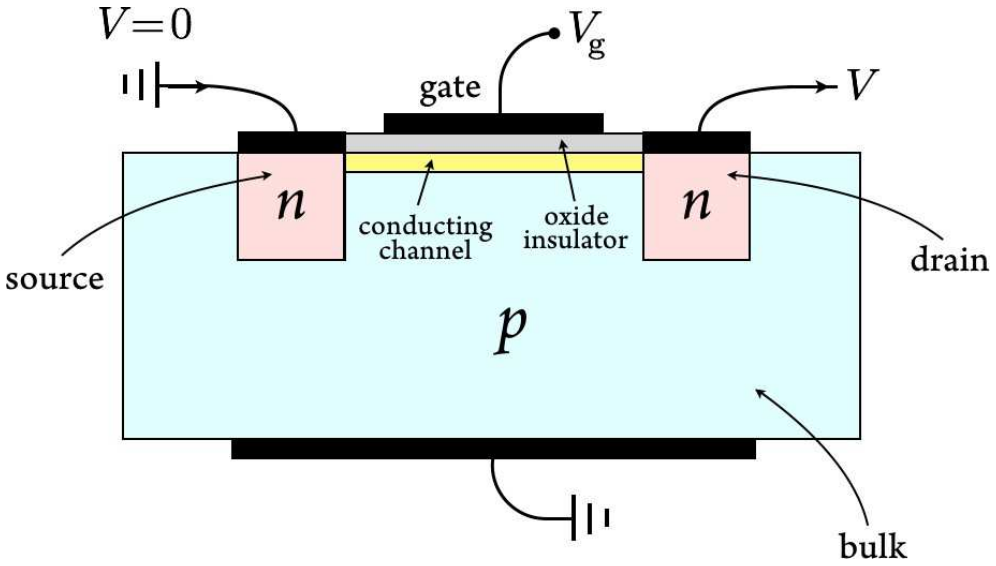


Figure 3.18: The MOSFET.

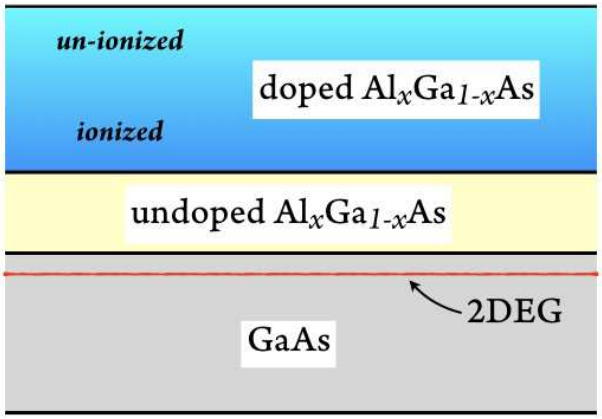


Figure 3.19: GaAs$-$Al$_x$Ga$_{1-x}$As heterojunction.

source and drain directly and resulting in a gate-controlled current flow. Although not shown in the figure, generally both source and drain are biased positively with respect to the bulk in order to avoid current leakage.

3.4.4 Heterojunctions

Potential uses of a junction formed from two distinct semiconductors were envisioned as early as 1951 by Shockley. Such devices, known as *heterojunctions*, have revolutionized the electronics industry and experimental solid state physics, the latter due to the advent of epitaxial technology which permits growth patterning to nearly atomic precision. Whereas the best inversion layer mobilities in Si MOSFETs are $\mu \approx 2 \times 10^4\ \mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$, values as high as $10^7\ \mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$ are possible in MBE-fabricated GaAs$-$Al$_x$Ga$_{1-x}$As heterostructures. There are three reasons for this:

(i) MBE (molecular beam epitaxy), as mentioned above, can produce layers which are smooth on an atomic scale. This permits exquisite control of layer thicknesses and doping profiles.

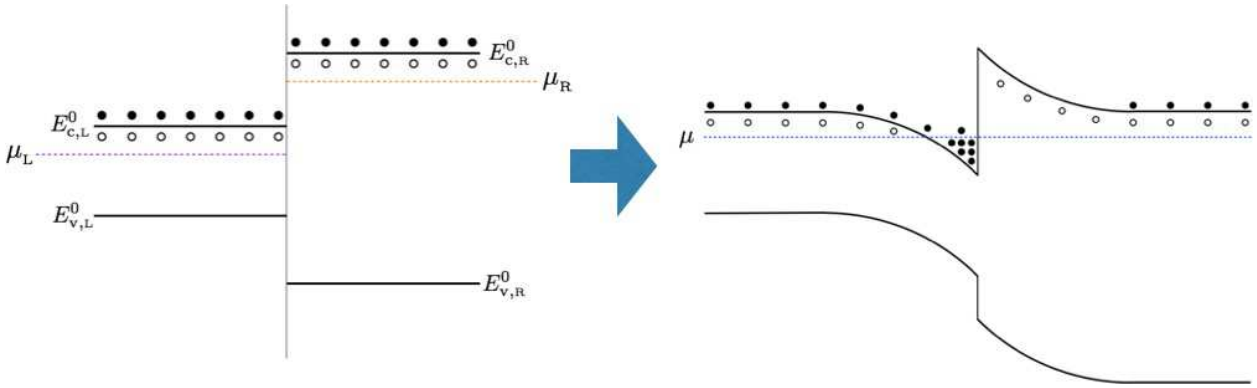


Figure 3.20: Accumulation layer formation in an n-n heterojunction.

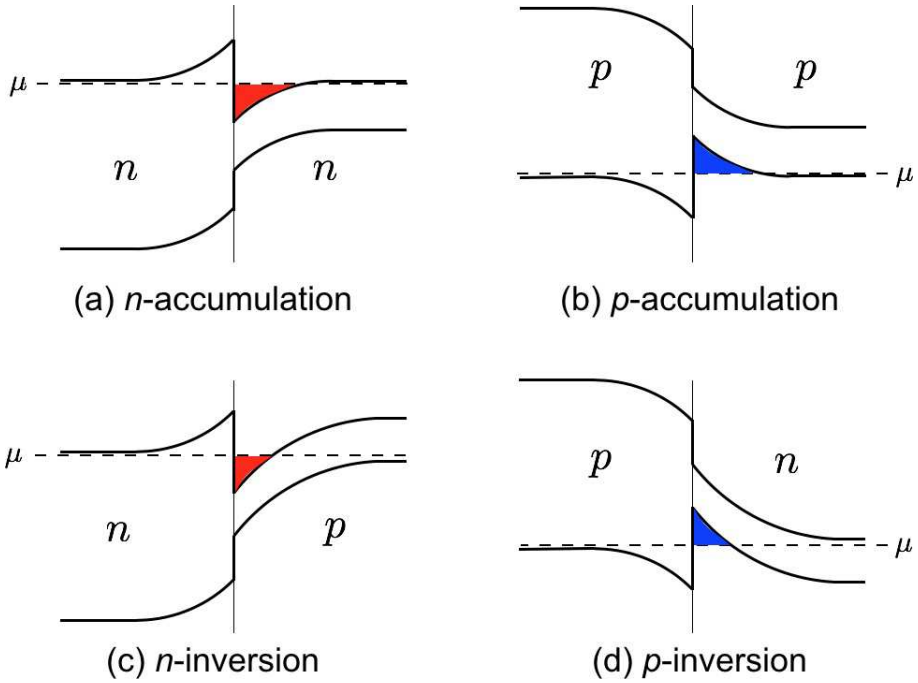


Figure 3.21: Accumulation and inversion in semiconductor heterojunctions. Red regions represent presence of conduction electrons. Blue regions represent presence of valence holes.

(ii) Use of ternary compounds such as $Al_xGa_{1-x}As$ makes for an excellent match in lattice constant across the heterojunction interface, *i.e.* on the order of or better than 1%. By contrast, the $Si-SiO_s$ interface is very poor, since SiO_2 is a glass.

(iii) By doping the $Al_xGa_{1-x}As$ layer far from the interface, Coulomb scattering between inversion layer electrons and dopant ions is suppressed.

Let's consider the chemical potential alignment problem in the case of an n-n heterojunction, sketched in Fig. 3.20. In the $GaAs-Al_xGa_{1-x}As$ heterojunction, GaAs has the smaller of the two band gaps. Initially there is a mismatch, as depicted in the left panel of the figure. By forming a depletion layer on the side with the larger band gap ($Al_xGa_{1-x}As$), and an accumulation layer on the side with the smaller gap (GaAs), an internal potential $\phi(x)$ is established which aligns the chemical potentials.

Fig. 3.21 shows the phenomena of accumulation and inversion in different possible heterojunctions. There are four possibilities: (a) n-n, (b) p-p (c) n-p with the n-type material having the larger gap, and (d) n-p with the p-type material having the larger gap.

3.5 Insulators

An insulator is a system in which there is an energy gap for charged excitations at $T = 0$. Here we shall consider a subclass known as *band insulators*, *i.e.* materials which may be adequately idealized as

noninteracting electrons in a crystalline potential with a Fermi level $\varepsilon_{\rm F}$ which lies in a band gap. Intrinsic semiconductors are then a subclass of band insulators. As we've seen above, extrinsic semiconductors can be modeled by including donor and acceptor levels to the intrinsic semiconductor band structure, and assuming that the dopant concentration is sufficiently low that the different donor/acceptor states associated with different ions may be considered noninteracting.

Insulators where the charge gap arises from strong electron-electron interactions are known as Mott-Hubbard insulators. Consider a tight-binding model with a single orbital per site, with the Hamiltonian

$$H = -\tfrac{1}{2}\sum_{i\neq j}\left(t_{ij}\,c^\dagger_{i\sigma}\,c_{j\sigma} + t^*_{ij}\,c^\dagger_{j\sigma}\,c_{i\sigma}\right) + U\sum_i n_{i\uparrow}\,n_{i\downarrow} \quad , \tag{3.64}$$

The second term on the RHS imposes an energy cost U whenever there are two electrons on the same site i. Consider now this model in which there is one electron per site. Assuming the crystal is a Bravais lattice, there is one band, and the Fermi level must cut through it in such a way that half the band is occupied and half is empty (since a filled band accommodates two electrons per site – one of each spin polarization). Thus, at $U = 0$ the system is a metal. But now consider the case where $U \gg W$, where W is the bandwidth at $U = 0$. Any state with one electron per site, with arbitrary spin polarization, *e.g.*, $|\,\psi\,\rangle = |\uparrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\cdots\rangle$ requires an energy $\Delta E \approx U - W$ to add an electron, since the added electron will necessarily result in double occupation of some site. This is called the *Mott-Hubbard gap*.

In this section, we will be concerned only with band insulators, and in particular with their dielectric properties. Mott-Hubbard systems are particularly interesting and important in condensed matter physics, but their analysis lies beyond the scope of this course.

3.5.1 Maxwell's equations in polarizable media

Maxwell's equations in the presence of sources are

$$\begin{aligned} \boldsymbol{\nabla}\cdot\mathbf{e} &= 4\pi\rho & \boldsymbol{\nabla}\times\mathbf{e} &= -\frac{1}{c}\,\frac{\partial\mathbf{b}}{\partial t} \\ \boldsymbol{\nabla}\cdot\mathbf{b} &= 0 & \boldsymbol{\nabla}\times\mathbf{b} &= \frac{4\pi}{c}\,\boldsymbol{j} + \frac{1}{c}\,\frac{\partial\mathbf{e}}{\partial t} \quad . \end{aligned} \tag{3.65}$$

Here, $\mathbf{e}(\boldsymbol{r})$, $\mathbf{b}(\boldsymbol{r})$, $\rho(\boldsymbol{r})$, and $\boldsymbol{j}(\boldsymbol{r})$ are the microscopic fields and sources. In a crystalline solid, the charge density $\rho(\boldsymbol{r})$ is a rapidly oscillating function of space, with positive delta-function like spikes at each ionic core and a negative contribution from the electron distribution which is localized for the core band states but more uniformly distributed between ions for the valence band states. The strongly fluctuating, microscopic description of electrodynamics at the atomic scale is not useful for our purposes[6].

Consider now a smoothing function $f(\boldsymbol{r})$ which satisfies $\int d^3r\; f(\boldsymbol{r}) = 1$, and define the smoothed fields

$$\boldsymbol{E}(\boldsymbol{r}) = \int d^3r'\; f(\boldsymbol{r}-\boldsymbol{r}')\,\mathbf{e}(\boldsymbol{r}') \qquad , \qquad \boldsymbol{B}(\boldsymbol{r}) = \int d^3r'\; f(\boldsymbol{r}-\boldsymbol{r}')\,\mathbf{b}(\boldsymbol{r}') \tag{3.66}$$

[6] In the following description of macroscopic electrodynamics, we follow the pellucid discussion in A. Garg, *Classical Electrodynamics in a Nutshell* (Princeton, 2012).

and the smoothed sources

$$\varrho(\boldsymbol{r}) = \int d^3r'\, f(\boldsymbol{r}-\boldsymbol{r}')\,\rho(\boldsymbol{r}') \qquad , \qquad \boldsymbol{J}(\boldsymbol{r}) = \int d^3r'\, f(\boldsymbol{r}-\boldsymbol{r}')\,\boldsymbol{j}(\boldsymbol{r}') \quad . \tag{3.67}$$

Now evaluate each of Eqns. 3.65 at $\boldsymbol{r}'$, multiply by $f(\boldsymbol{r}-\boldsymbol{r}')$, and integrate over $\boldsymbol{r}$. Integration by parts then yields the *macroscopic Maxwell equations*,

$$\begin{aligned} \boldsymbol{\nabla}\cdot\boldsymbol{E} &= 4\pi\varrho & \boldsymbol{\nabla}\times\boldsymbol{E} &= -\frac{1}{c}\,\frac{\partial\boldsymbol{B}}{\partial t} \\ \boldsymbol{\nabla}\cdot\boldsymbol{B} &= 0 & \boldsymbol{\nabla}\times\boldsymbol{B} &= \frac{4\pi}{c}\,\boldsymbol{J} + \frac{1}{c}\,\frac{\partial\boldsymbol{E}}{\partial t} \quad . \end{aligned} \tag{3.68}$$

These take the same form as the microscopic Eqns. 3.65, however, the meanings of the fields and the sources are quite different. Unfortunately, without a theory for the sources ϱ and $\boldsymbol{J}$, the above version of the macroscopic Maxwell equations is rather useless.

Macroscopic charge density

In a solid, charges may be *free*, such as conduction electrons or valence holes, or *bound*, such as the core electrons. Free charges may execute macroscopic motion in response to external fields, while bound charges remain local. To grasp the significance of bound charges, consider a system of neutral atoms or molecules, each of which has a dipole moment $\boldsymbol{d}$. If their number density is n, then the dipole moment per unit volume is $\boldsymbol{P} = n\boldsymbol{d}$, which is called the *electrical polarization*.

The electrical potential $\phi(\boldsymbol{r})$ due to a dipole at the origin is given by $\phi(\boldsymbol{r}) = \boldsymbol{d}\cdot\boldsymbol{r}/r^3$. Now consider a region Ω containing dipolar matter is then

$$\begin{aligned} \phi(\boldsymbol{r}) &= \int\limits_{\Omega} d^3r'\,\frac{\boldsymbol{P}(\boldsymbol{r}')\cdot(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3} = \int\limits_{\Omega} d^3r'\,\boldsymbol{P}(\boldsymbol{r}')\cdot\boldsymbol{\nabla}'\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} \\ &= \int\limits_{\partial\Omega} d^2r'\,\frac{\boldsymbol{P}(\boldsymbol{r}')\cdot\hat{\boldsymbol{n}}'}{|\boldsymbol{r}-\boldsymbol{r}'|} - \int\limits_{\Omega} d^3r'\,\frac{\boldsymbol{\nabla}'\cdot\boldsymbol{P}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} \\ &\equiv \int\limits_{\partial\Omega} d^2r'\,\frac{\sigma_{\rm pol}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} - \int\limits_{\Omega} d^3r'\,\frac{\varrho_{\rm pol}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} \quad , \end{aligned} \tag{3.69}$$

where $\sigma_{\rm pol} = \boldsymbol{P}\cdot\hat{\boldsymbol{n}}$ is the surface charge density and $\varrho_{\rm pol} = -\boldsymbol{\nabla}\cdot\boldsymbol{P}$ is the polarization charge density. We may now write $\varrho = \varrho_{\rm free} + \varrho_{\rm pol}$, and defining the electrical displacement field $\boldsymbol{D} \equiv \boldsymbol{E} + 4\pi\boldsymbol{P}$, we obtain the relation

$$\boldsymbol{\nabla}\cdot\boldsymbol{D} = 4\pi\varrho_{\rm free} \quad . \tag{3.70}$$

The polarization charge density is the bound charge density. While a body may be electrically neutral overall, the local charge density may vary. One writes $\boldsymbol{D} = \epsilon\boldsymbol{E}$, where ϵ is the *electric permittivity* (dielectric constant). Note that

$$\boldsymbol{P} = \frac{\epsilon-1}{4\pi}\,\boldsymbol{E} \quad . \tag{3.71}$$

Macroscopic current density

The macroscopic current density may be written $\boldsymbol{j} = \boldsymbol{j}_{\rm free} + \boldsymbol{j}_{\rm pol} + \boldsymbol{j}_{\rm mag}$, *i.e.* as a sum of contributions from the motion of free charges, from polarization currents, and from Amperean (magnetization) currents[7]. From $\varrho_{\rm pol} = -\boldsymbol{\nabla}\cdot\boldsymbol{P}$, we have that $\boldsymbol{j}_{\rm pol} = \partial\boldsymbol{P}/\partial t$, in order that the continuity equation $\partial_t \varrho_{\rm pol} + \boldsymbol{\nabla}\cdot\boldsymbol{j}_{\rm pol} = 0$ be satisfied.

Let $\boldsymbol{M}(\boldsymbol{r})$ be the magnetic dipole moment density. The vector potential due to a magnetic dipole $\boldsymbol{m}$ located at the origin is $\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{m}\times\boldsymbol{r}/r^3$, hence

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{r}) &= \int\limits_{\Omega} d^3r'\,\frac{\boldsymbol{M}(\boldsymbol{r}')\times(\boldsymbol{r}-\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|^3} = \int\limits_{\Omega} d^3r'\,\boldsymbol{M}(\boldsymbol{r}')\times\boldsymbol{\nabla}'\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} \\ &= \int\limits_{\partial\Omega} d^2r'\,\frac{\boldsymbol{M}(\boldsymbol{r}')\times\hat{\boldsymbol{n}}'}{|\boldsymbol{r}-\boldsymbol{r}'|} - \int\limits_{\Omega} d^3r'\,\frac{\boldsymbol{\nabla}'\times\boldsymbol{P}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} \\ &\equiv \frac{1}{c}\int\limits_{\partial\Omega} d^2r'\,\frac{\boldsymbol{K}_{\rm mag}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} - \frac{1}{c}\int\limits_{\Omega} d^3r'\,\frac{\boldsymbol{j}_{\rm mag}(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\quad ,\end{aligned} \tag{3.72}$$

where $\boldsymbol{K}_{\rm mag} \equiv c\,\boldsymbol{M}\times\hat{\boldsymbol{n}}$ is the surface current density and $\boldsymbol{j}_{\rm mag} = c\,\boldsymbol{\nabla}\times\boldsymbol{M}$ is the magnetization volume current. Defining the field $\boldsymbol{H} = \boldsymbol{B} - 4\pi\boldsymbol{M}$, then we have

$$\boldsymbol{\nabla}\times\boldsymbol{H} = \frac{4\pi}{c}\,\boldsymbol{j}_{\rm free} + \frac{1}{c}\,\frac{\partial\boldsymbol{D}}{\partial t}\quad . \tag{3.73}$$

One calls $\boldsymbol{E}$ the *electric field* and $\boldsymbol{D}$ the *(electric) displacement field*. Many texts call $\boldsymbol{B}$ the *magnetic induction field* and $\boldsymbol{H}$ the *magnetic field*. We will adopt the terminology of Garg (2012), who calls $\boldsymbol{B}$ the magnetic field and $\boldsymbol{H}$ the *magnetizing field*.

3.5.2 Clausius-Mossotti relation

In isotropic systems, we write $\boldsymbol{D} = \epsilon\boldsymbol{E}$ and $\boldsymbol{B} = \mu\boldsymbol{H}$. How to connect the electric permittivity (dielectric constant) ϵ and the magnetic permeability μ to microscopic quantities such the atomic polarizability α which relates the atomic or molecular dipole moment $\boldsymbol{d} = \alpha\mathbf{e}$ to the microscopic local electric field $\mathbf{e}$? We begin by writing $\mathbf{e} = \mathbf{e}_{\rm near} + \mathbf{e}_{\rm far}$ as a sum of contributions from nearby atoms and those which are far away. The distinction is that the far field is such that it is equal, to high precision, to its local average, *i.e.* $\mathbf{e}_{\rm far}(\boldsymbol{r}) = \boldsymbol{E}_{\rm far}(\boldsymbol{r}) = \int d^3r'\,f(\boldsymbol{r}-\boldsymbol{r}')\,\mathbf{e}_{\rm far}(\boldsymbol{r}')$, provided $r > R$. We expect R should be on the order of a hundred lattice spacings. Thus, $\boldsymbol{E}_{\rm far}(\boldsymbol{r})$ is the field that would exist at a point $\boldsymbol{r}$ if one were to remove all the material from a radius R about this point. The corresponding macroscopic near field, $\boldsymbol{E}_{\rm near}$, is that due to a uniformly polarized sphere of dipole density $\boldsymbol{P}$, which from elementary electrostatics is given by $\boldsymbol{E}_{\rm near} = -\frac{4\pi}{3}\boldsymbol{P}$. Since $\boldsymbol{E} = \boldsymbol{E}_{\rm near} + \boldsymbol{E}_{\rm far}$, we have that

$$\boldsymbol{E}_{\rm far} = \boldsymbol{E} + \tfrac{4\pi}{3}\boldsymbol{P} = \frac{\epsilon+2}{3}\,\boldsymbol{E} \tag{3.74}$$

[7]There is a fourth term, $\boldsymbol{j}_{\rm conv}$, due to convection currents, given by $\boldsymbol{j}_{\rm conv} = (\varrho_{\rm free} + \varrho_{\rm pol})\boldsymbol{u}$, where $\boldsymbol{u}$ is the velocity of the convective flow. Convection currents arise in liquids and gases, but not in solids.

and therefore

$$\mathbf{e} = \boldsymbol{E}_{\rm far} + \mathbf{e}_{\rm near} = \frac{\epsilon+2}{3}\,\boldsymbol{E} + \mathbf{e}_{\rm near} \quad . \tag{3.75}$$

In an isotropic system such as a liquid, or in a crystal with cubic symmetry, one has that $\mathbf{e}_{\rm near}(\boldsymbol{r}) = 0$ if $\boldsymbol{r}$ is an atomic position about which the symmetry is manifested. We then have

$$\boldsymbol{P} = \frac{\epsilon-1}{4\pi}\,\boldsymbol{E} = n\alpha\mathbf{e} = \left(\frac{\epsilon+2}{3}\right)n\alpha\,\boldsymbol{E} \quad , \tag{3.76}$$

and therefore

$$\frac{\epsilon-1}{\epsilon+2} = \frac{4\pi n\alpha}{3} \quad , \tag{3.77}$$

a result known as the *Clausius-Mossotti relation*. Solving for ϵ,

$$\epsilon = 1 + \frac{4\pi n\alpha}{1 - \frac{4\pi}{3}n\alpha} \quad . \tag{3.78}$$

If $|n\alpha| \ll 1$, which is typical, then we have

$$\epsilon \approx 1 + 4\pi n\alpha \quad . \tag{3.79}$$

3.5.3 Theory of atomic polarizability

Consider an atom with Z valence electrons. A crude classical model of atomic polarizability by writing $\boldsymbol{F} = m\boldsymbol{a}$ for each valence electron, *viz.*

$$m_{\rm e}\ddot{\boldsymbol{r}} = -k\boldsymbol{r} - e\,\mathbf{e}(t) \quad , \tag{3.80}$$

where $\boldsymbol{r}$ is the displacement from equilibrium and $\mathbf{e}$ is the microscopic electric field. The dipole moment is then $\boldsymbol{d} = Ze\,\boldsymbol{r}$. If all quantities are taken to vary as $e^{-i\omega t}$, we obtain $\hat{\boldsymbol{d}}(\omega) = \alpha(\omega)\,\hat{\mathbf{e}}(\omega)$, with

$$\alpha(\omega) = \frac{Ze^2}{m_{\rm e}(\omega_0^2 - \omega^2)} \quad , \tag{3.81}$$

where we've defined $k \equiv m_{\rm e}\omega_0^2$. Quantum mechanically, we expect $\hbar\omega_0$ to be on the order of an atomic transition energy, which is typically on the order of electron volts, and since $\hbar = 6.58 \times 10^{-16}\,{\rm eV}\cdot{\rm s}$, this corresponds to frequencies on the order of $10^{16}\,{\rm Hz}$. For $\omega \ll \omega_0$, we have $\alpha = Ze^2/m_{\rm e}\omega_0^2$, which is frequency-independent. This is valid up to frequencies $\nu_0 = \omega_0/2\pi$ which are in the ultraviolet regime.

In polar crystals, the unit cell consists of positively and negatively charged ions. Examples include III-V and II-VI semiconductors, for example. For the sake of simplicity, we analyze the case of one positive and one negative ion per cell, with charges $\pm q$, respectively. The equations of motion for the ionic vibrations are

$$\begin{aligned} M_+\ddot{u}_+^\alpha(\boldsymbol{R}) &= -\sum_{\boldsymbol{R}'}\sum_\beta \Big[\Phi_{++}^{\alpha\beta}(\boldsymbol{R}-\boldsymbol{R}')\,u_+^\beta(\boldsymbol{R}') + \Phi_{+-}^{\alpha\beta}(\boldsymbol{R}-\boldsymbol{R}')\,u_-^\beta(\boldsymbol{R}')\Big] + q\,\mathrm{e}^\alpha \\ M_-\ddot{u}_-^\alpha(\boldsymbol{R}) &= -\sum_{\boldsymbol{R}'}\sum_\beta \Big[\Phi_{-+}^{\alpha\beta}(\boldsymbol{R}-\boldsymbol{R}')\,u_+^\beta(\boldsymbol{R}') + \Phi_{--}^{\alpha\beta}(\boldsymbol{R}-\boldsymbol{R}')\,u_-^\beta(\boldsymbol{R}')\Big] - q\,\mathrm{e}^\alpha \quad , \end{aligned} \tag{3.82}$$

where

$$\Phi^{\alpha\beta}_{\eta\eta'}(\boldsymbol{R}-\boldsymbol{R}') = \frac{\partial^2 U}{\partial u^\alpha_\eta(\boldsymbol{R})\,u^\beta_{\eta'}(\boldsymbol{R}')} \tag{3.83}$$

are force constants for the lattice potential. Consider the $\boldsymbol{k}=0$ mode, where $\boldsymbol{u}_\pm(\boldsymbol{R})$ is independent of $\boldsymbol{R}$, *i.e.* all unit cell motions are in phase. Subtracting the second of the above equations from the first, we obtain

$$\begin{aligned}\ddot u^\alpha_+ - \ddot u^\alpha_- = &-\frac{1}{M_+}\sum_{\boldsymbol{R}}\Phi^{\alpha\beta}_{++}(\boldsymbol{R})\,u^\beta_+ - \frac{1}{M_+}\sum_{\boldsymbol{R}}\Phi^{\alpha\beta}_{+-}(\boldsymbol{R})\,u^\beta_- \\ &+\frac{1}{M_-}\sum_{\boldsymbol{R}}\Phi^{\alpha\beta}_{-+}(\boldsymbol{R})\,u^\beta_+ + \frac{1}{M_-}\sum_{\boldsymbol{R}}\Phi^{\alpha\beta}_{--}(\boldsymbol{R})\,u^\beta_- + \frac{q\,\mathrm{e}^\alpha}{M_+} + \frac{q\,\mathrm{e}^\alpha}{M_-} \quad .\end{aligned} \tag{3.84}$$

However, the fact that there is no restoring force for a uniform translation of the crystal requires

$$\sum_{\boldsymbol{R}}\Big[\Phi^{\alpha\beta}_{++}(\boldsymbol{R})+\Phi^{\alpha\beta}_{+-}(\boldsymbol{R})\Big] = \sum_{\boldsymbol{R}}\Big[\Phi^{\alpha\beta}_{-+}(\boldsymbol{R})+\Phi^{\alpha\beta}_{--}(\boldsymbol{R})\Big] = 0 \quad , \tag{3.85}$$

and therefore, with $\boldsymbol{\delta}\equiv\boldsymbol{u}_+-\boldsymbol{u}_-$, and assuming cubic symmetry,

$$\ddot\delta^\alpha = -\overbrace{\sum_{\boldsymbol{R}}\left(\frac{\Phi^{\alpha\beta}_{++}(\boldsymbol{R})}{M_+}+\frac{\Phi^{\alpha\beta}_{--}(\boldsymbol{R})}{M_-}\right)}^{\equiv\,\bar\omega^2\,\delta^{\alpha\beta}}\delta^\beta + \frac{q\,\mathrm{e}^\alpha}{M_-} \quad . \tag{3.86}$$

We may rewrite this as

$$\ddot{\boldsymbol{\delta}} = -\bar\omega^2\,\boldsymbol{\delta} + \frac{q\,\mathbf{e}}{M^*} \quad , \tag{3.87}$$

where $M^* = M_+M_-/(M_++M_-)$ is the reduced mass. Here $\bar\omega$ is the frequency of the $\boldsymbol{k}=0$ optical phonon. mode The polarization density is then $\boldsymbol{P}=q\boldsymbol{\delta}/v_0$, where v_0 is the unit cell volume. Solving for an oscillating electric field $\mathbf{e}\,e^{-i\omega t}$, we obtain

$$\boldsymbol{\delta}(t) = \frac{q\,\mathbf{e}}{M^*}\cdot\frac{e^{-i\omega t}}{\bar\omega^2-\omega^2} \quad . \tag{3.88}$$

We conclude that the *displacement polarizability* is

$$\alpha_{\rm disp} = \frac{q^2}{M^*(\bar\omega^2-\omega^2)} \quad . \tag{3.89}$$

Now in addition to the displacement polarizability, we also have to add in the individual atomic polarizabilities of the positive and negative ions, hence our final result is

$$\alpha(\omega) = \alpha_+ + \alpha_- + \alpha_{\rm disp}(\omega) \quad . \tag{3.90}$$

Thus, from the Clausius-Mossotti relation Eqn. 3.77, we have

$$\frac{\epsilon(\omega)-1}{\epsilon(\omega)+2} = \frac{4\pi}{3v_0}\left(\alpha_+ + \alpha_- + \frac{q^2}{M^*(\bar\omega^2-\omega^2)}\right) \quad , \tag{3.91}$$

provided $\omega \ll \omega_0$. We then have, at $\omega = 0$,

$$\frac{\epsilon_0 - 1}{\epsilon_0 + 2} = \frac{4\pi}{3v}\left(\alpha_+ + \alpha_- + \frac{q^2}{M^*\bar{\omega}^2}\right) \quad , \tag{3.92}$$

while for $\bar{\omega} \ll \omega \ll \omega_0$, which we call 'infinite' frequency for our purposes,

$$\frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} = \frac{4\pi}{3v}\left(\alpha_+ + \alpha_-\right) \quad . \tag{3.93}$$

These equations allow us to write

$$\epsilon(\omega) = \epsilon_\infty \cdot \frac{\omega^2 - \omega_{\text{L}}^2}{\omega^2 - \omega_{\text{T}}^2} \quad , \tag{3.94}$$

where

$$\omega_{\text{T}} = \left(\frac{\epsilon_\infty + 2}{\epsilon_0 + 2}\right)^{1/2}\bar{\omega} \qquad , \qquad \omega_{\text{L}} = \left(\epsilon_0/\epsilon_\infty\right)^{1/2}\omega_{\text{T}} \quad . \tag{3.95}$$

Note that $\epsilon(\omega_{\text{L}}) = 0$ and $\epsilon(\omega_{\text{T}}) = \infty$, and that

$$\left(\frac{\omega_{\text{L}}}{\omega_{\text{T}}}\right)^2 = \frac{\epsilon_0}{\epsilon_\infty} > 1 \quad , \tag{3.96}$$

a result known as the *Lyddane-Sachs-Teller relation*. The behavior of $\epsilon(\omega)$ is sketched in the left panel of Fig. 3.22.

3.5.4 Electromagnetic waves in a polar crystal

Consider now the propagation of electromagnetic waves in a polar crystal. Assuming the absence of free charges, we have $\boldsymbol{\nabla}\cdot\boldsymbol{D} = 0$ and $\boldsymbol{\nabla}\times\boldsymbol{E} = -c^{-1}\partial_t\boldsymbol{B}$ governing the macroscopic fields. We will ignore the $c^{-1}\partial_t\boldsymbol{B}$ term and justify this later on. A plane wave solution with $\boldsymbol{E}$ and $\boldsymbol{D}$ both proportional to $\exp(i\boldsymbol{k}\cdot\boldsymbol{r})$ thereby requires $\boldsymbol{k}\cdot\boldsymbol{D} = \boldsymbol{k}\times\boldsymbol{E} = 0$, which has no nontrivial solutions if $\boldsymbol{D} = \epsilon(\omega)\,\boldsymbol{E}$. *I.e.* either $\boldsymbol{E} = 0$ or $\boldsymbol{D} = 0$. Thus we have the following two possibilities:

- *longitudinal mode* : $\boldsymbol{E} \parallel \boldsymbol{k}$ with $\epsilon = 0$, hence $\omega = \omega_{\text{L}}$ and $\boldsymbol{D} = 0$.
- *transverse mode* : $\boldsymbol{D} \perp \boldsymbol{k}$ with $\epsilon = \infty$, hence $\omega = \omega_{\text{T}}$ and $\boldsymbol{E} = 0$.

These conclusions hold valid in the $\boldsymbol{k} \to 0$ limit. To find the dispersion for general $\boldsymbol{k}$, we need to solve Eqns. 3.82 under general conditions, *i.e.* not assuming all the unit cells are in phase. This yields a dispersion as shown in the right panel of Fig. 3.22.

Why were we allowed to drop the $c\,\partial_t\boldsymbol{B}$ term in Faraday's equation? This is because it is negligible in the limit $ck \gg \omega$. Since optical frequencies are on the order of that of zone edge phonons, this means we must satisfy $k \gg (\pi/a)\cdot(s/c)$, where a is the lattice spacing, s is the acoustic phonon velocity, and c the

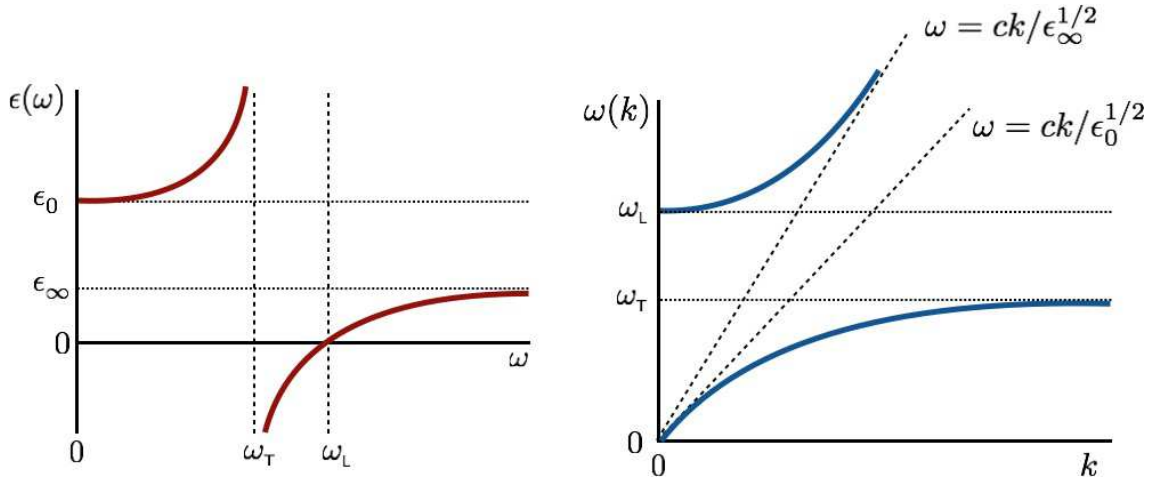


Figure 3.22: Left: dielectric function $\epsilon(\omega)$ in a polar crystal. Right: solution to the equation $\omega = ck/\epsilon^{1/2}(\omega)$ showing the polariton dispersion branches $\omega_\pm(k)$.

speed of light. Since $s/c \sim 10^{-5} - 10^{-4}$, we are in good shape provided k is not extremely close to the zone center. Finally, since the reflectivity is

$$R(\omega) = \left| \frac{\sqrt{\epsilon(\omega)} - 1}{\sqrt{\epsilon(\omega)} + 1} \right|^2 \quad , \tag{3.97}$$

for $\omega \in [\omega_{\rm T}\,,\,\omega_{\rm L}]$ the dielectric function $\epsilon(\omega)$ is purely imaginary and thus the crystal is purely reflecting.

Nota bene : The "charge" q is poorly defined, since it is spread out in a continuous distribution rather than a Dirac delta function. Thus, Eqn. 3.94 and the LST relation are much more useful than Eqn. 3.91, since we can always measure ϵ_0 and ϵ_∞.